Machine learning algorithms in the health sector : integration of functional knowledge toenhance the analysis of gut microbiota data

Baptiste Ruiz

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Informatique*

Par

Baptiste RUIZ

Machine learning applied to functional representation of patients' microbiota : robustness analysis and interpretability.

Thèse présentée et soutenue à Rennes, le 28/11/2024
Unité de recherche : INRIA Rennes

Rapporteurs avant soutenance :

Julien CHIQUET — Directeur de recherche, INRAE Paris Saclay, UMR Mathématiques et Informatique Appliquée
Nathalie VIALANEIX — Directrice de recherche, INRAE Toulouse, UMR Mathématiques et Informatique appliquées de Toulouse

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Examinateurs :	Laurent BREHELIN	Chargé de recherche, CNRS, LIRMM, Montpellier
	Julien CHIQUET	Directeur de Recherche, INRAE, MIA, Paris-Saclay
	Damien EVEILLARD	Professeur, Univ Nantes, LS2N, Nantes
	Yann LE CUNFF	Maitre de conférences, Univ Rennes, IRISA, Rennes
	Nathalie VIALANEIX	Directrice de recherche, INRAE Toulouse, UMR Mathématiques et Informatique appliquées de Toulouse
Dir. de thèse :	Anne SIEGEL	Directrice de recherche, CNRS, IRISA, Rennes
Co-dir. de thèse :	Isabelle LE HUËROU-LURON	Directrice de recherche, INRAE, Numecan, Rennes

Acknowledgement

Lectrice, lecteur, sois prévenu · e, ce document est rédigé principalement en anglais! Si je reprend ici l'usage de la langue de Molière, c'est surtout parce qu'il s'agit de la langue commune à toutes les personnes auxquelles je dois exprimer mes remerciements, et qu'après délibération, j'ai fini par décider que ce serait quand même plus sympa pour elles de comprendre ce que je raconte, même pour les moins adeptes de la langue de Shakespeare. Si vous êtes de ceux-là, cette section est faite pour vous, ainsi que le résumé qui suit. La suite sera peut-être un peu plus ardue...

Trois ans, même quand ça passe vite, mine de rien c'est long. Suffisamment en tout cas pour que beaucoup de choses se passent. Sur le plan personnel, comme sur le plan professionnel, je me retrouve à des kilomètres de là où j'ai démarré. Et si j'ai parcouru ce chemin, c'est surtout parce que j'ai été bien accompagné. J'aimerais donc adresser mes remerciements:

— aux membres du jury, pour avoir pris le temps de me lire, de m'écouter, et (peut-être? Je ne sais pas encore, c'est dans le futur proche...) de me torturer le jour de la soutenance. J'adresse une mention spéciale aux rapporteurs, pour avoir considéré que mon travail n'était pas suffisamment inintéressant pour m'épargner ladite soutenance.

— à mes encadrantes et encadrant. Je me rends compte de la chance que j'ai eu de travailler dans les conditions auxquelles j'ai eu droit, et ça je le dois en grande partie à vos conseils et à votre bienveillance. Le projet de thèse s'est construit sur des sujets qui vous tenaient à cœur, et je reste très touché que vous me l'ayiez confié pour ces trois dernières années. J'espère que le résultat est à la hauteur de vos attentes.

— plus spécifiquement, à Yann, qui a probablement le plus eu à me supporter pendant ces trois dernières années, et je ne me considère pas forcément comme un cadeau... Merci pour tous tes conseils, j'admire énormément ta patience et ta pédagogie. Tes élèves ont beaucoup de chance de t'avoir comme enseignant, et moi d'avoir eu droit à tes cours particuliers lorsque je bloquais.

— à Isabelle, pour m'avoir gardé un pied dans la biologie de laquelle je viens. Tu as

toujours été de bon conseil et de bon soutien lorsque j'avais mes paniques de fin de semaine. C'était toujours un plaisir de faire le trajet jusqu'à Saint-Gilles pour discuter!

— à Sylvie, pour ton enthousiasme et ton dynamisme. Lorsque tu me faisais faire le tour des locaux du CHU, ta fierté était contagieuse. Merci de m'avoir fait une petite place à côté de ton bureau, et pour tous les gens que tu m'as fait rencontrer en leur vendant mon projet mieux que je n'aurais pu le faire!

— à Anne, pour tout ce que tu as fait au cours non pas de ces trois dernières années, mais des cinq dernières. Si j'ai atterri à Rennes, c'est parce que tu m'as donné ma chance en répondant à ma bouteille à la mer, lancée lorsque je cherchais un stage de fin d'année. Et ce faisant, tu m'as permis d'écrire un des plus beaux chapitres de ma vie! Merci pour ça, et pour ton accompagnement tout du long, avec ta capacité à jouer à Tetris avec ton calendrier pour toujours trouver du temps et être à l'écoute. Vraiment, je ne sais pas comment tu fais...

— à l'école doctorale ~~MathSTIC~~ MATISSE, l'Université ~~Rennes 1~~ de Rennes et l'INRIA ~~Rennes Bretagne Atlantique~~ de l'Université de Rennes pour tout l'encadrement administratif de cette thèse.

— à Arnaud, sans tes contributions et ton soutien, ma thèse tiendrait sur deux pages. Ca aurait été moins de travail, mais probablement moins intéressant aussi.

— à Mahendra Mariadassou et Hédi Soula, pour leurs conseils et retours, toujours enrichissants, en CSID.

— à Olivier, Jeanne et Pauline, pour les discussions scientifiques enrichissantes sur le projet. Je n'ai pas pu intégrer tout ce dont on a discuté, mais ma section Perspectives vous doit beaucoup.

— à mes camarades doctorants, notamment mes partenaires de Sciences en Cour[t]s: Sandra, dont l'appartement a servi tour à tour de décor et de salle de cinéma, Roland, camarade cuistot parti pour le pays de la frite, Kerian, sauveur de patates et dealeur officiel de félins, et Khodor, dealeur officiel de pâtisseries libanaises. Partager ces moments de créativité avec vous (et aussi les angoisses au moment de la rédaction, mais ça on préfère oublier) a été un réel plaisir!

— à Marie, qui sauve quotidiennement l'ensemble de l'équipe (moi inclus) sur le plan administratif!

— à l'intégralité de l'équipe Dyliss, mais aussi GenScale et GenOuest. Merci pour cette ambiance formidable que vous cultivez tous ensemble, merci pour les discussions

(scientifiques ou non) en réunion, au bureau ou à la cafétéria, toutes mes excuses pour mon sens de l'humour en pause. Si je me retrouve dans un environnement à moitié aussi agréable pour la suite de mon parcours pro, j'aurai touché le jackpot!

— à l'équipe GenOuest encore, mais pour des aspects plus pratiques. Sans votre plateforme, je n'aurais pas eu de résultats, et ça m'aurait un peu mis dans l'embarras... Merci donc de m'avoir tiré de ce mauvais pas!

— à l'intégralité des équipes EAT et NuMeCan à l'INRAE Saint-Gilles. J'étais moins souvent de passage dans votre beau centre, mais j'ai énormément apprécié l'ambiance qui régnait chez vous. Merci pour les animations scientifiques, les discussions dans le bureau, et le café (en dosette, pas trop celui de la cafétéria...).

Sur un plan plus personnel maintenant, un immense merci à ma famille, pour m'avoir soutenu dans ce projet dès le début. Votre intérêt pour tout ce que je fais me touche beaucoup, j'espère que mes explications étaient suffisamment claires pour vous. Merci à Marie, pour les discussions et le soutien mutuel hors du cadre de nos équipes respectives, allant même jusqu'à synchroniser nos soutenances (désolé de ne pas assister à la tienne du coup, on se reverra à un concert pour rattraper ça). Et merci enfin à Alice, probablement le changement le plus important dans ma vie depuis le début de ce projet. Merci d'avoir accepté de venir me rejoindre ici alors que c'était compliqué pour moi de bouger chez toi pendant cette période, merci pour ton soutien inconditionnel pendant les périodes difficiles, et merci pour ta bonne humeur quotidienne. Je t'aime.

Merci finalement à vous, lectrice, lecteur, d'être parvenu · e à la fin de cette longue liste. J'arrête de m'étendre désormais sur ces considérations personnelles, et je vous laisse découvrir les travaux que j'ai pu mener grâce à toutes les personnes précédemment citées.

Table of Contents

Résumé de la thèse - Summary in French

Depuis plusieurs années, le domaine médical s'intéresse de près au microbiome intestinal humain et aux perspectives qu'il ouvre, car sa composition s'est révélée avoir un large éventail d'impacts insoupçonnés sur la santé de l'hôte. Initialement, l'intestin humain était considéré comme un conduit relativement passif pour la digestion et l'absorption des nutriments. Cependant, les progrès de la biologie moléculaire et des technologies de séquençage ont révélé que le microbiome intestinal est un écosystème complexe et dynamique, qui fait partie intégrante de divers processus physiologiques. Ce changement de paradigme a mis en évidence l'influence profonde du microbiome sur les fonctions métaboliques, immunitaires et neurologiques, le positionnant comme un acteur clé de la santé générale de son hôte. Ce constat a conduit à des recherches approfondies sur sa composition et son potentiel en tant que vecteur de compréhension, de traitement et de prévention des maladies. Au travers de l'application de diverses méthodes séquençage, telles que le séquençage métagénomique Shotgun (MGS) ou de l'ARNr 16S, il a été possible d'identifier et de quantifier la composition du microbiome intestinal *in silico*.

Ces profils, généralement construits au niveau taxonomique, ont été explorés en tant que base pour la classification supervisée des individus et l'identification de marqueurs bactériens associés à des états pathologiques spécifiques. En particulier les modèles de forêts aléatoires (RF), se sont révélées très efficaces pour classifier les données du microbiome intestinal et prédire l'état de santé de l'hôte. Bien que les descriptions taxonomiques aient été plus largement étudiées, il émerge un intérêt croissant à comprendre les aspects fonctionnels du microbiome intestinal: plutôt que de comprendre *qui* sont les espèces les plus influentes dans la communauté, il serait plus pertinent de comprendre *ce qu'elles font*. Des outils tels que HUMAnN et PiCRUSt ont été développés pour représenter le microbiome intestinal à l'échelle fonctionnelle, en déduisant les voies métaboliques et les molécules exprimées dans un échantillon donné. Ces approches, bien que largement plébiscitées et utilisées par la communauté, ont toutefois quelques limites. D'une part, ces approches sont spécifiques à un unique type de séquence (MGS pour HUMAnN, 16S pour PiCRUSt), ce qui limite leur applicabilité. Ces processus sont aussi particulièrement lourds, du point de vue de la taille des entrées requises comme du temps de calcul.

Une analyse des applications de l'apprentissage automatique au microbiome intestinal fait ressortir des possibilités d'amélioration du point de vue des performances, de la robustesse et de l'interprétabilité de ces approches. L'un des freins majeurs à l'application de méthodes d'apprentissage automatique à la composition du microbiome intestinal est la dimension de ces données, qui sont défavorables à l'entraînement des classifieurs. Pour résoudre ce problème, des méthodes de sélection de variables ont été utilisées pour réduire la dimension des données et améliorer les performances de classification. Ces approches se sont parfois même basées sur les classifieurs eux-mêmes, en exploitant les classements d'importance de variables issus des RF. De plus, de nombreuses autres pistes restent possibles pour améliorer la fiabilité et l'interprétabilité des études sur le microbiome intestinal par biais de l'apprentissage automatique, par exemple en mettant l'accent sur la robustesse de l'entraînement et de l'évaluation des modèles, et en approfondissant l'exploration de la signification biologique des variables sélectionnées. L'intégration d'information fonctionnelle à ces analyses reste également rare, alors que son exploitation, ainsi possiblement que sa mise en rapport avec l'information taxonomique, s'inscrirait mieux dans la lignée des exigences actuelles de la communauté médicale.

C'est dans ce cadre que s'inscrivent les travaux de cette thèse, au cours de laquelle a été développée une approche d'analyse du microbiote intestinal basée sur l'apprentissage automatique. Cette approche inclut une nouvelle méthode pour la construction de profils fonctionnels du microbiote. Cette nouvelle approche requiert uniquement des profils taxonomiques, ce qui la rend plus légère et rapide que les approches de la littérature basées sur les séquences brutes. Cela est suivi par une approche analytique incorporant un entraînement répété de classifieurs RF sur des données taxonomiques et fonctionnelles, associés à une sélection itérative et automatique de variables, ainsi qu'une reconduction complète du processus à plusieurs reprises. Il en ressort une optimisation des performances de classification et une sélection de variables robustes. Les variables obtenues sont ensuite analysées, d'une part pour évaluer leur pertinence au regard de la bibliographie, mais aussi en mettant en valeur, par la visualisation des intercorrélations entre les échelles fonctionnelle et taxonomique, des effets d'accumulation faisant ressortir des signatures fonctionnelles non dérivables de l'analyse des profils taxonomiques. Enfin, les approches précédemment décrites ont été implémentées, et rendues disponibles pour permettre une application plus large de cette méthode.

Une méthode légère pour le calcul de profils fonctionnels du microbiome.

Nous présentons d'abord une nouvelle méthode de représentation fonctionnelle du microbiome intestinal, qui s'appuie uniquement sur une profil taxonomique comprenant une description taxonomique et une quantification des taxons présents dans le microbiome séquencé. Contrairement aux méthodes existantes telles que PiCRUSt et HUMAnN, spécifiques à des types de séquençage particuliers, et qui nécessitent des informations supplémentaires ou les séquençages bruts, cette méthode est générique et peut être appliquée aux profils dérivés du séquençage de l'ARNr 16S ou MGS. Cette approche par référencement s'avère également beaucoup plus rapide que l'analyse de séquences brutes opérée par HUMAnN3, à laquelle nous nous somme comparés.

La méthode repose sur le pipeline EsMeCaTa, qui associe des annotations fonctionnelles (FA) aux unités taxonomiques données en entrée via l'interrogation de bases de données telles qu'UniProt ou eggNOG. En combinant ces informations avec les abondances taxonomiques d'origine, la méthode calcule des scores d'abondance pour les FA au sein du microbiome intestinal. Une comparaison avec les résultats générés par HUMAnN3 sur un jeu de données d'exemple a montré qu'EsMeCaTa extrayait davantage d'informations, récupérant la majorité des annotations trouvées par HUMAnN3 tout en découvrant des annotations supplémentaires. De plus, l'analyse a révélé que les profils fonctionnels et taxonomiques générés par la méthode ne sont pas redondants, ce qui suggère qu'ils offrent des perspectives complémentaires sur la composition et la fonction du microbiome intestinal.

Une classification robuste et interprétable du microbiome, permise par une correction des dimensions des données.

Nous abordons ensuite les défis posés par la dimension des données du microbiome en ce qui concerne la classification supervisée, et présentons une nouvelle méthode de classification et de sélection de variables adaptée pour les jeux de données aux dimensions déséquilibrées. Les descriptions taxonomiques du microbiome comportent souvent un nombre d'échantillons inférieur au nombre de descripteurs, ce qui, suivant le phénomène de Hughes, impacte négativement les performances des classifieurs. Ce problème est exacerbé dans le cas des profils fonctionnels, où le nombre de descripteurs augmente considérablement.

La méthode proposée utilise des scores d'importance de Gini moyens issus de multiples RFs entraînées pour effectuer une sélection automatique de variables, en relevant les caractéristiques les plus discriminantes pour faire la distinction entre les patients et les individus sains. Le processus de sélection opéré se distingue des propositions classiques par son automatisation complète: la séparation des variables se fait au point d'inflexion de la courbe des scores d'importance décroissants, ce qui ne requiert aucune intervention humaine externe et contourne donc les biais associés. Ce processus est répété de manière itérative, et son application à des jeux de données publics et classiques nous permet de mettre en valeur que cette sélection de variables a tendance à améliorer les performances de classification, notamment en ce qui concerne les profils fonctionnels. Nous montrons également qu'à leurs meilleurs niveaux de sélection respectifs, les profils taxonomiques et fonctionnels fournissent des performances de classification comparables.

Ce processus de classification et de sélection est répété entièrement plusieurs fois, à chaque fois avec un jeu de test différent mis de côté. Cela permet d'évaluer la robustesse des classifications, mais également des sélections de variables. En effet, la variation des conditions d'entraînement impliquée par le changement de jeu test fait que les sélections de variables divergent à chaque répétition. Nous proposons ainsi une nouvelle approche pour évaluer la robustesse d'une sélection de variables opérée dans ces conditions, en mesurant la fiabilité de chaque variable en fonction du nombre de fois où elles ont été sélectionnées. Celles qui constituent l'intersection de toutes les sélections sont qualifiées de 'robustes', celles qui ont été sélectionnées 75% du temps ou plus sont 'confiantes', et celles qui sont sélectionnées au moins une fois sont 'candidates'.

Une évaluation comparative avec limma, un outil statistique couramment utilisé pour l'analyse d'abondance différentielle, permet de démontrer la robustesse et la fiabilité de la méthode proposée. La méthode basée sur les RF donne des sélections robustes de variables dans tous les jeux de données testés, tandis que limma a donné des sélections vides ou minimes dans plusieurs cas. De plus, la méthode RF s'est avérée capable de sélectionner des variables non identifiées par limma, ce qui souligne sa capacité à découvrir des signaux non-linéaires et potentiellement nouveaux au sein des données du microbiome.

L'interconnexion entre taxons et fonctions met en lumière des expressions fonctionnelles cumulatives au sein du microbiome intestinal.

Par la suite, nous explorons la signification biologique complémentaire des profils taxonomiques et fonctionnels dans le contexte de l'étude du microbiome intestinal. Nous soutenons que si les taxons individuels peuvent fournir des informations précieuses sur la santé de l'hôte, la compréhension des fonctions métaboliques exprimées par le microbiome dans son ensemble est essentielle pour comprendre pleinement les interactions hôte-microbiome.

Nous présentons ainsi une analyse détaillée des sous-listes robustes de variables obtenues à partir de la méthode décrite à la section précédente, en utilisant un jeu de données concernant l'IBD (Maladie Inflammatoire Chronique de l'Intestin) comme exemple. Une recherche bibliographique approfondie des annotations robustes a révélé leur pertinence pour la maladie, cette sélection étant significativement enrichie en annotations directement liées à l'IBD selon la littérature.

De plus, cela met en évidence l'importance d'explorer les interconnexions entre les taxons et les FAs. En examinant les associations entre les taxons et les FAs robustes, nous démontrons l'existence d'un effet de cumul fonctionnel au sein des profils taxonomiques. En d'autres termes, plusieurs taxons non significatifs individuellement peuvent avoir un rôle significatif lorsqu'ils sont regroupés sur le plan fonctionnel, ce qui met en évidence la nécessité d'étudier le microbiome intestinal à ces deux niveaux. Nous soulignons ainsi le potentiel de la méthode proposée pour découvrir des signatures métaboliques cumulatives qui pourraient ne pas être apparentes lors de l'examen des données taxonomiques seules.

Implémentation d'un pipeline informatique pour la mise à disposition publique de la méthode.

Enfin, nous présentons le logiciel SPARTA (Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes), un pipeline informatique développé pour automatiser le processus d'analyse du microbiome intestinal décrit dans les sections précédentes. Le pipeline intègre les étapes de calcul d'un profil fonctionnel à partir d'un tableau d'abondance taxonomique, d'exécution d'une classification itérative basée sur les RF et d'une sélection automatique des variables, et d'évaluation de la robustesse des

variables sélectionnées.

SPARTA est implémenté de façon modulable, en séparant notamment en deux commandes le profilage fonctionnel et la classification et sélection de variables. Le pipeline produit en sortie le profil fonctionnel calculé, les performances de classification et les ensembles robustes de variables, avec les interassociations taxons-fonctions explicitées. Le pipeline offre des options permettant aux utilisateurs de personnaliser les paramètres, tels que le nombre d'exécutions et d'itérations, ainsi que de choisir leurs propres jeux de test. L'utilisateur peut également faire varier l'algorithme d'apprentissage (RF ou Support Vector Machines (SVM)), la métrique d'importance des variables (Gini ou SHAP), ou encore proposer une adaptation des données d'entrée (profils fonctionnels issus d'autres outils, sélection de variables personnalisée ou application à un seul type de profil, par exemple). Nous démontrons les performances du pipeline sur les cohortes précédement utilisées comme benchmarks, en fournissant des informations détaillées sur les temps d'exécution et les considérations relatives à l'utilisation des ressources. La mise en œuvre de SPARTA vise à améliorer la reproductibilité et l'accessibilité de la méthode proposée, permettant à la communauté d'effectuer des analyses complètes du microbiome intestinal à l'échelle taxonomique comme fonctionnelle.

Conclusion et perspectives.

Les travaux de cette thèse apportent plusieurs contributions au domaine de l'analyse du microbiome intestinal. Nous proposons notamment des innovations autour du calcul de profils fonctionnels, rendant le processus plus léger et donc plus accessible, de l'approche analytique du microbiome, en présentant une méthode intégrant une sélection de variables automatisée et un clacul interne de robustesse, et de l'interprétation biologique des sélections obtenues, avec une validation exhaustive des signatures fonctionnelles ressorties et une mise en relation des échelles taxonomique et fonctionnelle permettant de mettre en valeur des signaux fonctionnels émergents.

Ces résultats restent toutefois préliminaires, et ouvrent plusieurs voies pour de futures recherches et améliorations. Un affinement de l'approche de sélection, une expansion par le Web Sémantique des sorties ou un travail sur l'intégration de métadonnées médicales supplémentaires seraient des exemples de pistes à suivre pour pousser plus loin la performance et la compréhension des sorties de l'approche présentée.

Software: Le logiciel développé dans le contexte de cette thèse est ouvert et disponible publiquement au dépôt suivant: `https://github.com/baptisteruiz/SPARTA`.

Publication: Une grande partie du travail présenté dans cette thèse a servi de base à un article publié dans *PLOS Computational Biology* [1]. L'article contient notamment la présentation des méthodes de profilage fonctionnel et de classification ainsi que leurs résultats, l'analyse biologique des listes robustes obtenues, et une description du pipeline SPARTA.

Introduction

The importance and perspectives opened by the human gut microbiota have been at the forefront of the discussion in the medical field in the past years, as a wide array of unsuspected impacts on host health have been derived from its composition. Initially, the human gut was considered a relatively passive conduit for digestion and nutrient absorption. However, advancements in molecular biology and sequencing technologies have revealed the gut microbiota as a complex and dynamic ecosystem, integral to various physiological processes. This paradigm shift has underscored the microbiota's profound influence on metabolic, immune, and neurological functions, positioning it as a key player in both health and disease.

The gut microbiota: a marker of host health, and a lever for treatment.

The intuition of the gut microbiota's larger impact dates back to the early 20th century, with the works of Elie Metchnikoff, who notably first postulated on the health benefits of lactobacilli, based on the yogurt consumption of healthy rural Bulgarian populations [2]. His postulate was that the gut flora was composed of different sorts of microbes, some beneficial to the host, and some detrimental, leading to the conclusion that in order to attain better health, one should strive to replace harmful microbes by useful ones.

Today, advances in this field have been made, and imbalance in the composition of the gut microbiota (also known as *dysbiosis*) has been correlated to a wide array of diseases. Expectedly, diet-related disorders have been found to be impacted by the gut microbiota. Obesity, for example, can be encouraged by an overabundance of species involved in the production of Short-Chain Fatty Acids (SCFAs), which provide the host with an over-abundance of energy [3, 4]. Similarly, the gut microbiota was found to have an impact on its host environment, and a dysbiosis could encourage local inflammation, favoring Inflammatory Bowel Disease (IBD) [5] or carcinogenesis in the context of a colorectal cancer [6]. The influence of the gut microbiota ranges beyond its host organ however, as

other types of cancer, such as prostate [7] and breast cancer [8], have also been correlated to dysbioses. Most surprisingly, the gut microbiota was also found to have a strong influence on the host's brain: the innervation of the gut through the vagus nerve links both organs together, creating a "gut-brain axis" through which the brain can be affected by the state of the gut [9]. This, coupled with the gut microbiota's role in synthesizing hormonal precursors, has led to the discovery of an influence of the gut microbiota's composition on psychological disorders, such as depression [10] and schizophrenia [11], as well as neurodegenerative diseases, such as Alzheimer [12] or Parkinson [13].

In line with Metchinkoff's conclusions, therapeutic strategies involving the implantation of probiotics, bacterial strains known to be beneficial to host health, have been developed. Comparable practices can be traced further back: in the late 1800's for example, surgeon William B. Coley pioneered a therapy for sarcomas (cancers that develop in the bones and connective tissue) involving the inoculation of *Streptococcus pyogenes* and *Serratia marcescens* [14]. Nowadays, with a better understanding of the bacterial species that affect host health, probiotic-based therapies have been applied to the gut microbiota. Through an adaptation of diet, the direct ingestion of known benevolent strains, or fecal transplants through which the microbiota of a healthy individual is transplanted into the gut of an unhealthy one, the gut microbiota can be modified to positively affect one's health status [15]. The efficiency of these approaches varies strongly from case to case however, proving that in more complex cases, a more targeted approach could be required.

To do this, understanding the biochemical pathways through which the gut microbiota influences the host organism is a prerequisite. From the study of bacteria correlated to diseases, several such pathways have been uncovered. Metabolites of interest synthesized by the gut microbiota include, among others: bile metabolites, which notably affect energy metabolism and cell signaling pathways, SCFAs which affect insulin secretion and body mass maintenance, or a wide array of vitamins which are involved in DNA replication and repair and enhance immune functions [16]. This knowledge has opened the way for treatments that directly compensate the deficiencies of the host, through direct intake of a medication, or by creating targeted probiotics through genetic modification [17].

The gut microbiota can therefore be leveraged as a marker of host health, but also as a vector for therapy. In order to maximize the efficiency of a treatment, personalized approaches, based on an individual's specific microbiota composition, could be used to determine the most efficient cures. The complexity of the human and microbial

metabolisms makes this information difficult to apprehend, however. As such, computational approaches capable of handling such data and deriving information and decisions from it could help progress in this domain. Machine Learning (ML) approaches would therefore be viable candidates to tackle this task.

Machine Learning applications in the health sector: a tool for medical progress.

ML has emerged as a transformative technology across various sectors, with its impact profoundly felt in the medical community. ML algorithms enable systems to learn from data, identify patterns, and make decisions with minimal human intervention. This capability has ushered in a new era of precision medicine, where patient care is increasingly informed by vast and complex datasets.

At its core, ML involves the development of algorithms that allow computers to learn from and make predictions or decisions based on data. This learning process is typically categorised into three types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training an algorithm on a labeled dataset, where the input-output pairs are known, to predict outcomes for new data. Unsupervised learning, on the other hand, deals with unlabeled data and aims to identify hidden patterns or intrinsic structures within the data. Reinforcement learning involves an agent that learns to make decisions by receiving rewards or penalties based on its actions in an environment [18]. The application of ML in healthcare leverages these methodologies to analyze medical data, predict disease outcomes, optimize treatment plans, and improve overall patient care. The medical community's adoption of ML is driven by the increasing availability of health data, advancements in computational power, and the need for personalized medicine [19].

One of the most significant applications of ML in medicine is in the diagnosis and prediction of diseases. ML algorithms can analyze medical images with high accuracy, assisting radiologists in detecting conditions such as cancer and neurological disorders [20]. For instance, convolutional neural networks, a type of deep learning algorithm, have been utilized to interpret radiological images, achieving diagnostic accuracies comparable to human experts [21]. Furthermore, predictive analytics using ML can identify patterns and risk factors in patient data, enabling early intervention and personalized treatment plans. For example, ML algorithms can be used to assist medical staff in prognosis or diagnosis

of diseases by analyzing Electronic Health Records (EHR) and lifestyle data [22]. Clinical Decision Support Systems (CDSS) enhanced by ML provide healthcare professionals with data-driven insights to improve decision-making processes. These systems integrate patient data, clinical guidelines, and ML algorithms to offer evidence-based recommendations. For instance, CDSS can alert physicians to potential drug interactions, suggest and support diagnoses, and propose treatment options based on the latest research and patient history [23].

Another area where ML has made substantial contributions is the shift towards personalized medicine. Personalized medicine aims to tailor medical treatment to the individual characteristics of each patient, moving away from the traditional one-size-fits-all approach. ML algorithms can analyze genetic, environmental, and lifestyle factors to identify optimal treatment strategies for individual patients [24]. A patient's EHR, immunophenotype or serum metabolites, for example, can be leveraged to classify the individual into a more specific subcategory of the disease associated to adapted responses, or can be directly used to predict response to a therapy. In this regard, the gut microbiota constitutes another source of information with a lot of potential, due to it being unique for every individual and being a strong indicator of the host's health status, as previously discussed.

Exploring the gut microbiota through Machine Learning: an untapped potential.

The recent advancements in both the understanding of the gut microbiota and in the development of Machine Learning approaches to process medical data converge towards the question: could ML be used to enhance the medical potential of gut microbiota data? This could involve using the gut microbiota as a basis for diagnosis, through the use of an automatic classifier. It could also imply expanding the interpretability of trained models, to provide further insights into the role of the gut microbiota in human health [25]. Such approaches could prove to be opportunities for the medical community to deepen their understanding of this complex question, and find a basis for novel and adaptable therapies based on the gut microbiota.

Such applications come with methodological challenges, however. The need to integrate data from several sources makes the problem complex, and increases the dimensions of the data. This causes issues with classification performance [26], and diminishes the potential for downstream interpretation, as the amount of information to handle becomes

overwhelming. As such, it is important to establish methods for integration of knowledge in ML that take account of classification performance and robustness, as well as interpretability. Variable selection can be leveraged to achieve this, being a known solution for classifier performance enhancement in these conditions [27] that would also alleviate the amount of information generated by the model. Issues related to robustness can also be tackled through methods such as repetition of the training process [28].

During the course of this thesis, we contributed advancements to this subject by developing and implementing a method for integrating the functional annotation of the gut microbiota into an automatic classification process and facilitating downstream interpretation of its results. The process takes as input taxonomic composition data, which can be built from 16S or whole genome sequencing, and links each component to its functional annotations through interrogation of the UniProt database. A functional profile of the gut microbiota is built from this basis. Both profiles, microbial and functional, are used to train Random Forest classifiers to discern unhealthy from control samples. Our method explores the classifiers' inherent variability by extending state-of-the-art methods in three dimensions: increased number of trained Random Forests, selection of important variables with an iterative process, repetition of full selection process from different seeds. This process shows that the translation of the microbiota into functional profiles gives non-significantly different performances when compared to microbial profiles on 5 of 6 datasets. Through repetition, it also outputs a robust subset of discriminant variables. These selections were shown to be more reliable than those obtained by a state-of-the-art method, and their contents were validated through a manual bibliographic research. The interconnections between selected taxa and functional annotations were also analyzed and revealed that important annotations emerge from the cumulated influence of non-selected taxa.

Chapter 1

State of the art

1.1 Understanding the gut microbiota by quantifying its contents.

1.1.1 Describing the gut microbiota: the limits of an ecosystemic description, and the potential of a metabolic network.

The gut microbiota has been the subject of many recent studies, as its influence on host health has been found to be much more important and complex than previously envisioned. When studying the gut microbiota, the taxonomic scale, which can be accessed from sequences at a lesser computational cost, has generally been favored. In recent years, however, some voices in the medical community have called for increased inclusion of the gut microbiota as a functional system in coming analyses [29]. Specifically, taxonomy-based approaches do not properly account for functional redundancies between species and, in turn, might fall short in identifying novel biochemical pathways that should be targeted by innovative therapies. This observation around taxonomic functional redundancy also raises the question of whether some important functions could be derived from a cumulative influence of several less detectable taxa, and therefore cannot be correlated to remarkably differentiating taxonomic units.

> "It is what microbes *can do*, not *who they are*, that is finally important for ecosystem functions." (Inkpen et al.) [30]

As such, the transition to the functional paradigm from taxonomic profiles and the exploration of the links between both levels of description are a central theme of this thesis. These questions raise stakes around the methods employed for the profiling of the gut microbiota.

1.1.2 Different approaches to sequencing the gut: whole genome or 16S.

Profiling the gut microbiota has been made possible by the development of sequencing methods over the years [31, 32]. From a sample, generally fecal, an extraction procedure can isolate the bacterial genetic material, of which the genetic sequences can then be explicitated through sequencing. This sequencing step can be applied directly to the entirety of the obtained genetic material, in a process called Shotgun Metagenomic Sequencing (MGS). Another approach is to focus on targeted sequences that are known to be a signal for phylogenetic affiliation. The 16S ribosomal rRNA molecule is known to have these properties, and can therefore be separated from the rest of the genetic material, then amplified to be sequenced by itself.

Each method has its advantages, with MGS being a more expansive characterization of the microbiota's genome and 16S being less demanding in terms of computational resources. Both are widely applied by clinicians, and can be used as basis to quantify the contents of the gut microbiota, on the taxonomic or functional level.

In summary

Modern sequencing technologies, when applied to stool samples, make it possible to sequence the gut microbiota's genetic material, either in its entirety (MGS), or with a focus on on the 16S rRNA sequences (16S). The former method is more thorough, but also more costly and computationally demanding. The latter is a lighter procedure, but does not achieve the same level of precision down the line.

1.1.3 From sequences to quantifications: methods for the profiling of the gut microbiota, for each sequencing approach.

1.1.3.1 The common approach: building taxonomic profiles.

The gut microbiota can be defined as an ecosystem consisting of several different microbes that are present in different abundances. Profiles on this level, identifying and quantifying microbial taxa, can be built from sequences using computerized tools adapted to each sequencing approach. For sequences obtained through MGS, the tool of reference is MetaPhlAn [33–35], whereas for 16S data several pipelines are commonly used, such as

QIIME2 [36], mothur [37], or FROGS [38].

MGS sequences can be directly translated into species, through interrogation of databases. In the case of MetaPhlAn's implementation[33–35], the tool interrogates a database of unique species-specific genetic markers, specifically built for it by means of the ChocoPhlan tool. Through sequence alignment using bowtie2 [39], these markers are identified within the input sequences. By quantifying the average robust coverage of each species' specific markers over the input sequences, the pipeline can give a measurement of each detected species' abundance.

Processing 16S sequences requires a clustering step beforehand, as a same organism can have several differing 16S rRNA genes. This clustering step can be done in direct reference to an external database, as implemented by QIIME2 [36] for example, which clusters 16S sequences together based on similarity to the contents of databases such as Greengenes [40] or SILVA [41] using VSEARCH [42].Another option is to rely on unsupervised models, as does the FROGS pipeline which clusters 16S sequences through the unsupervised classifier SWARM [43], before being affiliated to a taxonomy from one of the previously mentioned databases through the blastn+ tool [44], or the Ribosomal Database Project's naive Bayesian classifier [45]. Depending on the interrogated database, the specific clusters can be associated to taxa grouped as Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs).

The size and contents of the obtained clusters, crossed with the information from the referenced databases, allows these tools to give a measurement of the frequency, as well as representative sequences, of the recognized OTUs and ASVs.

Limitations have however been pointed out for both of these approaches, notably in relation to the quality of the sequence reference databases which can be prone to mistakes concerning taxonomic affiliation, sequence errors, or inaccuracies surrounding inclusion and exclusion criteria for example [46]. As such, it is usually recommended to refer to the most popular databases, which receive the most feedback and implement corrections with updates most frequently.

1.1.3.2 Linking taxa to functional annotations: building functional profiles.

The gut microbiota can also be defined as a context for diverse biochemical reactions, dictated by the microbes' metabolism and the host environment: the gut lumen, the contents of which are themselves dictated by the host organism's metabolism. Profiles at this scale can be built using tools such as HUMAnN [34, 47, 48] or PiCRUSt2 [49,

50], specifically adapted for MGS and 16S sequences respectively. Characteristics of these approaches are resumed in Table 1.1. All of these tools associate functional annotations (FAs) to taxa via the interrogation of internal or external databases, creating a link between the taxonomic and functional paradigms.

HUMAnN takes as input MGS sequences, on which it runs MetaPhlAn (see Section 1.1.3.1). Following this, the reads are mapped against an annotated pangenome database, built specifically for the sample through ChocoPhlan's interrogation of the NCBI [51] and UniRef's [52] resources about the recognized species, at the nucleotidic level, using bowtie2 [39]. Unmapped reads are directly looked up in UniRef databases (UniRef 50 or 90, depending on the user's input) using the DIAMOND search binary [53] by default. From this, an association is built between the strains detected in the sample and reference genes. This association is quantified by HUMAnN through a count of the maps between the input reads and the reference sequences, weighted accordingly to the quality of the mapping, and normalized by the alignable length of the reference sequence, resulting in an abundance measurement in Reads per Kilobase. The reference genes can then be annotated to functional hierarchies such as COGs [54], Pfams [55], the Kyoto Encyclopedia of Genes and Genomes' (KEGG) [56] Enzyme Commission (EC) numbers and Orthologs, or Gene Ontology (GO) terms [57, 58]. Quantification of said annotations can be obtained by summing the abundances of all reference sequences associated to an annotation. These abundances can be calculated at the level of a species, or of an entire sample. If annotated with KEGG, this information can then be used to reconstruct metabolic pathways, referencing MetaCyc [59, 60] for example. For this, the MinPath tool [61] is mobilized to recover a set of pathways that cover the extracted annotations parsimoniously. For each of these pathways, the top half most abundant annotations are retained, and the mean of their abundance scores is calculated to quantify the pathway's abundance.

The PiCRUSt2 pipeline's inputs are a table of taxonomic abundances (OTUs or ASVs), and the representative sequences of each OTU or ASV in question, obtained from 16S sequences as described in Section 1.1.3.1. The first step is to align the reference sequences, then place the obtained alignments into a reference tree of full 16S rRNA genes from the Integrated Microbial Genomes (IMG) [62] database, annotated from the KEGG [56] Ortholog and EC databases. Sequence alignment is performed using the HMMER model [63], which exploits profile hidden Markov models to find homologs between the reference sequences, and group them together. These alignments are then added to the reference tree, using the EPA-NG [64] or SEPP [65] tool. The resulting tree is exported to the newick

format. From this tree file, gene family abundances and marker genes for each taxonomic component are predicted. This is done through a hidden state prediction approach [66], whereby the ancestral traits on the tree are inferred from the values of the leaves before being propagated forward to predict or correct the values of the leaves. In this manner, we obtain a tree on which all OTUs or ASVs are assigned a normalized amount of gene families and marker genes, predicted from the values of the initial tree. Having predicted a metagenome for each taxonomic unit, the abundance of annotations per taxon and per sample is calculated, by multiplying the input taxonomic abundances by the amount of genes correlated to a given annotation in the associated metagenome. These annotation abundances can then be transcribed into pathway abundances through the same method as HUMAnN.

In summary

Contents of a sample can be characterized at the taxonomic level, by quantifying the microbial taxa recognized from the genetic material. This can be achieved using tools such as MetaPhlAn for MGS, or QIIME / FROGS for 16S for example. It is also possible to build a functional profile of the microbiota, by measuring the expression of biochemical pathways by the microbiota. This can be done using tools such as HUMAnN for MGS, or PiCRUSt2 for 16S.

1.1.4 Making functional profiling more accessible: a lighter process compatible with all sequencing methods.

1.1.4.1 The EsMeCaTa pipeline: a lightweight method to associate functional annotations to taxa regardless of the sequencing approach.

Working directly from sequenced reads has a cost in terms of computational resources, as the methods to process these inputs are complex and the sequences themselves can require an extensive amount of disk space to be stored. For example, the MGS sequences of the datasets presented in Table 2.1 and exploited in subsequent chapters, takes up between 303 GB and 1.5 TB, for a total amount of 5.1 TB of disk space required to host all six. Previous works have, however, made available taxonomic profilings of several datasets. These files make for a more comfortable entry point for building a functional representation of the microbiota, as they require less disk space (3.1 MB of disk space

for all of the datasets presented in Table 2.1), and spare the user the step of building a taxonomic representation from sequences, which is compulsory for each of the previously cited tools.

There isn't an existing tool for calculating a functional profile from a taxonomic profile alone however, as neither HuMANn nor PiCRUSt accept taxonomic profiles as a sole input. The EsMeCaTa pipeline [67] is an alternative method to associate FAs directly to a list of input species or OTUs, relying on their taxonomic description to query UniProt [68] and fetch all proteomes associated to the most precise recognized taxonomy. Downloaded proteomes are then clustered with the mmseq2 tool [69], and a meta-proteome including only proteins that are present in at least 50% of all species included in the input taxonomic unit by default. The kept proteins are then annotated, either through a second interrogation of UniProt, or through the eggnog-mapper tool [70, 71]. A list of proteomes annotated with GO terms [57, 58] and EC numbers [56] is given as output.

As such, though it does not directly calculate an abundance score for each annotation, EsMeCaTa could be a step toward a quality of life improvement in handling metagenomic datasets, by allowing a shift towards the direct handling of taxonomic abundance tables. It taking a taxonomic description as input also means that it is agnostic as to the sequencing method the data is taken from, unlike HuMANn and PiCRUSt which are respectively specialized for MGS and 16S data. This feature can also be exploited as a reference point for comparing functional profilings of 16S and MGS sequencings of a same sample, as unlike the respective approaches of PiCRUSt and HuMANn, it would ensure that the profiles are built using the same resources.

1.1.4.2 Contributing a method for the quantification of functional abundances from EsMeCaTa's associations.

In light of the stakes presented here, one of the contributions of this thesis is a novel method for building functional abundance profiles directly from taxonomic profilings by relying on EsMeCaTa, translating the interconnections between taxa and functions into a quantified measurement of the abundance of expression of these annotations.

The EsMeCaTa approach was favored because it focuses on associating annotations to a functional profile, which is the aspect of functional profiling at the heart of our manipulations. When compared to HUMAnN, it also involves a lesser cost in terms of storage and computation time, allowing us to effectuate benchmarks on all six MGS datasets presented in Table 2.1, whereas a functional profiling directly from raw sequences

was not possible for all datasets due to the limits of the resources at our disposal. A comparison with results obtained from a profiling of the IBD dataset with HUMAnN3 will however be made in Chapters 2 and 3. The explicit linking of taxa and annotations will also be an important for the uncovering of dynamics of functional cumulation between taxa, notably explored in Chapter 4.

In summary

Functional profiling using the classic approaches has downsides, notably related to the size of the input they require and their specialization towards either MGS or 16S sequencing. EsMeCaTa circumvents these issues by taking taxonomic affiliations as input, therefore being compatible with both sequencing approaches. It does not quantify annotation expression from this basis however; as such, a method to operate this quantification was put together for this thesis.

Method	Technology		Compulsory inputs	Required pre-treatment	Consulted databases	Sequence alignment	Output	
	MGS	16S					Annotations per taxon	Functional abundances
HUMAnN	x		Raw sequences		UniRef, NCBI	x	x	x
PiCRUSt2		x	Taxonomic profile, representative sequences	x	IMG		x	x
EsMeCaTa (UniProt)	x	x	Taxonomic affiliations	x	UniProt		x	
EsMeCaTa (eggNOG)	x	x	Taxonomic affiliations	x	UniProt, eggNOG	x	x	

Table 1.1 – **Comparison of the existing tools for functional profiling of the gut microbiota.** HUMAnN and PiCRUSt2 can quantify the functional annotations linked to a profile, whereas EsMeCaTa can only list them. The latter tool is the only one that can work directly from a taxonomic profile, and can be applied to 16S or MGS data indiscriminately. HUMAnN can only work from MGS raw sequences, and PiCRUSt2 requires both taxonomic abundances and representative sequences, obtainable through a pre-processing of 16S sequences by a tool such as FROGS or QIIME (see Section 1.1.3.1)

1.2 Classifying the gut microbiota with Machine Learning: promises and levers for improvement.

1.2.1 The gut microbiota's potential as input for prediction.

The development of Machine Learning (ML) methods, referring to methods that learn automatically from data, and notably supervised classification approaches, has opened up many possibilities in how to approach the question of using microbiota data to predict a host status. A first application of ML to sequenced microbiota data came in a 2011 study by Knights et al. [72], which provided an exploration of the applicability and efficiency of such approaches when applied to the composition of the microbiota. This work showed that supervised classification of the microbiota was possible using several already well-known approaches. Classic ML-driven feature selection methods, notably elastic net [73], are also shown to be applicable. The tests consisted in five benchmarks on taxonomic abundance datasets from Costello et al. (2009) [74] and Fierer et al. (2010) [75]. From this, the study established benchmarks on classification tasks predicting the body habitat (ear, gut, hair, nose, mouth or skin), skin site (forearm, foot, forehead, palm...), and identity of the subject that the samples were taken from.

This work was built on by Statnikov et al. in 2013 [27], which conducted a thorough evaluation of classification methods on the Knights et al. datasets, but also on other datasets exploiting microbiota from the skin (Alekseyenko et al., 2013 [76]) and diverse regions of the digestive track (Nossa et al., 2010 [77]) to predict body sites and diagnose Psoriasis and Esophagitis. The tested classifying methods were Support Vector Machines (SVM) [78], Kernel Ridge Regression [79], Regularized Logistic Regression, Bayesian Logistic Regression, Random Forests (RF) [80], K-nearest Neighbors [81] and Probabilistic Neural Networks [82]. Each method's parameters were selected by a nested cross-validation.

Overall, both studies converge however in highlighting RFs, which most consistently yielded top results in both articles' benchmarks, singling itself out as an especially efficient method for the classification of microbiota data.

1.2.2 Random Forests: a classifier that adapts well to microbiota data.

1.2.2.1 Principle of the algorithm.

RFs are an ensemble learning method, consisting of a collection of decision trees trained individually on a subset of the input data and making collegial classification decisions.

Decision trees are classifiers that hierarchically split the initial dataset based on successive splitting rules applied to the data's features, namely comparing a feature's value to a learned threshold. This results in a tree-like structure, the leaves of which are assigned a label inherited from the input data. A tree will classify new data by applying the rules inherent to its successive nodes, until the data is assigned to a subset defined by a leaf.

When training a decision tree, each node's splitting criterion is learned through the Gini Impurity metric [83], which evaluates the probability of misclassifying a sample chosen at random based on a given criterion. Over all variables from the node's subset, the threshold criterion that minimizes this index is retained to split the dataset. After training, this metric can then be used to measure how influential a variable is within a tree's decision path, by summing the impurity decreases generated by each node where said variable is leveraged (Gini importance [80, 84]). This variable importance can also be measured through other metrics, such as SHAP importance [85] which calculates each variable's contribution to a decision from the basis of a trained classifier.

RFs are generated by training an ensemble of decision trees, each on a representative subset of the original data [80]. Once all trees are trained, they will classify new data by popular vote.

1.2.2.2 Insights on Random Forests' performance and robustness.

A classifier's performance can be measured through different metrics. One of the most commonly used is the area under the Receiving Operator Characteristic curve (ROC AUC), which plots the true positive rate as a function of false positive rate, at increasing classification thresholds [86]. This metric notably has the advantage of being less influenced by the proportions of each class within the test dataset compared, notably, to model accuracy.

Beyond its performance, a model must also be evaluated on its robustness, meaning its ability to maintain its classification performance in varying contexts. RFs are reputed to be overall robust models, notably due to their resistance to overfitting [80]. This characteristic

can be estimated by repeating the training and testing process on different subsets of the input dataset.

In summary

Applications of ML approaches to microbiota data have highlighted Random Forests (RF) as the best adapted models for this type of data. RFs are a robust model, with good potential for interpretability thanks to their intrinsic metric for variable importance (Gini importance score).

1.2.3 The methodological stakes of gut microbiota classification with Random Forests.

With RF models proving to be the most effective approach to apply to gut microbiota data, several studies have explored its specific usage as a predictor of host health from the composition of the gut microbiota. These approaches are resumed in Table 1.2. These works each have their own strengths and limitations, highlighting the potential of the current approaches as well as giving insights on how to improve classification on gut microbiota.

1.2.3.1 Repetition and resampling for enhanced robustness.

The prediction of host health on the basis of the gut microbiota's composition, which is at the heart of this thesis' subject, was first notably explored by Pasolli et al. (2016) [87].

This paper tackled the question of using the gut microbiota as a predictor for diseases. As such, they developed a prediction tool, MetAML, which automatizes the training of predictive models to differentiate different categories of individuals based on the composition of their gut microbiota. Their method was tested on several datasets: Colorectal [88], Cirrhosis [89], Obesity [90], Inflammatory Bowel Disease (IBD) [91], and Type 2 Diabetes in the context of Chinese (T2D) [92] and European (WT2D) [93] cohorts. Benchmarks were conducted to discern an individual's health status (healthy control, or unhealthy) on the basis of these cohorts' taxonomic abundance profiles, but also of the presence of strain-specific microbial features, both generated using the MetaPhlan2 tool [33]. RF models were trained with the following fixed parameters: 500 trees, a number of features

to consider when splitting the training data set to the root of the number of features in the dataset, and the quality of splits was measured using the Gini criterion. Classification performances were evaluated over 20 independent runs, with 10-fold cross-validation with no external test data subset. This approach has been criticized due to the risk of a data leak, meaning that measured performances could overestimate the models' true capacities due to being interrogated on data subsets that have played a part in training it.

This limitation has notably been identified in a study by Oh & Zhang in 2020 [94], which introduced another tool for classification on the basis of the gut microbiota named DeepMicro. Aiming to correct the previously mentioned faults with MetAML's training method, DeepMicro's performance results were measured on a separated test set, and optimizes the parameters of the models through a grid search [95] involving 5-fold cross-validation processes. This approach corrects the risk of a data leak, and improves overall performance through parameter optimization. However, these additions combined with the Representation Learning aspect of the approach made the training process more costly in time and resources, leading to models being evaluated over over 5 runs compared to MetAML's 20. The lesser amount of repetitions of the evaluation process reduces the robustness of DeepMicro's results when compared to its predecessor.

These approaches illustrate a first level on which gut microbiota classification can be improved: the robustness of the results. This can be tackled by augmenting the amount of repetitions of the training process [28], with a resampling involving sanctuarized test datasets to avoid data leaks.

1.2.3.2 Reducing the dimensions of the data for better performances.

A characteristic of the gut microbiota that makes it difficult to use for classification are its dimensions. Indeed, datasets usually have few samples, each of which contains a lot of information. This characteristic impedes automated classification performances, as per the Hughes phenomenon, also known as curse of dimensionality [26], which postulates that there is an optimal amount of features for a set amount of observations, and that augmenting data dimensionality past this point decreases classifier performance.

DeepMicro is notable for circumventing the curse of dimensionality by applying projection methods to the input data in order to enhance performance. The tool added an element of Representation Learning, prefacing the ML training with a reduction of the data's dimension through diverse methods such as Principal Component Analysis (PCA) [96] or autoencoders [97]. When tested on the same abundance profiles as MetAML, this

transformation did not yield significant improvement in classification. This dimensionality reduction proved to be very efficient however when applied to profilings based on the presence or absence of strain markers, which contained 200 times as much information per patient as their taxonomic counterparts. In this case, AI-based dimensionality reduction consistently improved classification performances, whereas classical PCA and Gaussian Random Projection based representations decreased performances in a majority of cases.

In MetAML's case, it was found that the taxonomic profiles were best discerned by RFs when coupled with a feature selection. The selection in question was performed on the basis of the feature ranking performed following the RF's classification (see Section 1.2.2.1), where the top k features are selected and used as basis for re-training a RF model, for k in 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200. The top performing set is retained as the optimal number.

As such, correcting the dimension of the data is another major stake when it comes to improving classification performance. While DeepMicro's Representation Learning-based approach to this issue has shown a lot of promise, the solution of variable selection has the added advantage of keeping the data's readability, boosting the results' interpretability.

1.2.3.3 Expanding on interpretability: the advantages of the functional paradigm.

Classification based on Random Forests opens possibilities when it comes to understanding the impact of each variable on model performance, through its capacity to generate a ranking of variable importances (see Section 1.2.2.1). In the case of MetAML for example, after training, the authors include a discussion around model interpretability, notably illustrating and commenting the top 25 species by Gini importance score averaged over 20 runs for the Cirrhosis and Colorectal datasets. The biological coherence of certain top features from the Cirrhosis dataset is also explored, highlighting known pathogens of this disease such as *Veillonelle* and *Streptococcus* strains within this list. Some taxonomic units are also identified as generic markers of an unhealthy gut microbiota. The interpretation stops shy of covering the biological significance of all highlighted species however. The fact that it conducts an evaluation on the taxonomic scale means that it also does not align with the more recent demands of the medical community, which sees more potential in the exploration of the gut microbiota at the functional scale (see Section 1.1.1).

In studying health status classification from functional profilings of the gut microbiota, two studies stand out. First to mention is Jones et al.'s publication in 2020 [98]. This study explored RF classification of Crohn's disease severity using both 16S and MGS sequences,

the former converted to taxonomic profiles with QIIME2 [36], and the latter functionally profiled with HUMAnN2 [48]. The results showed that the functional profiles performed worse than their taxonomic counterparts when it came to classification, but remained significant classifiers. This was followed by an interpretation of the top 30 pathways and taxa by feature importance, which established their coherence notably with previously identified taxa, such as *Ruminococcus gnavus*.

The earlier works realized in 2018 by Douglas et al. [99] are perhaps more closely related to this thesis' subject. This study once again focused on Crohn's disease, and once again retained the RF classifier, but this time aimed to predict disease state and treatment response via RF classifiers. An originality of this approach was that it exploited and compared both 16S and MGS profiles, with integration of the taxa's metabolic functions using PICRUSt [49] for the former, and HUMAnN2 [48] for the latter. The overall conclusions, for disease state classification, were that for both methods, functional profiles could significantly classify patients, but couldn't match the best performances obtained by taxonomic information. This classification was further exploited by extracting the variables' importance scores, in terms of mean decrease in accuracy, in order to rank them by decreasing importance. This allowed to highlight important taxa, like *Akkermansia muciniphila*, and pathways, such as the biosynthesis of amino-acids. A joint analysis of the top 3 features from each profile showed that 16S OTUs were the most informative class of features.

These studies showed the limitations of exploiting FAs to augment microbiota datasets in the context of ML, as the translation negatively impacts classification performance in both cases. This can however be explained by the augmentation in data dimensionality that is induced by this shift. For example, in Douglas et al.'s study, the 16S functional profile contained 200 times as many variables as the taxonomic abundance table. Previously established dimensionality reduction techniques could be a viable response to this issue. It should also be noted that these results were obtained on a single instance of model training with a fixed amount of 1001 trees and other parameters left at default, using Leave One Out Cross-Validation (LOOCV) [100, 101], meaning that performance and robustness were not the primary focus of these works.

These studies were however efficient in highlighting an advantage of this new representation: the subsequent analyses based on RF feature importance scores, taken from the best classifiers, singled out impactful metabolic functions, which is an important gain in terms of the models' interpretability. These results are in line with demands raised

by the voices in the medical community mentioned in Section 1.1.1 [29], in that they provide insights on the more directly exploitable metabolic scale. The issue of functional redundancy, however, remains unexplored here. These observations combined lead us to believe that by pushing Douglas et al. and Jones et al.'s interpretations further, with rigorous model fitting and feature selection and a heightened awareness of the correlations between the highlighted taxa and functions, we could expect to uncover some biologically relevant features from ranking and filtering markers based on how informative they are.

In summary

Applications of RF models to various profilings of gut microbiota samples have established benchmarks for classification, and highlighted levers for improvement in terms of robustness through repetition, classification performance through variable selection, and interpretability through the exploration of top features, notably functional.

1.2.4 Contributing a novel approach: enhanced robustness with an expansion of state of the art methods in three dimensions, and integration of the taxonomic and functional scales.

In light of these observations, we developed an approach during the course of this thesis that incorporates the strengths of all of the previously mentioned approaches. This method, called Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes (SPARTA), applies RF-based approaches to both taxonomic and functional profiles. The method is introduced in Chapter 3, and an implementation in the form of a computational pipeline is described in Chapter 5.

The models are trained over several runs (10 by default) with dedicated test subsets, each involving 20 instances of model training with separate validation subsets. This is done to avoid issues related to potential data leaks, in the same vein as DeepMicro, while striving for better robustness through an amount of repetitions of the training method more comparable to MetAML's. Inspiration was also taken from MetAML's approach to performance enhancement in the form of a variable selection based on variable rankings and iterated to reach peak classification performance, and which can then be expanded

into biological interpretation. As such, this approach enhances robustness and performance by extending the MetAML and DeepMicro [87, 94] procedures in three dimensions, with an increased number of trained RFs, selection of important variables with an iterative process, and repetition of the full process from different seeds, to allow exploration of the inherent variability in performances due to changes in training hyperparameters.

Obtaining a set of discriminant functions is one of the major aspects when turning to FAs instead of taxa descriptors. As such, the interpretability aspect is also pushed further, through a thorough exploration of both taxonomic and functional signatures, in the vein of Douglas et al. [99] and Jones et al.'s [98] works, but also taking in account the inter-correlations between both profiles so as to highlight variables issued from a cumulation effect. These results are presented in Chapter 4.

In summary

This thesis contributed a method for ML-driven analysis of microbiota data with enhanced focus on robustness, achieved through an extension of the method in three dimensions: increased number of trained classifiers, selection of important variables with an iterative process, and repetition of the process several times. This new method also integrates both the functional and taxonomic paradigms to the analysis. This allows for a more expansive downstream biological interpretation, which takes account of the impact of both paradigms as well as their intercorrelations.

1.3 Leveraging Machine Learning for variable selection: performance and interpretability.

In dealing with the biological interpretability of a classifier's results, the capacity to evaluate a variable's importance and operate a selection of significant information based on this criterion facilitate the process immensely. In the case of microbiota data, this is especially true for functional profiles, the dimensions of which are not easily tractable. This was previously illustrated by the data used by Douglas et al. [99] (see Section 1.2.3.3), but also in the presentation of the HUMAnN3 tool [34], which was applied to an illustrative example consisting of a meta-analysis of Colorectal Cancer cohorts containing 121 species, from which the tool derived 2,895 EC numbers [56]. This approach has also improved the

	Studied profile	Datasets	Tested models	Performance enhancement approach	Robustness	Interpretability
MetAML [87]	**Taxonomic** (MGS)	Various studies: Cirrhosis [89], Colorectal [88], Obesity [90], IBD [91], T2D [92], WT2D [93]	**RF**, SVM	Variable selection: top k features ranked by Gini importance metric.	20 independent runs, with 10-fold cross-validation. Issue: absence of an external test set, so parameters may be selected based on a model which has seen the test set.	The top 25 species by Gini importance score are given and commented: the biological significance of certain species in the context of the Cirrhosis dataset notably, and the signaling of certain taxa as general markers of an unhealthy microbiome.
DeepMicro [94]	**Taxonomic** (MGS)	Various studies: Cirrhosis [89], Colorectal [88], Obesity [90], IBD [91], T2D [92], WT2D [93]	**RF**, SVM, MLP	Representation Learning: projecting the data (autoencoders, PCA, Gaussian). Parameter optimization through Grid Search.	5-fold cross-validation on a training set, then evaluated on a separated test set over 5 runs. This corrects the issue of a potential data leak, but is arguably insufficiently robust.	The study is fully performance-focused, and does not comment on variable importance or biological significance of the results.
Douglas et al. [99]	**Taxonomic** (16S, MGS), **Functional** (16S, MGS)	Crohn cohort (BISCUIT cohort [102])	**RF**	-	LOOCV evaluation.	**T**he top 3 ranking taxa and pathways are singled out and commented by the paper, including in regard to their biological significance. Taxonomic and functional profiles are also compared to one another in terms of significance, by commenting on the variable importances of models trained on a combination of both profiles.
Jones et al. [98]	**Taxonomic** (16S), **Functional** (MGS)	Crohn cohort (MAREEN [98])	**RF**	-	LOOCV evaluation.	The top 30 features of the top performing trained models are given, and the significance of some of the top taxa is briefly discussed.
SPARTA (Work from this thesis)	**Taxonomic** (MGS), **Functional** (MGS), **Taxonomic and functional associations**	Various studies: Cirrhosis [89], Colorectal [88], Obesity [90], IBD [91], T2D [92], WT2D [93]	**RF**	Variable selection: automated iterative selection based on features' Gini importance scores (by default). Parameter optimization through GridSearch.	10 independent runs, each with a dedicated test subset. Each run consists of 20 instances of model training with 5-fold cross-validation, each evaluated on separate validation sets.	The robustness of the selected variables is evaluated over 10 runs. The most robust variables were bibliographically validated in full on an example. The interconnections between robust taxa and functional annotations are explored and discussed.

Table 1.2 – **Comparison of notable applications of RF classification to gut microbiota data.** MetAML and DeepMicro explored different ML applications to predict patient health state based on the taxonomic composition of the gut microbiota, with a primary focus on performance. Douglas et al. and Jones et al. put more focus on interpretability, by exploring the functional scale in parallel. The SPARTA pipeline, which was developed in the context of this thesis, integrated the strengths of the cited approaches, in terms of performance, robustness and intepretability.

performances of classifiers dealing with high-dimension data [27, 87]. Several approaches can be leveraged to evaluate variable importance and perform selection (see Table 1.3).

1.3.1 Linear approaches to variable selection.

When faced with the problem of identifying markers of a person's health state using the microbiome, one's first instinct would be to statistically compare the composition of microbiotas sampled from healthy and unhealthy individuals, and highlight markers that are differentially expressed between both profiles, to be used as reference for future diagnostics. This approach, akin to a linear regression, is usually applied by default to datasets [88–90, 92, 93] and has permitted the identification of several disease markers, both taxonomic and functional. By selecting the variables with the most significant differ-

ential expression, this approach can also be used for variable selection. This can be done, for example, as implemented for the DESeq2 [103] tool, which is best adapted for RNA-seq data, or limma [104], which has broader applicability. These tools fit a Generalized Linear Model over the data before testing whether each variable's regression coefficient is significantly different from zero through a Wald test applied to shrunken logarithmic fold change estimates. The p-values of the test in question, after adjustment by the Benjamini-Hochberg method for multiple testing, can be used as basis for selection [105–107].

1.3.2 Random Forests: a basis for variable selection in non-linear problems.

RF models, notably through the calculation of variable importance scores (Gini or SHAP for example, see Section 1.2.2.1), are also a potent tool for evaluating variable importance and operating selection. This approach differs from DESeq2 or limma's in the sense that profiles are evaluated in a non-linear fashion, which allows RF importance rankings to take account of more complex distinctive criteria. In terms of classification performance, they also notably outperform the classic regression-based approaches [27], which limma is based on, marking it as a comparatively innovative approach. RFs are also becoming more prevalent for variable selection as well, notably in the domain of bioinformatics [108].

Several approaches exist to perform variable selection based on RF importance rankings. This has been explored in the previously cited MetAML approach [87], which searches for the optimal top-k features that maximize classification performance, but also by Statnikov et al. [27], where RF models were found to be more precise when coupled to a variable selection based on RF-based backward elimination procedure (RFVS), as described by Svetnik et al. [109], wherein a set fraction of the dataset's variables, chosen at the bottom of the Gini Importance Score's ranking, is iteratively removed until the model reaches peak performance. In Statnikov et al.'s case, 20% of the variables would be eliminated at each iteration. This approach, coupled with MetAML's variable selection (see 1.2.3.2) illustrate the possibilities offered by Gini importance rankings when it comes to selecting variables. Its capacities in comparison to the classic statistical approaches are seldom mentioned however. As a basis for variable selection, RF-based approaches have also rarely been directly compared to linear approaches such as limma [104].

In summary

Variable selection is a potent and effective approach to enhance ML models' performance and interpretability when applied to the gut microbiota. This issue can be tackled linearly, through approaches such as DESeq2 or limma. It can also be approached through the prism of RF models' variable importance rankings.

1.3.3 Human bias, robustness and exploitability: challenges surrounding variable selection.

One could however regret that the previously presented variable selection methods require to choose discrete parameters: limma [104] requires a p-value threshold to be defined, RFVS iteratively selects a predetermined percentage of the dataset, and MetAML's approach covers an empirically chosen set of top k values. A fully automated selection process could be preferable, as it would remove user-induced bias altogether.

Another aspect of these selection processes to take into account is that of the variable selection methods' robustness, which can be measured through the coherence of repeated selection tasks for example. On this aspect, DESeq2 and limma have already been evaluated as having good reliability [110, 111], though an application to RNA-sequencing data [112] has also shown that non-parametric variable selection methods could prove more robust. RF models have also been proven to be coherent in the right conditions, but their robustness is also highly dependent on the data and chosen approach [113]. As such, an internal measurement of the RF selections' robustness should be envisaged to add transparency if we are to exploit these selections for downstream biological interpretation. This aspect of the method is evaluated by neither Statnikov et al. [27] or MetAML [87].

Finally, improving the exploitability of a selection of variables is also important, to make their downstream handling by biological experts more comfortable. This can be achieved through visualization. In the case of functional annotations, tools such as REVIGO for GO terms [114] or KEGG-mapper for EC numbers [115] provide options to represent shortlists of annotations. However, it is much more challenging to concurrently represent different ontologies. The measurement of the selections' robustness could provide advances in this department as well, as they would give the user the possibility to prioritize the most reliable elements of the shortlist.

1.3.4 Contributing a fully automated Random Forest-based selection approach for variable selection and associated robustness measurement.

A novel approach that covers these issues is integrated to the SPARTA pipeline, presented in Chapter 3, which we also implement in Chapter 5. This method fully automates the selection process by relying on a cutoff at the inflection point of the curve of decreasing variable importances, rather than selecting a set amount of features. It also repeats the training and selection process to establish a measurement of the robustness of the selected variables, while keeping account of the evolution of classification performance. Our approach makes use of the RF's capacity to handle non-linear problems, which is an advantage over DESeq2 and limma. We will however compare results obtained with our method with those of limma during the course of our manipulations. By default, the process will be based on Gini importance scores, though benchmarks made with SHAP as a basis for variable importance will also be established. The method's accent on robustness also facilitates downstream biological exploration. Being applicable to interlinked taxa and annotations, the output's format, when represented as a table or as a bipartite graph, also facilitates the detection of cumulation effects within the microbial community, as illustrated in Chapter 4.

In summary

During this thesis, we contributed a variable selection method based on RF variable importance, which can handle non-linear distinctive criteria. The novelty of this approach is that it is fully automated, as it does not require any discrete parameters to be specified by the user. It also provides an internal evaluation of the selection's robustness, based on a repetition of the selection process.

1.4 Conclusion

The importance and perspectives opened by the human gut microbiota have been at the forefront of the discussion in the medical field in the past years, as a wide array of unsuspected impacts on host health have been derived from its composition. When

studying the gut microbiota, the taxonomic scale has generally been favored, to identify biomarkers for various conditions [3, 5, 6]. In recent years, however, some voices in the medical community have called for increased inclusion of the gut microbiota's functional paradigm in coming analyses. Specifically, taxonomy-based approaches do not properly account for functional redundancies between species and, in turn, might fall short in identifying novel biochemical pathways that should be targeted by innovative therapies [29].

Functional profilings can be built with several approaches, depending on the upstream sequencing method. For raw MGS reads, various tools have been developed for functional analysis, notably including the HUMAnN pipeline [34, 47, 48] which can quantify FAs in a sample based on sequence alignments. For processed 16S sequencing data, PiCRUSt2 [49, 50] stands as one of the most popular tools for functional profiling. Other tools can be agnostic in regard to the sequencing method, such as the EsMeCaTa pipeline [67], which functionally annotates an input list of taxonomic affiliations according to the content of the UniProt database. All of these tools associate FAs to taxa via the interrogation of internal or external databases, creating a link between the taxonomic and functional paradigms.

The resulting functional profiles constitute a basis for uncovering functional markers within the gut microbiota, provided these markers can be ranked or filtered based on how informative they are. Such a ranking can be handled through a linear approach, for example using the DESeq2 [103] or limma tools [104], which fit a Generalized Linear Model over the data before testing whether each variable's regression coefficient is significantly different from zero [105–107]. Previous studies in clinical predictive modeling have also highlighted the potential for tree-based methods to perform such a variable selection, such as RFs [116] thanks to their inherent aptitude for variable ranking through the Gini feature importance metric [80]. RFs are also particularly relevant in this regard, due to their proven efficiency in classifying microbiota data [27], outperforming other classic techniques, such as SVMs [27, 87, 94].

In terms of performance, the biggest hurdle to be cleared pertains to the microbiota data's dimensions, which do not favor classifier training. Many previously cited approaches (see Section 1.3) have explored methods to diminish the input data's dimensions, with variable selection standing out as one of the most effective. This approach is compatible with the RF's intrinsic evaluation of variable importance, however none of the cited studies have tried a fully automated selection process based on this metric.

The works cited in 1.2.3 also put a differing amount of focus on the robustness of their results, with varying approaches to performance measurement and different amounts of repetitions of the training process. This concept of robustness can also be extrapolated to the notion of variable importance and selection, where little has been done in the cited works to allow for transparency around the selectors' internal coherence. It would be important to define a framework to evaluate the robustness of a feature's importance, so as to have a stronger basis for interpretation of these results.

In terms of interpretability finally, obtaining a set of discriminant functions is one of the major aspects when turning to FAs instead of taxonomic descriptors. Though some studies developed the biological implications of their outputs, these interpretations remained restrained to a subset of important variables, and no exhaustive examination of the obtained lists of important variables' biological pertinence was made. These observations are also generally restrained to the taxonomic scale, though the works of Douglas et al. and Jones et al. offer a first glimpse into the potential of applying this protocol to functional profiles. While the shift to functional profiles leads to a decrease in classification performance, the subsequent analyses based on RF feature importance scores singled out impactful metabolic functions. However, the usual number of FAs identified in biological samples (2895 ECs derived from 121 species with HUMAnN3 in context of a meta-analysis of Colorectal Cancer cohorts for example [34]) is not easily tractable. The potential implications surrounding the links between taxa and their expressed biological functions also remain underexplored.

During the course of this thesis, we will put together an approach that makes it possible to exploit the RF as an automated variable selector to improve its performances, but also to internally evaluate a variable's robustness as a predictor, for better interpretability of the model. To achieve that goal, this approach extends the MetAML and DeepMicro [87, 94] procedures in three dimensions (increased number of trained RFs, selection of important variables with an iterative process, repetition of full selection process from different seeds) to ensure full reproducibility and exploration of inherent variability in performances due to changes in training hyperparameters. We will also show that cumulative phenomena can be identified by leveraging the relationships between taxa and their expressed FAs. Finally, we will show that this analysis can be done even without access to raw sequence data. These contributions are made through the implementation of the Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes (SPARTA) pipeline.

In a first chapter, we will describe how this approach will integrate a novel method for building a functional profiling of the gut microbiota directly from an input taxonomic abundance table, based on the EsMeCaTa pipeline and therefore compatible with all upstream sequencing methods. Through a comparison of the outputs with those of HUMAnN3, we will show that the information gathered by this new approach is more thorough, while covering the majority of the annotations gathered by HUMAnN. This new method will allow us to work on six MGS databases, making the process sufficiently light to be compatible with our computing resources through its exploitation of taxonomic tables as input instead of raw sequences. This approach's compatibility with MGS and 16S profiles alike make it a more versatile approach than other classic tools for functional profiling.

In a second chapter, we will introduce a method to conduct a RF-based classification and analysis of the microbiota data, both on the taxonomic and functional scale, while accounting for the interconnections between taxa and functions. This method will be based on the re-training of multiple classifiers, iterated in the context of a succession of classification and variable selection applications, itself repeated several times for better robustness of the results. Through a post-processing method designed to accentuate genericity and robustness, it will also output a curated list of important variables, selected through a fully automated and non-parametric RF-based approach, alongside an internal metric for evaluation of a variable's robustness in terms of its importance based on repetition of the selection process. This will be done while ensuring consistent classification performances when switching from taxa to FAs as a basis for classification. These operations will be tested on six datasets, used for benchmarking by several previous studies in this domain. The approach's performance in terms of classification will be compared with a reproduction of DeepMicro's results, and its robustness as a variable selector will be compared to limma's.

In a third chapter, the notable functional variables identified in this manner will be bibliographically explored in detail on an example, and discussed in regard of their relationship with notable taxa. This will allow us to confirm the biological viability of our selection, but also show how the visualization of the interconnections between taxa and annotations can highlight functional variables that gain in importance through the cumulated influence of non-discriminant taxonomic counterparts.

Finally, we will describe a publicly available implementation of the SPARTA pipeline, making the approach developed during the course of this thesis reproducible and accessible

for future applications to other data.

Publication

The majority of the work presented in this thesis is also the basis of a scientific article, published at *PLOS Computational Biology* [1]. The article notably covers the presentation of the method and results presented in Chapter 3, the functional profiling approach described in Chapter 2, the biological interpretations of Chapter 4 and a presentation of the implementation described in Chapter 5.

Software

The software developed in the context of this thesis is open-source, and available on the GitHub website. This notably concerns the implementation of the SPARTA pipeline, described in Chapter 5, which can be found with accompanying README instructions, a testing suite and material for result reproduction at the repository hosted at `https://github.com/baptisteruiz/SPARTA`.

		Robustness	Interpretability	Parameterization
DESeq2 [103] / limma [104]		Usually applied to RNA-seq data, DESeq2 and limma have been evaluated as an approach with good within-method consistency and robustness [110, 111]. However, they have also been found to be less robust than non-parametric approaches by some studies [112].	DESeq2 and limma give a ranking and an evaluation (adjusted p-value) of each variable's importance in linearly differentiating profiles.	Requires a human input for variable selection (p-value threshold)
RF	RFVS (i.e: Statnikov et al. [27])	This aspect of the approach is not evaluated in the referred articles. Self-consistency of Random-Forest-based variable selection is shown by Kursa et al. to be highly dependent on the data and on the selection approach, but is capable of having high retention rates in the right circumstances [113].	Variable importance scores (usually Gini) give a ranking of each variable's non-linear importance in the trained model's reasoning.	Requires a human input for variable selection (fraction of the data to be dropped iteratively)
	Optimal top k features (i.e: MetAML [87])			Requires a human input for variable selection (list of empirical top feature thresholds to be tested)
	Iterative automatic threshold (i.e: SPARTA)	The self-consistency of the method is evaluated internally, and is an area of focus		Fully automated variable selection

Table 1.3 – **Comparison of notable approaches for variable ranking and selection in the context of the gut microbiota composition.** The question of ranking variables by importance can be approached with linear methods, as implemented with limma and DESeq2, or non-linearly through RF variable importance metrics. Several approaches can then be applied to use these rankings for variable selection, with the method developed for SPARTA in the context of this thesis being the only one to include full automation.

Chapter 2

A novel method for computing functional profiles.

The gut microbiota can be described on several different levels. On a first basis, it can be defined as a community of micro-organisms that populate the gut. As such, it is most commonly described on the taxonomic scale, with measurements of the abundances of the different taxonomic units that constitute the population. This sort of description has been made available with technological advancements in gene sequencing [31, 32], and in the tools for subsequent *in silico* analysis of the data, such as the QIIME [36], mothur [37], and FROGS [38] pipelines for 16S rRNA sequences, or MetaPhlAn [33–35] for Shotgun Metagenomic Sequencing (MGS) data. These taxonomic profiles have also been widely explored as a basis for supervised classification of individuals [27, 72, 87, 94].

Beyond this conception of the gut microbiota as an ecological community, one could also view it from a mechanistic point of view: as a stage for exchange, consumption and transformation of metabolites in accordance with the micro-organisms' metabolism, and in interaction with the host environment. Thus, another paradigm through which the gut microbiota can be described is the quantification of the biochemical reactions that take place in its context. These mechanics are decomposed, described and made available in databases such as the Gene Ontology (GO) terms [57, 58] or the Kyoto Encyclopedia of Genes and Genomes' (KEGG) Enzyme Commission (EC) numbers [56, 117], among others. Methods to build such functional profiles from sequenced reads have been implemented, notably PiCRUSt [49, 50] and HUMAnN [34, 47, 48], respectively designed for treatment of 16S and MGS reads.

In the medical community, discussion has emerged in the past years around the potential benefits of shifting from taxonomic analysis to understanding the functional aspects of the gut microbiota [29]. Many consider that host-microbiota interactions can only truly be understood on the functional level, and that comprehension of the microbiome on this scale is essential to envision strategies to improve host health via the gut microbiota. In

this regard, taxonomic profiles on their own provide insufficient information, as a taxonomic unit can be functionally redundant and therefore have more relevance in regard to their cumulated influence on the community's metabolome. Stakes are therefore emerging around the implementation of methods to translate taxonomic descriptions of the microbiota into functional ones. Indeed, with an abundance of previous works on taxonomic profilings of the gut microbiota, many such profiles have been made available and, if transformed into functional profiles, could be a resource for further exploration of the issue at hand. PiCRUSt and HUMAnN can operate such conversions, but with caveats, as both of these processes require extra information (a phylogenetic table for PiCRUSt, and access to the original reads for HUMAnN) and are specialised to profiles derived from one method: 16S rRNA sequencing for PiCRUSt, and MGS for HUMAnN.

A first contribution of this thesis was to implement a novel method for translating taxonomic descriptions of the gut microbiota into functional profiles, with a larger focus on genericity and traceability. Genericity of the method implies that it can be applied to a profile derived from 16S rRNA sequencing or MGS indiscriminately, and without the need for additional information. Traceability refers to the conservation of the information linking together taxa and annotations, to expand upon later on. To do so, we relied on the EsMeCaTa pipeline [67], which correlates taxonomic units to annotations, and calculated functional abundance scores based on these outputs.

2.1 From taxonomic profiles to functional descriptions of the microbiota: a new methodology for functional quantification on the basis of a reference-based approach.

The method described in this chapter, and illustrated by Figure 2.1, only requires as input a table describing a microbiota sample as relative or absolute abundances of taxonomic units, of which the taxonomic affiliation is detailed. This process involves two main steps: the first is to associate functional annotations (FAs) to the taxonomic units, and measure the importance of its expression by said taxonomic units. This is achieved through the EsMeCaTa pipeline, as described in 2.1.1. The second is to combine this information with the original taxonomic abundances to generate a score measuring

the total expression of each gathered annotation within the community. This step is explained in 2.1.2. After the scores are calculated, they can be normalized to better highlight discriminant features using the TF-IGM method [118], which is explained in 2.1.3.

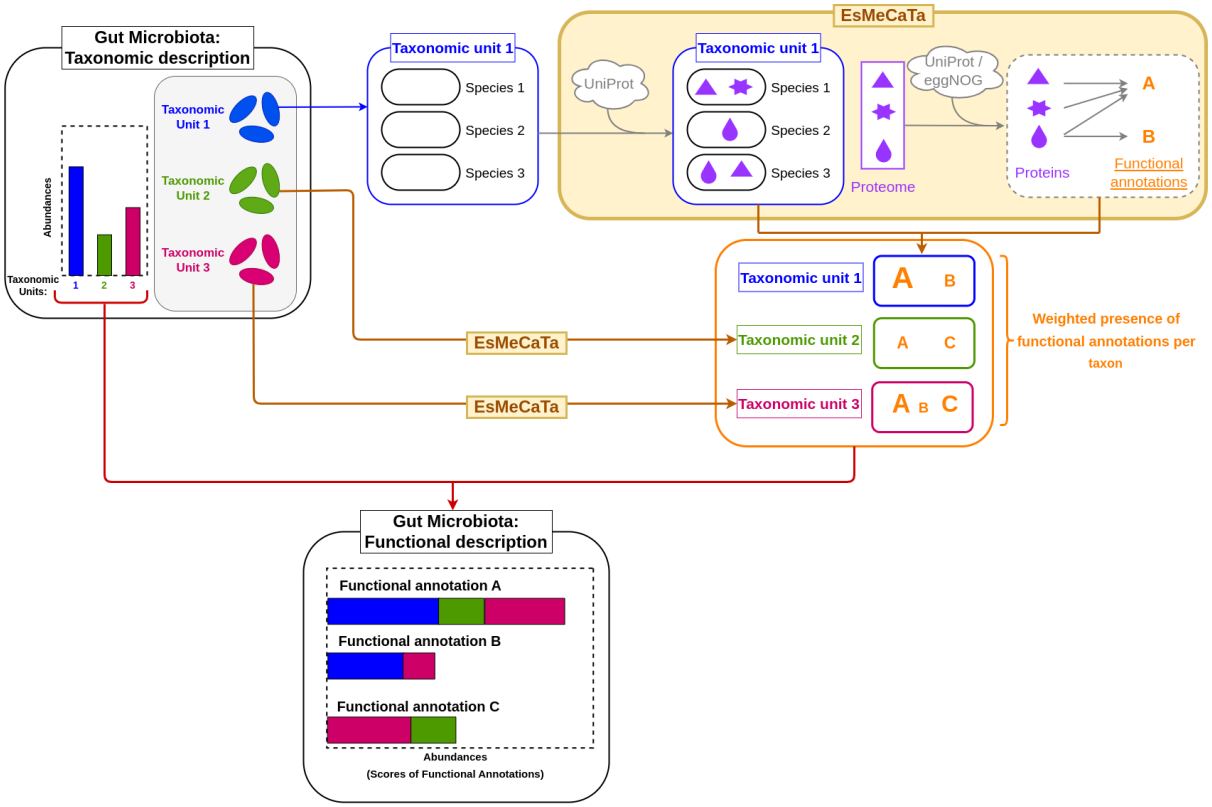


Figure 2.1 – **Method and processes for the calculation of a functional representation of the gut microbiota from taxonomic affiliations and abundances.** The method's only input is a taxonomic description of the gut microbiota, in the form of a table of taxonomic abundances (abundance of taxa or taxonomic units). For each taxonomic unit, the EsMeCaTa pipeline will recover the proteomes associated to the most precise recognized taxonomic description of the taxon in the UniProt database. The proteins that are representative of the taxonomic unit are then annotated, either through another interrogation of UniProt, or through the eggnog-mapper tool. From this information, we can estimate the importance of the expression of an annotation by a taxonomic unit. By combining this with the taxa' original abundances, we can calculate an abundance score for functional annotations within the gut microbiota.

2.1.1 Associating functional annotations to taxonomic affiliations: the EsMeCaTa pipeline.

The EsMeCaTa pipeline follows three steps. The first step, 'proteomes', takes as input a tabular that associates a given name for all the studied bacteria to their exact taxonomy. From this, EsMeCaTa interrogates the UniProt database [68] for proteomes associated with the taxon in question. If none can be found, the step is re-iterated with the superior taxonomic rank, until at least one proteome can be associated with the unit. In the event that more than 99 proteomes are associated with a taxon, a random selection of around 99 proteomes will be made, with respect to the taxonomic diveristy of the initial proteomes set. The selected proteomes are then downloaded from UniProt.

The second step, 'clustering', selects protein clusters that are representative of the taxonomic unit within the downloaded proteomes. To do so, the MMseqs2 tools [69] is used to create clusters of similar proteins from the proteomes. If a protein cluster contains similar proteins from 95% of the proteomes attributed to the taxonomic unit, it will be retained as part of its meta-proteome.

The final step, 'annotation', fetches the FAs (GO terms and EC numbers) of the retained protein clusters. It can do so by interrogating the UniProt databases, or by using the eggnog-mapper tool [70, 71]. The former option simply queries UniProt and fetches the information relative to the annotation of the proteins retained in the cluster. The latter is more demanding in terms of computation, as it involves an alignment step to match the retained proteic sequences to the contents of the eggNOG database [71] during a search step, before proceeding to protein annotation in the context of orthologs inferred by the tool in a scope defined by the user [70]. The final output is an ensemble of tabulars, one per taxonomic affiliation in the input, that contains all of the protein clusters kept in the taxon's meta-proteome and their FAs.

2.1.2 Calculating a functional representation of the patient's microbiota from taxon-annotation pairings.

EsMeCaTa, being reference-based, does not provide a quantification of the functional annotations within the sample itself. As such, in order to compute a representation of the gut microbiota on the scale of the FAs, mixing information concerning its specific composition with the associated metabolic mechanisms, we give each annotation (F) a score, labeled as a Score of FA (SoFA), within a subject sample (i), similarly to PiCRUSt

[49], according to the following formula:

$SoFA_{F,i} = \sum_t n_{t,i} \times x_{F,t}$

where $n_{t,i}$ is the abundance value of taxon t within sample i, and $x_{F,t}$ is the number of proteins within taxon t's proteome that are linked to the function F.

As such, each annotation's SoFA is equal to the sum of the abundances of all taxa that express it, weighted by the strength of said expressions as measured by EsMeCaTa.

In summary

This thesis' first contribution is a new method for building a functional description of the gut microbiota directly from a quantified taxonomic profile. This method involves running the EsMeCaTa pipeline on the input taxa's affiliations to generate weighted associations between them and functional annotations. The second step is to measure the total expression of each gathered annotation as the weighted sum of the abundances of the taxa that express them.

2.1.3 Normalizing and scaling data based on expected relevance with TF-IGM.

The TF-IGM method [118] is used to normalize the results presented in this article. It was originally exploited in Natural Language Processing, as a method to highlight terms in a corpus of texts that are significantly present within a text while penalizing those that are too widespread. The formula had to be re-adapted to fit our data and circumstances, and in our pipeline it is calculated based on the following two components:

— TF (Term Frequency): equivalent to the frequence of an annotation within the totality of a sample i:
$tf_{f,i} = \frac{SoFA_{f,i}}{\sum_{j \in J} SoFA_{j,i}}$
where $SoFA_{f,i}$ is annotation f's score within sample i, and J is the ensemble of the annotations recorded within sample i.

— IGM (Inverse Gravity Moment): for each annotation f, the calculated values for $tf_{f,i}$ are ranked in decreasing order and noted as $T(f)_1$,...,$T(f)_n$, so that $T(f)_1 > T(f)_2 > ... > T(f)_n$, n being the total number of samples. We then have:
$igm(f) = \frac{T(f)_1}{\sum_{r=1}^{n} T(f)_r \times r}$
where r is the rank of the T(f) score in the previously defined order.

The total TF-IGM score of an annotation f within a sample i will then be:

$tf\_igm(f,i) = \sqrt{tf_{f,i}} \times (1 + \lambda \times igm(f))$

where λ is a value between 5 and 9. As per Chin et al.'s [118] recommendation, its value was set to 7 by default.

2.1.4 Presentation of the test datasets.

This method for functional profiling was tested using publicly available species-level abundance profile datasets from the MetAML repository [87] and post-processed for DeepMicro [94], concerning subjects diagnosed with a variety of diseases: Cirrhosis [89], Colorectal Cancer [88], Inflammatory Bowel Disease (IBD) [91], Obesity [90], and Type 2 Diabetes on a Chinese [92] (T2D) and an European [93] cohort (WT2D). Each subject in these datasets had their gut microbiota sampled and sequenced with whole-genome shotgun and Illumina paired-end sequencing. The results were processed as per the standard procedure described by the Human Microbiome Project[119], then converted to species-level relative abundance profiles via the MetaPhlAn2 tool [33] with default parameters. Sub-species level features were then filtered using the MetAML tool [87].

Each cohort includes a portion of healthy control individuals, in addition to those who suffer from the disease in question. The proportions of each group in our cohorts are detailed in Table 2.1. These cohorts were also used to test and benchmark the methods presented in subsequent chapters.

Disease	**Dataset**	**Total samples**	**Control samples**	**Patient samples**	**Raw sequences file size**
Liver Cirrhosis	Cirrhosis	232	114	118	1.1 TB
Colorectal Cancer	Colorectal	121	73	48	985 GB
Inflammatory Bowel Disease	IBD	110	85	25	442 GB
Obesity	Obesity	253	89	164	1.5 TB
Type 2 Diabetes	WT2D (European Women Cohort)	96	43	53	303 GB
	T2D (Chinese Cohort)	344	174	170	796 GB

Table 2.1 – **Distribution of samples within the datasets of reference.**

2.2 Comparison with sequence-based approaches.

2.2.1 EsMeCaTa is a faster and lighter approach to functional assignation.

In order to position our method with the state of the art, we compared the outputs of our functional profiling method applied to the IBD dataset with a profile obtained through the application of the HUMAnN3 tool [34], applied directly to the MGS reads of the IBD dataset. This process involved MetaPhlan4 [35] for initial taxonomic profiling, used with default parameters, with reference to the ChocoPhlan database [34] (version mpa_vJan21_CHOCOPhlAnSGB_202103), and the UniRef90 database [52]. Some reads from sample V1_UC-19 were corrupted, blocking HUMAnN3's application to this specific sample. As such, it was also removed from the functional profiles fed to EsMeCaTa in the context of a comparison between both approaches.

Functional profiling tool	**Duration**
HUMAnN3	27d 22:06:00
EsMeCaTa (UniProt)	2d 04:12:45
EsMeCaTa (EggNOG)	6d 16:03:51

Table 2.2 – **Running time of the functional profiling tools.** Benchmarks were made on the IBD dataset, using the raw MGS reads as input for HUMAnN3, and taxonomic profiles obtained from Pasoli et al. [87] for EsMeCaTa. All jobs were launched on a calculation cluster, with 10 CPUs and 150 GB of memory at their disposal.

Of all the tested processes, HUMAnN stands out as the heaviest one to run in terms of input size, which consisted of 442 GB of reads in the case of the IBD dataset (see Table 2.1), but also in terms of run time, as illustrated by Table 2.2. Indeed, EsMeCaTa runs from a 301.5 kB input, in the form of a taxonomic abundance table, and is shorter by a factor of around 4 in the case of EsMeCaTa with eggnog-mapper [70, 71], and 13 in the case of EsMeCaTa with UniProt [68]. This can partially be explained by the fact that EsMeCaTa's input has already been processed by MetaPhlan, itself run by HUMAnN3 in its first half [34]. However, it also clearly illustrates the benefit in terms of computational workload of relying on EsMeCaTa instead of working directly from raw reads when a processed taxonomic profile is already available, as is the case for the datasets presented in Table 2.1. Within EsMeCaTa, a difference can also be observed between the runtimes of the UniProt and EggNOG versions of the pipeline, the latter taking 3 times as long to run as its counterpart. This difference is entirely due to the 'annotation' step of EsMeCaTa (see

Section 2.1.1), seeing as the first two steps of the pipeline are common to both versions. This stems from the extra processing required by eggnog-mapper, notably concerning the alignment of the proteic sequences to the contents of the referred database (see Section 2.1.1). It should be noted however that, though it is a less complex process, the speed and reliability of the interrogation of the UniProt database is most susceptible to fluctuate as it is dependent on the quality and stability of the connection with the online resource. Some applications of this process during the course of this thesis have failed altogether due to issues in the interrogation of the online client.

2.2.2 EsMeCaTa and HUMAnN recover similar information.

EsMeCaTa recovers more information than HUMAnN, both when interrogating UniProt or using EggNOG for annotation retrieval. As shown by Figure 2.2, the EggNOG version of EsMeCaTa recovers the most information, with a total 16,340 annotations associated to the dataset. This is over twice as much information as HUMAnN3, with 7,973 annotations, was able to retrieve. These lists have a total consensus of 6,385 annotations, meaning that 80.1% of the annotations gathered by HUMAnN are also found by EsMeCaTa using EggNOG. This consensus is smaller with the UniProt version of EsMeCaTa, covering 5,903 total annotations, which accounts for 74% of the list gathered by HUMAnN. This shows however that EsMeCaTa is capable of finding the majority of the FAs that HUMAnN3 does, along with extra new information, without needing access to the original sequence files.

The list of annotations gathered by EsMeCaTa includes a higher proportion of GO terms than HUMAnN, accounting for 81.5% and 82.9% of the UniProt and EggNOG retrieved lists respectively against 71.7% for HUMAnN. They are also the main source of novelty within the annotation lists: 44.9% of the UniProt-gathered GO terms and 66.7% of those obtained through EggNOG were not found by HUMAnN, against mirroring proportions of 29.8% and 33.0% for EC numbers.

Figure 2.3 illustrates the amount of GO terms from each namespace (Molecular Function, Biological Process or Cellular Component) recovered by each method. We can see that the increased amount of GO terms retrieved by EsMeCaTa in comparison to HUMAnN3 is mostly explained by an increase in the amount of cellular components and biological processes within the recovered lists. The former category is 2.7 times as present in the output obtained from EggNOG compared to HUMAnN, and 1.8 times as present in the one obtained from UniProt, while the latter category shows respective increase

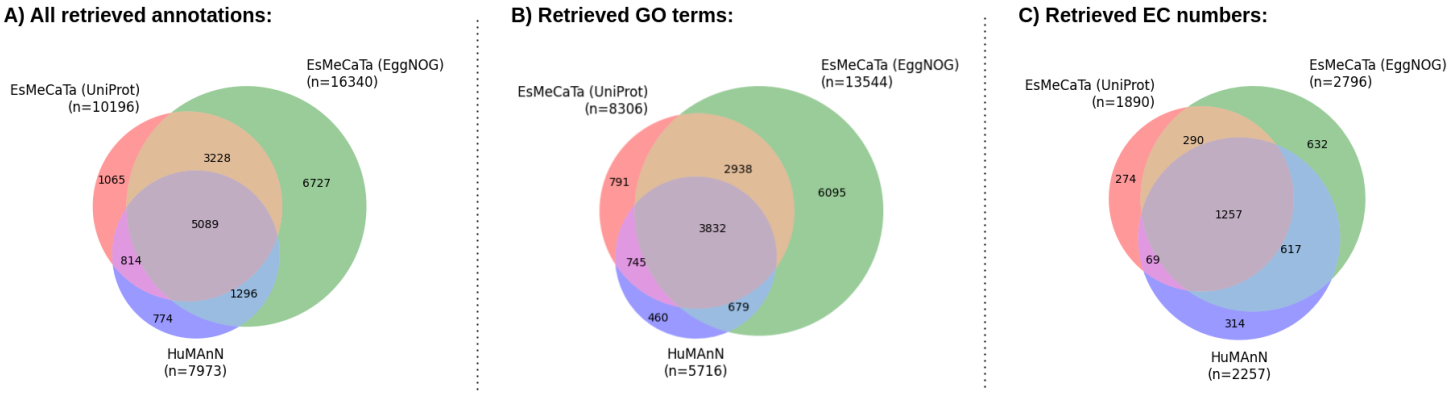


Figure 2.2 – **Size and overlap of the functional annotation lists retrieved from taxonomic profiles using EsMeCaTa, both using UniProt or EggNOG for the *annotations* step, and from MGS sequences using HUMAnN3 .** The numbers represented on the graphs give the absolute amount of annotations in the overlap. The total amount of annotation included in each category is indicated by its label, as n. **A)** Representation of all annotations retrieved by each method. **B)** Same representation, focusing on the GO terms retrieved by each method. **C)** Identical to B), but focusing on the retrieved EC numbers.

ratios of 4.0 and 1.8, making it the most represented category in both of EsMeCaTa's associations. The Molecular Function namespace remains constant throughout the lists, increasing by a factor of 1.2 in both of EsMeCaTa's outputs when compared to HUMAnN. This increase in information could have several implications: on the one hand, an increase in information increases the precision of the community's functional description. On the other, it makes the resulting functional descriptions less understandable, as the quantity of information becomes overwhelming. While the EsMeCaTa-based functional profiles have the potential to be more thorough as descriptors, a complementary variable selection would be necessary to make it viable as a resource for biological interpretation. These questions will be addressed in the subsequent Chapter 3.

In summary

We compared EsMeCaTa to HUMAnN3 in terms of performance and on the nature of the information gathered by both methods. EsMeCaTa uncovers more functional information than HUMAnN3, notably more GO terms of the "biological process" category. At the same time, the large majority of the information gathered by HUMAnN3 (∼75% of the annotations) is also found by EsMeCaTa, at a much lesser cost in terms of input size and computation time.

EsMeCaTa's performances are however dependent on the annotation approach that it leverages. With eggnog-mapper, it will gather more information, but at a higher computational cost. With UniProt, the process is lighter and quicker, but is less stable and gathers less information.

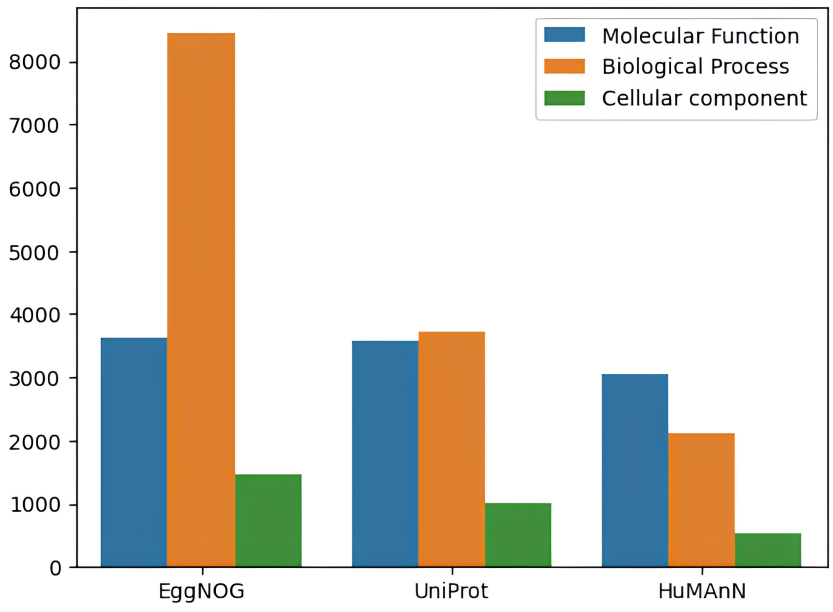


Figure 2.3 – **Amount of GO terms from each namespace obtained from taxonomic profiles using EsMeCaTa, both using UniProt or EggNOG for the *annotation* step, and from MGS sequences using HUMAnN3.**

2.3 A first exploration of the taxa's functional expression.

2.3.1 Exploring the associations between taxa and annotations exposes their non-redundancy.

The availability of both the taxonomic and functional scales as support for analysis brings the question of the redundancy between both paradigms. In other words, there is limited novelty to be derived from functional annotations that are exclusively associated to a taxon, as this functional signature could be derived from the taxonomic signal without need for an exploration of the functional paradigm.

We evaluated this prospect by looking into the taxa-function associations obtained with EsMeCaTa UniProt applied to the IBD dataset, which yields 10,196 annotations from 443 taxa. On average, annotations are associated to 47.8 taxa. One annotation is associated with the most taxa (437 taxa out of 443): GO:0016021, which is attached to the cellular membrane component, and is therefore expected to be extremely widespread. Unique associations account for 37.5% of all annotations, thus a majority of annotations are associated to more than one taxon. Overall, no function is perfectly ubiquitous, and the majority of functions are linked to several different taxa, and are therefore not directly redundant with the taxonomic information. Furthermore, the fact that most functions are expressed by several different taxa illustrates the functional redundancy evoked previously [29], and therefore confirms the possibility that functional signatures could be derived from a cumulation of several taxa's influences

2.3.2 Evaluating the functional proximity of taxa highlights the prevalence of unique functional profiles.

Another aspect of the functional profile that could limit its independence from the taxonomic scale is the prevalence of taxa with the same metabolic profiles. If the annotations that are expressed by several taxa are only ever expressed by species that have the exact same metabolism, this would amount to having a functional profile that is redundant with the taxonomic scale if it was clustered based on ecological roles. To quantify functional redundancy among taxa, we used Jaccard proximity [120] to measure the similarity of their functional associations based on the associations obtained with EsMeCaTa UniProt

applied to the IBD dataset. An excerpt of the calculated distances are illustrated in Figure 2.4. The full heatmap was separated in four parts, available in Appendix A.

Taxa with a Jaccard distance of 5% or less were considered functionally identical. For example, on Figure 2.4, this is shown to be the case for *Bacteroides gallinarum* and *Bacteroides sp 1 1 30*, as well as *Clostridiales Family XIII Incertae Sedis unclassified* and *Clostridiales bacterium 1 7 47FAA*. As such, these taxa would be grouped together and considered functionally redundant.

Figure 2.5 illustrates all of the groups of functionally redundant taxa obtained in this manner on the whole dataset. There are a total of 32 groups, represented in red, containing between 2 and 7 taxa each. In total, 101 taxa have at least one functionally redundant taxon in the dataset, which amounts to 22.8% of all taxa. A closer look at these groups shows that they largely englobe taxa that are phylogenetiacally close, as all but two of the groups only contain taxa of the same genus. This is coherent with what we would expect from EsMeCaTa, which conducts a taxonomy-based approach to annotation.

The remaining 342 taxa, accounting for 77.2% of the total, do not have such close neighbors however, meaning that there is a majority of functionally singular taxa in this dataset. This further cements the idea that the information contained in the functional set could not be easily derived from the taxonomic data, marking it as potentially innovative. It is interesting to note that if the taxa maintain distinct functional profiles, it is in spite of sharing many annotations with each other, as previously mentioned in Section 2.3.1. This means that functional cumulation could happen between species that occupy otherwise entirely different ecological niches within the microbiota.

In summary

The majority of functional annotations are expressed by more than one taxon, yet in spite of this, 77.2% of taxa have a unique functional profile within the community. This illustrates that the functional information does not align with the taxonomic scale, and that the eventual importance of most functional annotations couldn't be attributed to the influence of only one taxon or group of taxa *per se*.

Figure 2.4 – **Jaccard distances calculated between the functional profiles of a sample of taxa from the IBD dataset, annotated by EsMeCaTa with UniProt.** The details of the distances calculated on all taxa are available in Appendix A.

Figure 2.5 – **Representation of the functional redundancy between taxa of the IBD dataset.** Taxa are considered to be functionally redundant if they have a Jaccard distance of 5 % or less in terms of the functional annotations that are assigned to them. Taxa in red have at least one taxon that is functionally redundant with them, and are grouped with them in the representation. The others, in blue, have no other redundant taxon. The blue group accounts for 77.2% of all taxa, while the red contains the remaining 22.8%.

2.4 Conclusion and discussion.

We have established a new method for functional representation of the gut microbiota, which requires only a descriptive quantification of taxonomic units to be calculated. This feature differentiates our approach from the state of the art methods, notably PiCRUSt and HUMAnN, which cannot make this conversion from the taxonomic profiles alone. This approach is based on the EsMeCaTa pipeline, which explicitates and quantifies the links between taxa and their FAs through the retrieval of the annotations associated with a given taxon in the UniProt or eggNOG database. This contributes to the approach's genericity, making it applicable to profiles derived from 16S rRNA and MGS sequencing alike, whereas the other cited tools are specialized in only one of these profiles. A comparison with the outputs generated by HUMAnN shows that the EsMeCaTa pipeline retrieves more information than its counterpart, and recovers the majority of the information gathered by HUMAnN. Further investigation into the specifics of the calculated profiles has allowed us to determine that functional and taxonomic profiles generated by our method were not redundant.

The limits of reference-based annotation approaches.

It should be noted that EsMeCaTa's exploitation comes with caveats, that also need to be addressed. The pipeline's reliance on UniProt for at least the association of proteomes to the taxonomic units means that any bias in the remote database would impact the tool as well, such as the inclusion of proteomes not adapted to the samples' environment of origin. Any information missing from UniProt concerning the taxa can also only be inferred, through the referral to the consensus proteins of the upper echelon of the taxonomy. As such, imprecisions in the annotation should be expected. This setback can however be mitigated in light of the comparison made in this chapter with HUMAnN3's results, which show that a significant portion of the information gathered by EsMeCaTa is also validated by other sources. This referral to external databases is also a feature of HUMAnN with UniRef and NCBI resources, and PiCRUSt with IMG, and is therefore a limitation that can be applied to functional profiling approaches in general.

Dependency of the functional score calculation on taxonomic abundances.

Another limitation to account for with our approach is inherent to the use of taxonomic profiles as an input. Though this approach makes the process lighter in terms of computational resources, it also make the tool reliant on the quality of the preprocessing steps applied to the profile, as there is no referral to the original reads. The taxonomic

abundances given in the input are a central component of the functional scores' calculation, therefore any biases introduced during the taxonomic profiling would have an impact on the scores calculated for the functional annotations. Knowledge of the data's origin therefore remains important when applying the method presented here. Generally speaking, it should be noted that there is an inherent bias to the exploitation of metagenomic data, notably concerning the taxa of lower abundance which are susceptible to be false positives.

Taxonomic and functional paradigms for the exploration of the gut microbiota.

The development of a lightweight approach for functional profiling opens up many possibilities for the exploration of gut microbiota data, as it makes functional descriptions, which are more in line with the demands of the medical community, more widely accessible. On this basis, we can better explore the functional intricacies of the microbiota, but also evaluate the comparative pros and cons of each paradigm in analytic tasks. The following chapter will cover one aspect of this question, by comparatively using both profiles to classify host disease state.

Chapter 3

A method for robust classification in situations of unbalanced dimensionality.

Several studies have evaluated the gut microbiota's potential as a basis for predicting an individual's health status [87, 94]. Among the potential strategies for improving classification performance, the perspective of correcting the data's dimensionality has a lot of potential. Indeed, the dimensions of taxonomic descriptions of the microbiota do not favor classification, as the amount of samples available for training in the dataset is on average half as high as the number of descriptors (around 200 samples for 400 taxa, see Tables 2.1 and 3.2). As such, previous studies have consistently found that variable selection enhances classifiers' performances: both Statnikov et al. [27] and Passoli et al. [87] enhanced classification performances by selecting taxa based on their importance within trained RF classifiers. These results confirm that the gut microbiota, as an input, is impeded by a Hughes phenomenon [26], also known as "curse of dimensionality". This issue is even more prevalent when tackling functional profiles, as for a similar amount of training samples, the number of descriptors augments in scale from around 400 taxa to around 10,000 functional annotations (see Table 3.2), though the effect of variable selection on functional profiles has not been explored as thoroughly.

Different approaches for variable selection have been employed in the context of supervised classification from microbiota samples, often based on RF importance rankings [80], seeing as these models have proven to be among the best adapted for these tasks, notably outperforming SVMs (see 1.2.1). The previously mentioned works of Statnikov et al. [27] and Passoli et al. [87] operated a selection of taxonomic units from this basis: removing a fixed percentage of the least informative variables for the former, and retaining the top k features, with k the amount in a set list of values that maximizes model accuracy, for the latter. Among other notable studies, Jones et al. [98] explored the top 30 features obtained from RF classifiers, trained on both taxonomic and functional profiles, and Douglas et al. [99] looked up the top features of a model trained on a profile combining both types

of features. However, these results were not used for variable selection, but rather as a basis for discussion over the relative influence of taxonomic and functional profiles. There is therefore a lack of concrete results in the literature concerning the effects of variable selection on the classification performance of models trained on functional profiles of the gut microbiota. We can also note that the presented methods relied on fixed thresholds, be they an absolute amount of variables or a percentage of the total list. No method akin to a fully automatic selection was explored.

From this observation, we developed a method to train classifiers on microbiota data, both taxonomic and functional, which integrates an adaptation to the data's dimensionality through an automated variable selection process, but also with an enhanced focus on robustness, both in terms of the classification performances, but also of the selected variables list's contents. To achieve that goal, we extended the MetAML [87] and DeepMicro [94] procedures in three dimensions, with an increased number of trained RFs, selection of important variables with an iterative process and repetition of the full selection process for the exploration of the inherent variability in performances due to changes in training conditions.

3.1 Methodology for robust classification.

This section details this thesis' second contribution: a methodology for classification and variable selection which is adaptable to datasets of unbalanced dimensions, and is therefore suited for using descriptions of microbial communities as an input. In the same vein as our previously mentioned references [27, 87], a variable selection on the basis of RF importance scores was used to tackle issues related to the dimensionality of the data, with the introduction of a novel method for automatically computing a selection threshold, which is then iterated to obtain an optimal level of selection. The process is described by Algorithm 1 and Figure 3.1.

The method's robustness was another area of focus. In order to evaluate this characteristic, several repetitions of the training process are required. Aggregating the results of several classifiers is also a known approach to enhance the robustness of RFs as a means for variable selection [28]. Our approach is based upon the MetAML [87] and DeepMicro [94] procedures which describe the average results of, respectively, 20 and 5 RFs trained from a predefined seed. To gain in robustness, we train 20 independent RFs to predict the patient's status, and extract the average classifier performances and variable importance

rankings from this re-training procedure. Our approach then automatically extracts a shortlist of important features from variable importance, and trains 20 new RFs on the selected features. A run of the method consists of 5 iterations of this selection procedure. These runs are repeated 10 times, each time with a different test set put aside for performance measurement. After this, the selection level that gives the best overall classification performance is retained as the optimal iteration. This extension of MetAML's procedure in 3 dimensions (re-training of 20 RFs, iteration of the variable selection process, and repetition of the entire procedure over 10 runs with a different test set for each) is a guarantee for robustness, and allows for the exploration of inherent variability in performances due to changes in training hyperparameters (see Figure 3.2).

3.1.1 Random Forests for reliable classification.

For ML classification, we used RF models [80]. These classifiers are known from the literature to be one of the best performing models in tasks related to microbiota classification [27, 87, 94]. Unlike other models with notable performances in these studies, such as SVMs [78], RFs can also be used for feature selection and model interpretation through measurement of feature importances [80].

In our method, RF classifiers are trained to sort individuals in two classes (patients or controls, for example), based on the relative abundance profiles of their microbiota or on their calculated mechanistic representation. Before any training, a subsample of 20% the size of the full dataset is set aside as a test set. During training, the remaining data is randomly split into a training set and a validation set, with a respective 80% / 20% distribution. In order to account for the disparity in representation between the different categories of individuals within the datasets, both classes were given weights proportional to their frequency, as implemented by scikit-learn's 'balanced' class weight parameter [121]. When measuring the performance of our classification algorithms, the metric used was the median Area Under the Receiver Operating Characteristic Curve (AUC) [86] over 20 re-trainings of the models, measured on the test set initially set aside.

Though RFs are the basis of our classification approach, our method involves the integration of a re-training process with a resampling of the training subsets, in order to assess robustness (see Section 3.1.3).

Algorithm 1 Methodology for iterative classification and variable selection

Require: $DataTable$ ▷ Abundance table of taxa/FAs in gut microbiota samples
Require: $DataLabels$ ▷ Label associated to each sequenced sample
Require: $n(runs) \geq 1$ ▷ Default Value: 10
Require: $n(iterations) \geq 1$ ▷ Default Value: 5
Require: $n(classifiers) \geq 1$ ▷ Default Value: 20

Phase 1 – ***Train n(classifiers) RF classifiers n(runs) times, each time with a variable selection iterated n(iterations) times***

Repetition:
$run \leftarrow 1$
while $run \leq n(runs)$ **do**
$TrainingSamples(run)$, $TestSamples(run)$, $TrainingLabels(run)$, $TestLabels(run) \leftarrow SetAsideTestIndividuals($ $DataTable$, $DataLabels$) ▷ See 3.1.1
Iteration:
$iteration \leftarrow 1$
while $iteration \leq n(iterations)$ **do**
Re-training:
$classifier \leftarrow 1$
while $classifier \leq n(classifiers)$ **do**
if $iteration = 1$ **then** $ValidationSamples(classifier) \leftarrow SetAsideValidationIndividuals($ $TrainingSamples(run)$, $TrainingLabels(run)$) ▷ Validation sets are defined on the first iteration for each classifier, and are re-used for all iterations of the same run (see Figure 3.2)
end if
$TrainingSamples(run, classifier)$, $ValidationSamples(run, classifier)$, $TrainingLabels(run, classifier)$, $ValidationLabels(run, classifier) \leftarrow SetAsideValidationIndividuals($ $TrainingSamples(run)$, $TrainingLabels(run)$, $ValidationSamples(classifier)$)
$TrainedClassifier$, $VariableImportances[classifier] \leftarrow RandomForestTraining($ $TrainingSamples(run, classifier)$, $TrainingLabels(run, classifier)$) ▷ See 3.1.1
$AUC[classifier] \leftarrow TestingClassifier($ $TestSamples(run)$, $TestLabels(run)$) ▷ See 3.1.1
$classifier \leftarrow classifier + 1$
end while
$MedianAUC[run, iteration] \leftarrow Median(AUC[1 : n(classifiers)])$
$SelectedVariables[run, iteration] \leftarrow VariableSelection(VariableImportances[1 : n(classifiers)]$) ▷ See 3.1.2
$TrainingSamples(run)$, $TestSamples(run)$, $TrainingLabels(run)$, $TestLabels(run) \leftarrow KeepSelectedVariables($ $TrainingSamples(run)$, $TestSamples(run)$, $TrainingLabels(run)$, $TestLabels(run)$, $SelectedVariables[run, iteration]$) ▷ See 3.1.2
$iteration \leftarrow iteration + 1$
end while
$run \leftarrow run + 1$
end while

Phase 2 – ***Post-processing: aggregating the results of the multiple training instances, by iteration level***

for $i \leftarrow [2 : n(iterations)]$ **do** ▷ The optimal iteration must have effectuated at least 1 selection
$IterationMedianAUC[i] \leftarrow Median(MedianAUC[1 : n(runs), i])$
end for
$MaxRunAUC$, $MaxIteration \leftarrow Max(IterationMedianAUC)$
$RobustVars, ConfidentVars, CandidateVars \leftarrow IntersectShortlists($ $SelectedVariables[1 : n(runs), MaxIteration$]) ▷ See 3.1.3
return $MedianAUC[1 : n(runs), MaxIteration]$, $RobustVars$, $ConfidentVars$, $CandidateVars$

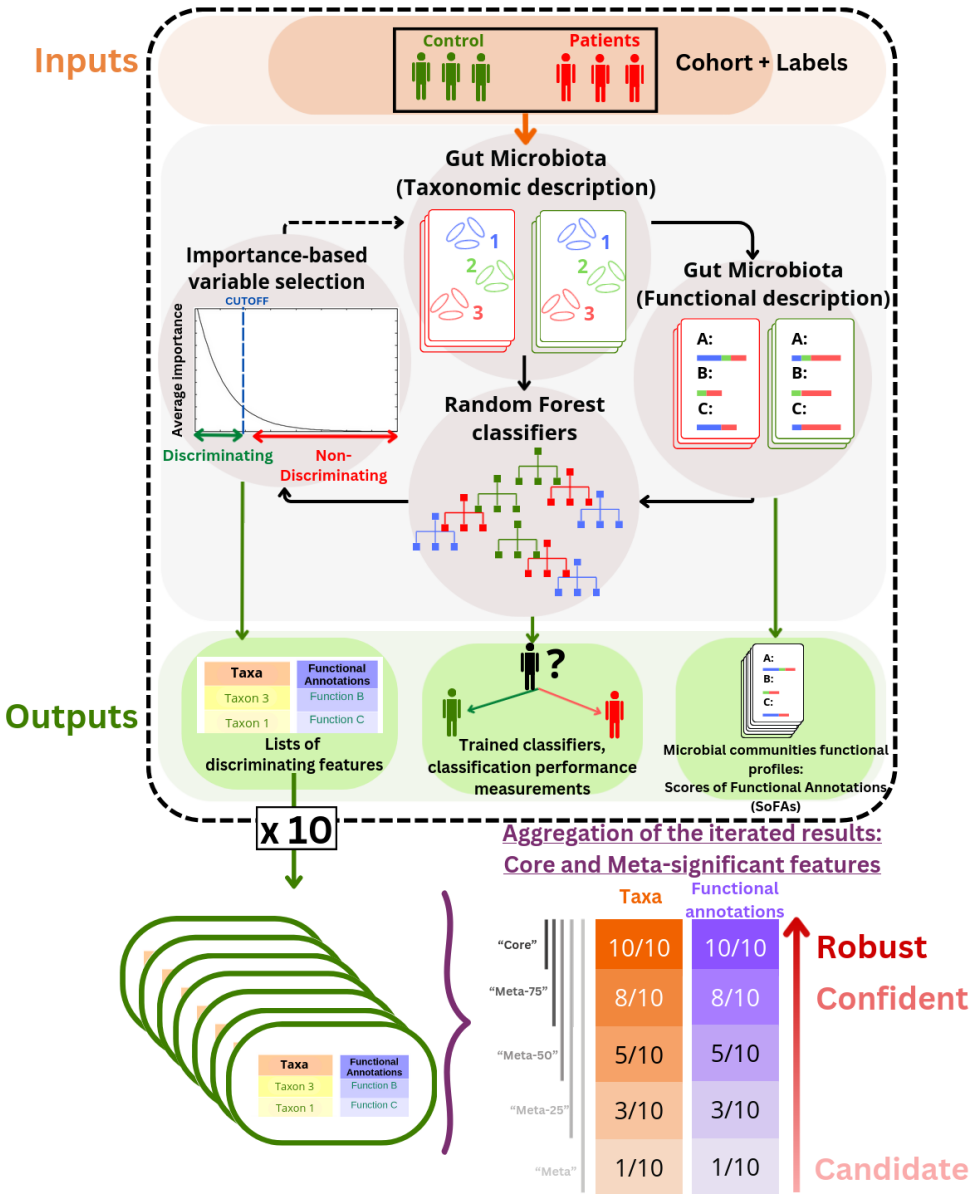


Figure 3.1 – **A schematic representation of a pipeline for iterative and selective classification and variable selection.** From taxonomic abundance tables and their associated labels as inputs, the pipeline produces functional descriptions of the microbiota samples via the EsMeCaTa pipeline. Both of these profiles are then used as basis for the training of RF models to discern Control from Patient profiles. The average importance scores (Gini by default) of these variables over all trained forests is then used as basis for a selection of discriminating variables, which can then be processed again iteratively, or passed as an output. For robustness, the process is repeated 10 times, leading to 10 different lists of discriminating taxa and FAs. These lists can be compiled into different categories, which group variables by level of robustness based on the frequency of their appearance in the significant lists. Thus, unanimous variables are considered to be "robust" discriminators, those agreed on by 75% or more of the classifiers are considered "confident", and those that are selected at least once are considered "candidates". Internally to the pipeline implementation, robust features are labeled "Core", and the others are labeled as "Meta-X", X being the percentage of discriminating variable lists that include them.

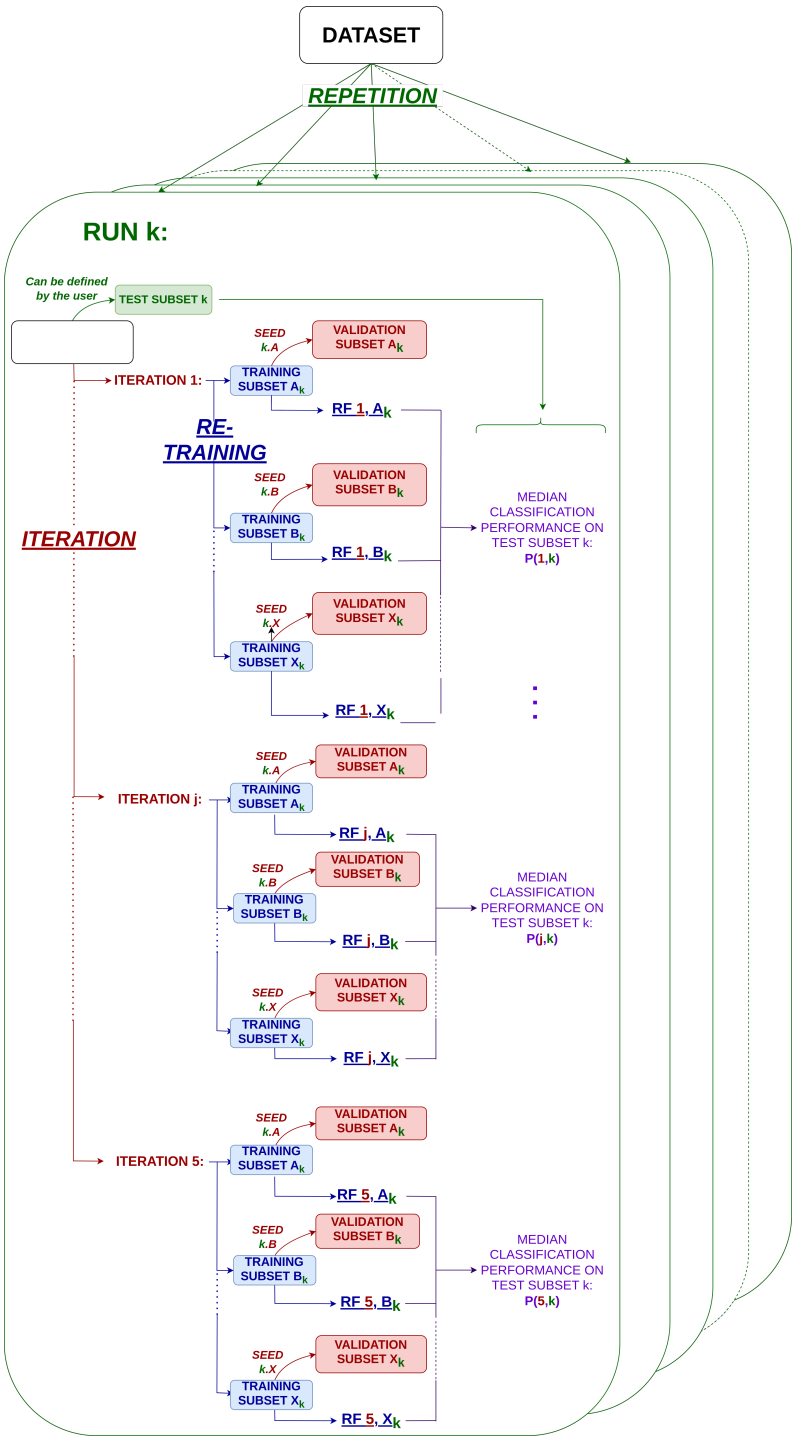


Figure 3.2 – **Classification algorithm as implemented for our approach:** For a given run k, a test subset is randomly selected within the initial dataset and set aside. A given iteration j consists in training X random forests (20 by default), each having a dedicated validation subset. These 20 forests are used to compute a median classification performance $P(j,k)$ and a shortlist of important features. This list is used to train the X random forests of iteration $j+1$. By default, 10 runs and 5 iterations are launched.

3.1.2 Automatically extracting discriminating information from trained classifiers.

Following the classifier's training, the resulting feature importances are extracted. By default, we used the Gini Importance metric, calculating the mean accumulation of the impurity decrease within each tree, as implemented in the Scikit-learn Python library [121]. The feature importances of all 20 trained models are averaged, and features are then ranked based on this metric in decreasing order. These scores, when ordered from highest to lowest, display a kink-like shape.

Once ordered, we aim to distinguish a separation between the features that were essential to the clasifier's functionality, and those with a lesser impact. We place this threshold at the inflection point of the curve representing the decreasing importance scores, determined via an implementation of the Kneebow method [122], with all features above this point being labeled as "Significant", and those below as "Non-Significant". Only Significant features are retained for the following selective iteration.

An example of this process is illustrated by Figure 3.3. In this example, all annotations are given a rank X based on their position in the decreasing order of Gini importance scores, averaged over 20 trained RFs. The resulting curve displays an obvious inflection, which is placed by the Kneebow method at the level of the variable ranked at position 541. As such, the annotations between rank 0 and 541 are considered to be "Significant" and are therefore selected, and those ranked at 542 onward are removed, reducing the subset to 5% of its initial size.

3.1.3 Iteration and repetition of the process.

This selection process is iterated 5 times, and the full process is repeated over 10 runs. Each repetition involves the selection of a new test subset, as per the procedure described in Section 3.1.1, and each iteration involves the re-training of 20 classifiers with the setting aside of the same validation subsets as those defined during the first iteration (see Algorithm 1 and Figure 3.2). Following this, the optimal level of selection that is retained is the one that yields the best classification metric after at least one variable selection over 10 repetitions.

This process generates shortlists of discriminating features that can be combined for a robust consensus. With 10 applications, each time with different test subsets, variations in the contents of these shortlists consistently occur. To address this, variables are categorized

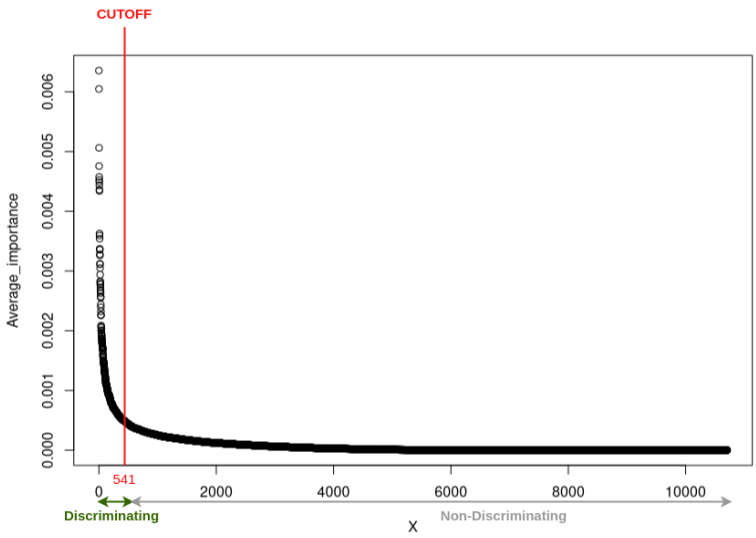


Figure 3.3 – **Illustration of the automatic variable selection process.** The IBD dataset's annotations are given a rank (X) based on their average Gini importance score. The cutoff is operated at the inflection point of the curve, according to the Kneebow method. Annotations above the inflection point are labeled as "Discriminating", and are selected. Those below are labeled as "Non-Discriminating", and are removed.

as follows: **(i)** "Robust" if unanimously deemed discriminating in all runs (above the variable selection threshold). This category contains the variables that are most essential to the discernment of both patient profiles. **(ii)** "Confident" for the variables that were considered discriminating by at least 75% of the different runs (in our case, by 8 or more runs out of 10). This category contains variables that are likely to be important for profile discrimination, and could be a complement to the robust shortlist for interpretation. **(iii)** "Candidate" for variables shortlisted in at least one run. These are variables that should not be fully excluded from consideration when it comes to interpretation, but that are unlikely to be influential. More generally, across all of these categories, the robustness of a selected variable can be evaluated in the light of the number of different runs that list it as discriminting.

In summary

This thesis' second contribution is a novel approach for ML-driven classification and variable selection suited for the gut microbiota. This approach involves:

- **Re-training** several RF classifiers, and extracting their median classification performance as well as the average importance scores of all variables,
- **Iterating** the re-training process after an automatic variable selection, made with a novel strategy based on importance scores,
- **Repeating** the entire procedure, each time introducing variablility through the setting aside of a new test set.

This method gives a robust classification of the input data, but also outputs several selections of important variables, the most relevant of which is defined as the selection on which the best classification performances are obtained. Variations in the training conditions, with different test subsets selected for each repetition result in several different shortlists of discriminating features. We label as **'robust'** the features that constitute the intersection of these shortlists, as **'confident'** those that are present in 75% or more of them, and as **'candidate'** those that are present in at least one of them. The amount of times a variable is labeled as discriminating by the optimal level of selection is an indicator of how reliable it is for the distinction of the differentiated profiles.

3.2 Application of the method to publicly available datasets.

The method described in the previous section was applied to the previously presented publicly available datasets (see Table 2.1), which were notably also explored as supports for disease state classification by Passoli et al. [87] and Oh & Zhang [94]. Classification was done on the basis of the taxonomic profiles, but also on functional profiles derived from them through EsMeCaTa [67], as described in 2.1.2. This was done using an implementation of the process in Python (SPARTA pipeline, see Chapter 5). The following section will present the results of these applications.

3.2.1 Classification performances and impact of the variable selection.

Figure 3.4 and Table 3.1 illustrate the classification performances of RF classifiers [80] trained to distinguish between patients and healthy individuals, per profile and dataset. For each trained RF, the AUC is calculated. Seeing as 20 RFs are trained within an iteration, the median of these 20 AUCs is retained to represent the performances of the iteration as a whole. The full iterative process is repeated 10 times, giving 10 median performance metrics per level of iterative selection (see Fig 3.2). Classification on the taxonomic datasets prior to selection is analogous to the classification without representation learning method implemented in DeepMicro [94], with 20 RFs instead of 5 and dedicated test sets.

Figure 3.4 represents the performances obtained without any selection alongside the optimal selection, defined as the non-zero selection level that maximizes the median of this metric over 10 repetitions. For the latter profiles, the number of corresponding iterated selections are given in the 'Optimal Selection' column. For each dataset, a Mann-Whitney U-test was conducted comparing the performances based on the taxonomic and functional profiles at respective optimal selection levels. For example, the Colorectal dataset's functional (purple) and taxonomic (green) profiles have been tested over 10 runs. These tests have allowed us to detect the level of variable selection that yields the best median classification scores for each profile, which were then chosen for this representation. In this case, as shown in the 'Optimal selection' column, the functional dataset gives its best performance after 2 iterations of variable selection, whereas the taxonomic dataset gives its best performance after just one. The performances of RFs trained on taxonomic and functional profiles without selection are also represented, in red and blue respectively. Each of the 10 runs yields an average classification performance score, corresponding to the plotted dots. The boxplots represent the associated distribution and notably show that the functional profile has a median AUC of 0.85, against 0.86 for the taxonomic profile. The difference between both distributions was not found to be significant by a Mann-Whitney U-test, as shown by the absence of an asterisk symbol on this row.

Overall, we can see that taxonomic profiles yield better median classification performances than their functional counterparts, with the T2D dataset being the only exception. However, the difference in performance between both profiles is only significant in the case of the WT2D dataset, showing that though converting our data to the functional level

comes at the cost of some performance, both profiles perform comparably as basis for classification.

Table 3.1 gives the details of the impact of the iterative selection process on classification performances. By showing the performances obtained on the validation subsets, this table also illustrates the consequences of a data leak after variable selection. Indeed, seeing as the validation set changes for every RF trained during an iteration of our process, the following variable selection is based on variable importance scores derived in part from forests that have been trained on samples included in the validation sets. As such, once at least one level of selection has been conducted, we can see that performances on the validation test become increasingly superior relatively to those obtained on the test sets.

This table, along with Figure 3.4, also illustrates the asymmetrical benefit of variable selection. Functional profiles systematically benefit from a reduction of dimensionality, as their median performances after iterative selection (purple on Figure 3.4) are always superior to those obtained without variable selection (blue on Figure 3.4). For taxonomic profiles however, variable selection leads to a decrease in median results for three of the six datasets (Cirrhosis, WT2D, and IBD).

These results confirm the efficiency of variable selection as a way to improve classification performances in cases where the input dataset's dimensions are unbalanced, as is notably shown by how this approach consistently benefits the functional profiles, which are most unbalanced. The varying effects on taxonomic profiles on the other hand illustrate the importance of having an adaptive approach to the selection process, as different datasets can benefit most from different degrees of selection.

In summary

An application of our method to the datasets of the MetAML study has shown that translating taxonomic profiles into functional profiles usually comes at a cost in terms of classification performance, though the difference in performance between both profiles is rarely significant. The application of variable selection enhances performance on most of the taxonomic profiles, and on all of the functional profiles.

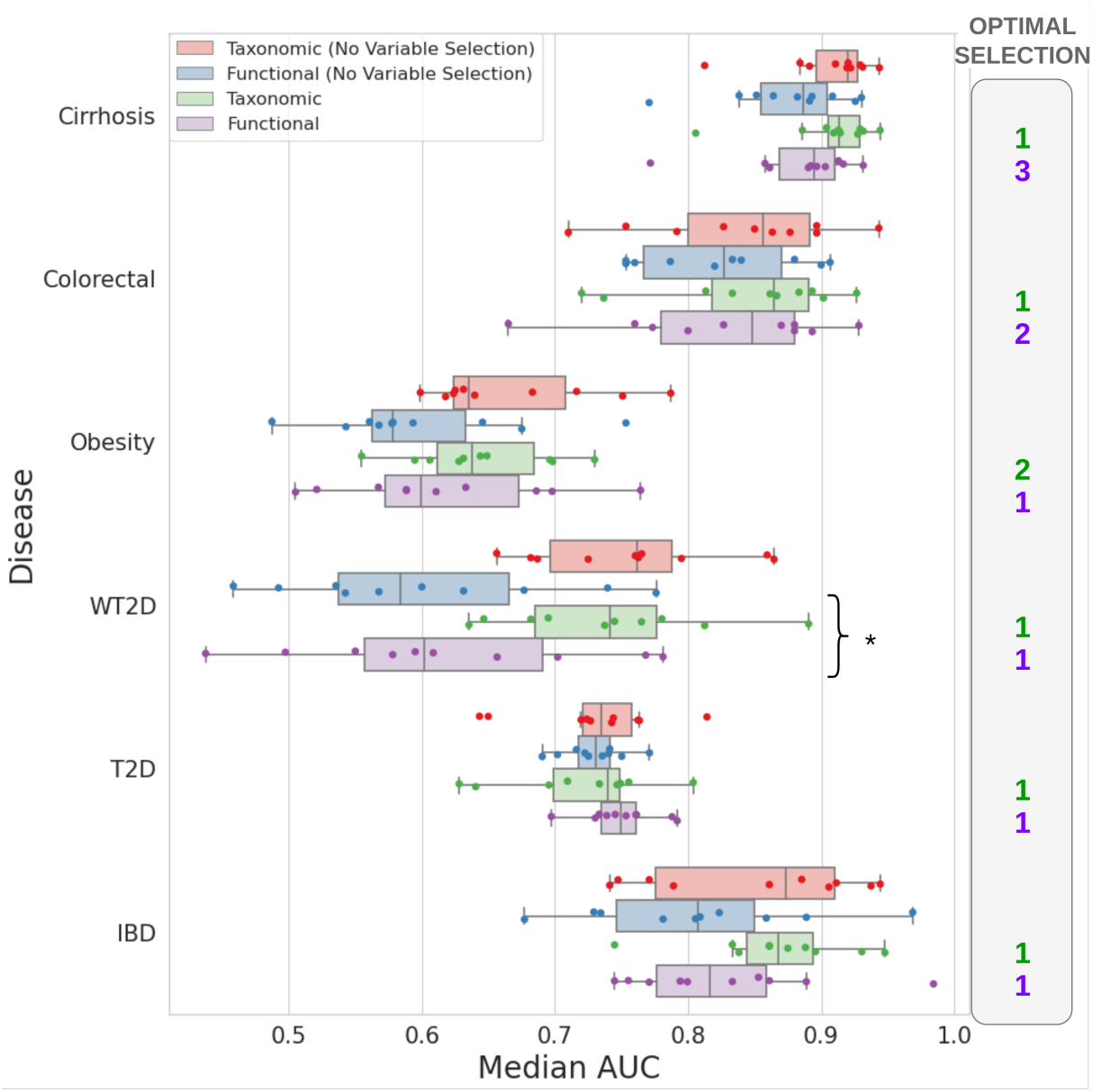


Figure 3.4 – **Classification performances of RF models trained on taxonomic and functional profiles, and impact of the variable selection on performance.** Median classification performances (AUC) for all types of profiles and each dataset, on the original datasets as well as at the optimal level of selection over 10 full runs of the pipeline. Each of these runs involved a different randomly selected test set of individuals, which was used for both profiles. Performances and importance scores for each run were computed and averaged over 20 distinctly trained RF models. The amount of selection iterations required to obtain the best average among these median AUCs are represented beside each plot. Instances when the difference in performance between functional and taxonomic profiles after variable selection is significant for a same dataset (based on a Mann-Whitney U-test) are signaled by a * symbol.

Dataset	**Cirrhosis**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.88675	0.91985	**0.91985**	0.952675	0.890175	0.93345	0.9137	0.951775	0.893775	0.9366	0.91075	0.93975	**0.89445**	0.932975	0.901075	0.90985	0.88135	0.91395	0.8995	0.903325
Dataset	**Colorectal**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.826675	0.814975	0.856675	0.86	0.840825	0.87	**0.864175**	0.886675	**0.84835**	0.906675	0.838325	0.89665	0.83	0.92165	0.8083	0.85335	0.7567	0.8825	0.713325	0.829975
Dataset	**Obesity**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.577875	0.566925	0.6355	0.626475	0.59955	0.6435	**0.630475**	0.6292	0.58525	0.74285	**0.637625**	0.717175	0.5602	0.76135	0.6191	0.677625	0.529475	0.6673	0.5724	0.643125
Dataset	**WT2D**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.58385	0.568225	**0.761525**	0.746225	**0.6018**	0.7336	0.7411	0.7879	0.57925	0.8207	0.733325	0.82325	0.54865	0.81945	0.6479	0.7677	0.461875	0.798	0.583375	0.724775
Dataset	**T2D**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.7304	0.69835	0.734675	0.720575	**0.749375**	0.75985	**0.74005**	0.72605	0.742575	0.777325	0.711325	0.756	0.689925	0.73885	0.608225	0.697875	0.681075	0.717025	0.5854	0.663875
Dataset	**IBD**																			
Iteration	0				1				2				3				4			
Profile	Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic		Functional		Taxonomic	
Measurement	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
AUC	0.807325	0.836775	**0.87325**	0.88825	**0.8164**	0.9412	0.86805	0.94705	0.7882	0.94705	0.8073	0.94705	0.79295	0.954375	0.756	0.885325	0.7888	0.92645	0.7208	0.833825
Average ratio: validation / test	0.99501		1.0005		1.0904		1.0335		1.1752		1.0975		1.2134		1.1106		1.2364		1.1374	

Table 3.1 – **Evolution of the average median AUC scores per dataset, on the validation and test sets, at increasing levels of variable selection, for taxonomic and functional profiles.** Column 0 shows performances obtained before variable selection for each profile. The top-performing selection levels on the test sets are highlighted in bold. The bottom row records the average ratio between performances obtained on the test and validation subsets.

3.2.2 Variable selection for a more tractable amount of information to explore.

The datasets used in the previous section contained on average 484 taxa. Through EsMeCaTa's [67] pipeline and its interrogation of UniProt [68], these taxa were linked to a total average of 10,510 FAs per dataset, resulting in a 22-fold mean increase in the amount of information, as shown in Table 3.2. For example: in total, the sequenced samples of the Cirrhosis dataset covered 542 taxa, which were associated by EsMeCaTa to a total of 10,434 FAs. Following the application of our method, 72 of these taxa and 33 of these annotations were included in the candidate sublists. Among these, 32 taxa and 7 annotations were in the confident subset, and 23 taxa and 4 annotations were in the robust subset.

This dimensional increase is counterbalanced by a selection of variables based on RF importance scores. These scores, when ordered from highest to lowest, display a kink-like shape. Selection is done by automatically operating a cut-off at the inflection point of the kink and probing whether classification performances are improved (see Section 3.1.2). This selection aims to correct the redundancies and the dimensionality of the original dataset for better classification. It also generates one of the pipeline's main outputs: a list of ranked features (either taxa or FAs) based on their average importance scores,

Dataset	Features	Initial number of taxa	Predicted Functions	Robust subset	Confident subset	Candidate subset
Cirrhosis	Taxa	542	-	23	32	72
	FAs	-	10,434	4	7	33
Colorectal	Taxa	503	-	24	37	109
	FAs	-	10,635	1	17	355
Obesity	Taxa	465	-	136	154	188
	FAs	-	11,341	26	169	3,199
WT2D	Taxa	381	-	27	51	136
	FAs	-	10,180	8	69	3,150
T2D	Taxa	572	-	117	136	202
	FAs	-	10,275	139	307	1,575
IBD	Taxa	443	-	22	29	100
	FAs	-	10,196	59	167	1,883

Table 3.2 – **Application of the presented selection process to identify signature taxa and functions on 6 reference datasets.** Total amount of features (taxa and FAs) in the original dataset ("Initial Number" column) and in the robust, confident, and candidate selections at the optimal selection threshold (Calculated over 10 runs of the pipeline).

and including an automatically computed cutoff that distinguishes discriminating and non-discriminating information.

The amount of information retained per run for all functional datasets is illustrated in Figure 3.11 (in purple). The figure shows that the average amount of information to retain for optimal classification performance varies depending on the dataset. For instance, the IBD dataset approximates our approach with the top 500 annotations ranked by average Gini importance, whereas the Obesity dataset requires the top 1,000 annotations for a comparable selection. This underscores the advantage of an adaptive method over a fixed threshold, as it adjusts to problem complexity. Additionally, our method's selection thresholds diverge from traditional thresholds, such as the top 30 features explored in Jones et al. [98], offering insights into optimal information consideration for discerning microbiota profiles.

In summary

The variable selection included in our approach also highlights robust markers of each of the tested diseases in the gut microbiota. The size of these marker lists adapts automatically to fit the problem instead of relying on a pre-established threshold, and is much more manageable than the initial profiles as an entry point for biological interpretation.

3.2.3 Timed benchmarks of the repeated and iterated classification and selection process.

Table 3.3 records the time taken to execute the process on each dataset. Overall, excluding EsMeCaTa which is variable because it is dependent on the quality of the available network (see separate benchmarks in Section 2.2), the pipeline takes between 25.7 hours (92,549.7 s, Colorectal dataset) and 45.8 hours (16,4972.2 s, WT2D dataset) to complete on a calculation cluster, with 10 CPUs and 100 GB of RAM at disposal, when using the default Gini metric for the calculation of importance scores.

	Cirrhosis	Colorectal	Obesity	T2D	WT2D	IBD
Total amount of samples	232	121	253	344	96	110
Total amount of taxa	542	503	465	572	381	443
Functional score calculation time (s)	598.445459	385.065978	536.508328	970.565677	249.550255	334.447807
Average time per run (s)	9489.6373645	9216.4661335	9865.6487073	10333.1304005	16472.2608586	9383.6936185
Total	95494.819104	92549.727313	99192.995401	104301.869682	164972.158841	94171.383992

Table 3.3 – **Runtime of the application of our method to all datasets.** The total run times include data formatting, functional score calculation, 10 runs of iterative classification (5 iterations, 20 forests per iteration), and post-processing. The runtime of the EsMeCaTa pipeline is not included.

The time taken to calculate the functional scores scales with the dimension of the information given as input. That is however not the case of the average time per run, with the WT2D dataset in particular standing out as having the longest run time in spite of being the dataset with the least amount of information (381 taxa over 96 samples). This could indicate that SPARTA does not perform as well when the amount of information is under a certain threshold, or it could be indicative that the WT2D classification is a particularly complex case. Further studies on low-dimension datasets should be envisaged to answer this.

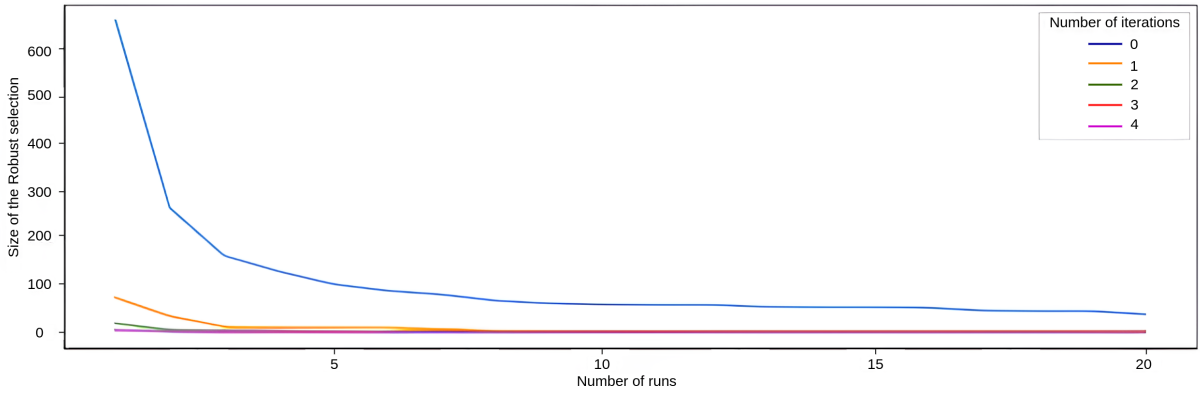
3.2.4 Impact of the iteration and repetition of the method.

Our results highlight the need to perform at least one iteration and several repeated runs to reduce the dimensionality of the functional datasets, while maintaining the classification performance, and derive a list of robust variables. The number of required iterations depends on both the dataset and the user needs in terms of classification performance and interpretability. Figures 3.5 and 3.8a illustrate the impact that a variation of these parameters can have on the results. When it comes to iterative selection, Figure 3.8a showcases that the first selection is always by far the most important, and there is little variation in selection sizes past the second selection. Therefore, 2 selections could also be perceived as an upper limit, though some of our datasets have shown better classification performance beyond this level of selection. Figure 3.5 illustrates, in the case of the IBD dataset, that the sizes of both the functional and taxonomic robust selections stabilize and hit a plateau after only a few runs. In both cases, 10 runs is sufficient to attain a stable content for the robust selection. As such, we presented results obtained over 10 runs, comprising 5 iterative selections each. These values were chosen as a compromise between execution time and robustness of the results. This conclusion could however only be attained *a posteriori*, once the results had been obtained. Someone aiming to apply the same method to their data may want to reduce the amount of operated runs, but should bear in mind that these results may vary depending on the dataset.

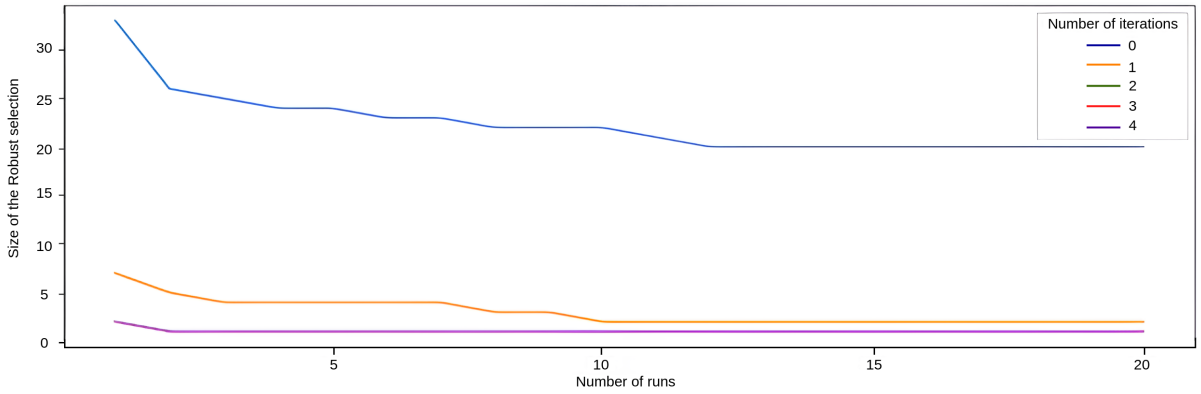
3.3 Exploring alternative approaches.

3.3.1 Impact of the TF-IGM scaling.

The functional scores on which the previous results were based were processed with the TF-IGM normalization, presented in Section 2.1.3. This manipulation exacerbates the scores of the most differentially expressed annotations, heightening their highest scores, and lowering their lowest, to facilitate classification. A caveat of this approach however is that, as a cost for making the profiles more discriminating, it can enhance biases inherited from the database or from the taxonomic profiling. In order to measure the impact of this process on our classification results, we trained RFs on functional profiles both non-transformed, and normalized with TF-IGM. Figure 3.6 shows the performances of all profiles in this context, for all levels of selection, over 5 runs. On this plot, we can see that the profiles normalized with TF-IGM are consistently marked as the better performing



(a) Functional profile.



(b) Taxonomic profile.

Figure 3.5 – **Sizes of the robust selections obtained at each iteration level on the IBD dataset (functional and taxonomic).**

functional profiles on average, yielding better top performances in all of the datasets aside from Obesity. Though none of these differences were significant according to a Mann-Whitney U-test, the top TF-IGM results are also shown to be more consistent, with smaller standard deviations in all datasets except for Colorectal. The combination of these factors has led to the adoption of TF-IGM normalization by default in all subsequent manipulations of the functional profiles in this context.

3.3.2 Impact of the variable importance measuring approach.

In our method, the variables' importances hold a central place, as they are the basis for the iterative variable selection. As such, though we relied on Gini importances by default, other such metrics could be considered for this calculation. One such other option is the SHAP importance [85], which calculates each variable's contribution to a decision from the basis of a trained classifier.

A reproduction of the previous results was obtained with a substitution of Gini by SHAP as implemented by the SHAP Python package [85], the results of which are presented in Figures 3.7, 3.8 and 3.9. Figure 3.7 shows that neither metric allows for better classification performance than the other, as both give very similar results. Comparison of the performances' distribution with a Mann-Whitney U-test confirmed this observation, as none of the results obtained with SHAP (in red) were found to be significantly different from those obtained with Gini (in green for taxonomic and purple for functional), as pictured on Figure 3.7 by the absence of an asterisk over all of the SHAP results. The sizes of the individual selections obtained with Gini and SHAP are consistently comparable, as shown by Figure 3.8a, with a few exceptions notably from the taxonomic profiles: the taxonomic Gini-based selections of Obesity and T2D's first selections are notably larger than their SHAP counterparts. In spite of this, Figure 3.8b shows that the contents of the lists tend to differ: the IBD dataset's first selections with SHAP and Gini were the most similar on average, with a mean similarity percentage of 60%. Of the two options, the contents of the Gini-based selections proved to be the most robust, consistently providing larger robust selections than the SHAP-based version of the approach, be it on the functional or taxonomic data, as shown by Figure 3.9. SHAP's robust selections are particularly small, only exceeding 10 features in the case of IBD and Cirrhosis' first functional selections. In both these cases, between 80 and 100% of the selected annotations, and between 50 and 100% of the selected taxons (the 50% case only accounting for 2 selected taxons) were also found in Gini's robust selection. Considering all of these results, Gini importance was kept

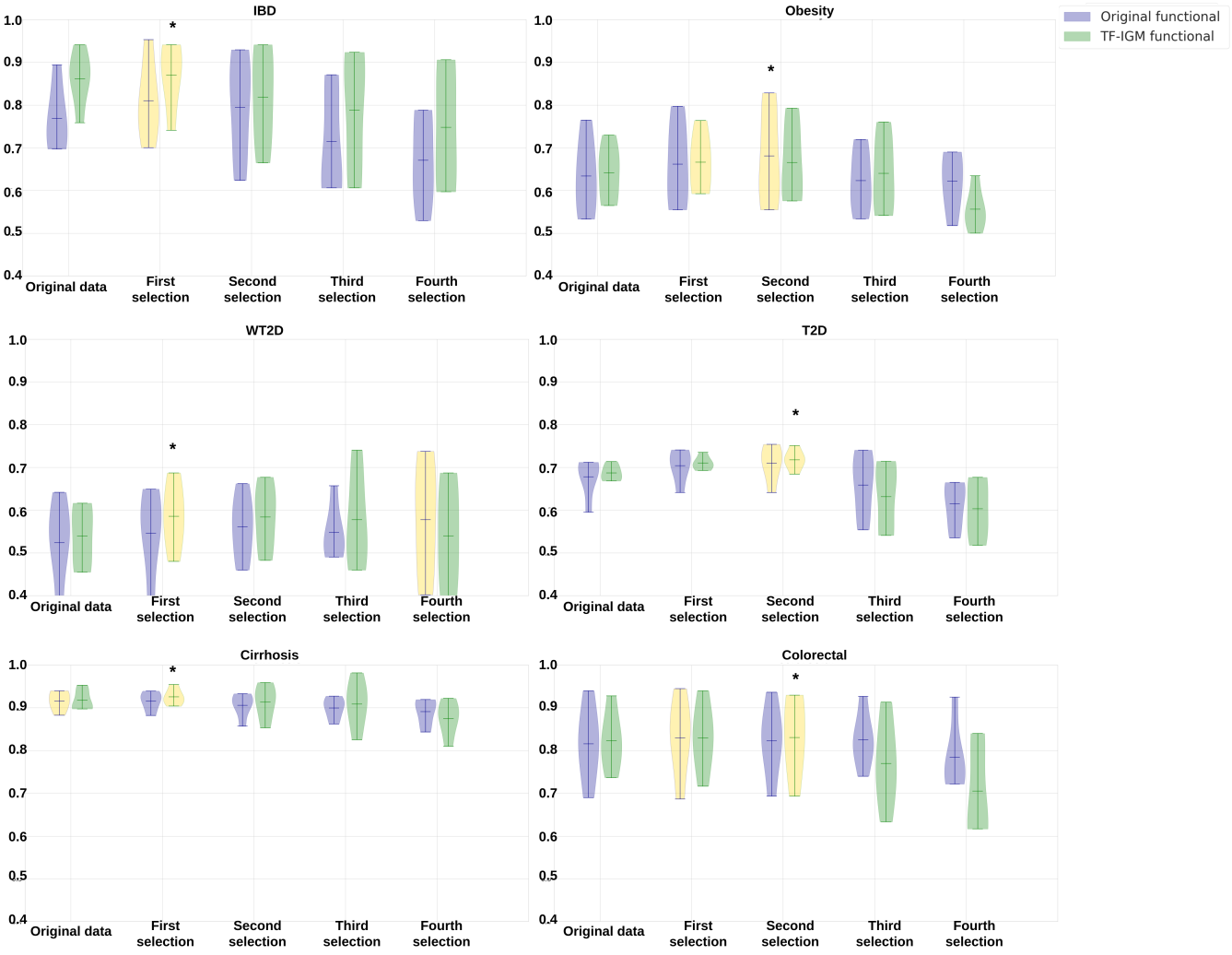


Figure 3.6 – **Classification performances of Random Forest models trained on functional profiles with and without TF-IGM scaling.** Similarly to Figure 3.4, the average classification performances (AUC) for all types of profiles and each dataset are represented, for all levels of variable selection, and over 5 full runs of the pipeline. For each run, 20 RF classifiers were trained, and the median AUC was retained. Both profiles were tested with identical test and validation sets. The top performance for each profile, meaning the one with the highest average, is plotted in yellow. Of both distributions drawn in yellow, the one with the highest average is marked with an asterisk.

as the default metric for our approach, as it proved to be more robust experimentally.

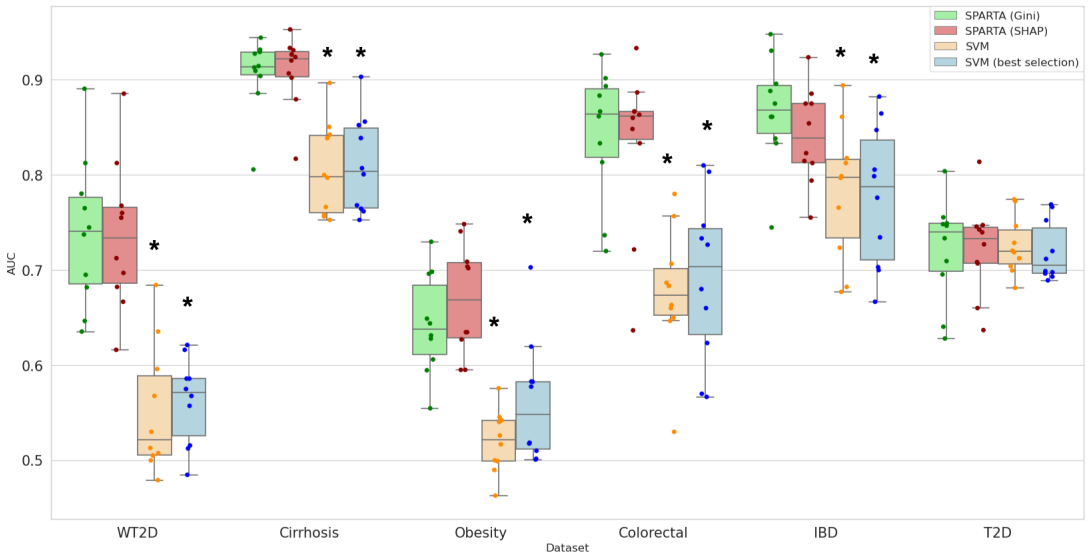
3.3.3 Impact of the classification algorithm.

Though RFs were consistently highlighted in the literature as one of the best adapted models for classification on the basis of the microbiota, SVMs have also shown good performance in such tasks [27, 87, 94]. As such, we compared the performances of SVM models applied to our data to those obtained with our RF-based approach, both on the full datasets and on the optimal selections derived from the iterative selections of our approach, using the same amount of iterations for each of which the same number of models were trained with the same test and validation sets as previously used to measure the performances of our approach. The results, plotted in Figure 3.7, show that SVMs always perform below our RF-based approach on average, for both taxonomic and functional profiles on all datasets. These differences are significant on all datasets aside from T2D when it comes to functional profiles, but are less pronounced in the case of functional profiles, where both SVM classifications perform significantly below their RF counterpart in the case of the Obesity dataset alone. However, the Colorectal and IBD datasets are also the only ones on which neither of the SVM functional classification tasks performed significantly below the RF-based method. Based on these results, RFs were considered to be the best adapted classification approach for our method, and were kept as the default for all of the results obtained afterwards.

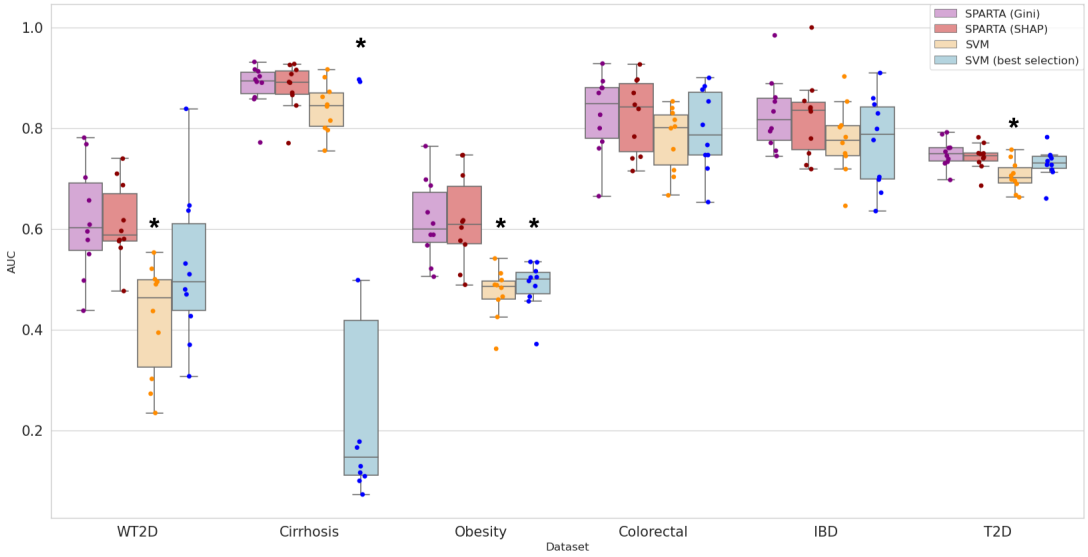
3.3.4 Comparison with sequence-based approaches: the impact on classification.

Throughout Section 2.2, the functional profiling of the IBD dataset by EsMeCaTa, which our previous results are based on, were compared to those obtained with HUMAnN3. This allowed us to highlight the benefits of the former, reference-based, approach over the latter, which is sequence-based, in terms of computation time and resources, for little deviation in terms of the outputs' informative contents. However, the question of whether one approach made for a better basis for classification compared to the other hadn't been explored then. One could expect HUMAnN, which consistently refers to the original sequences, to provide a more precise characterization of the functional microbiota, and therefore to be a more reliable basis for classification.

To answer this question, a comparative classification was made based on the functional

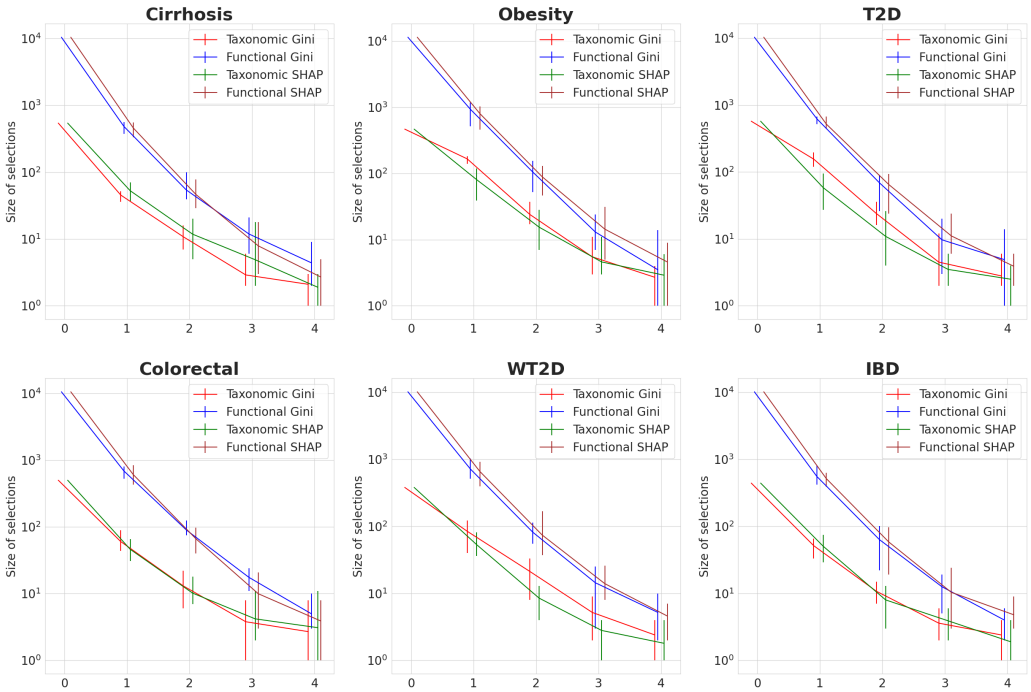


(a) Performances obtained on the taxonomic profiles.

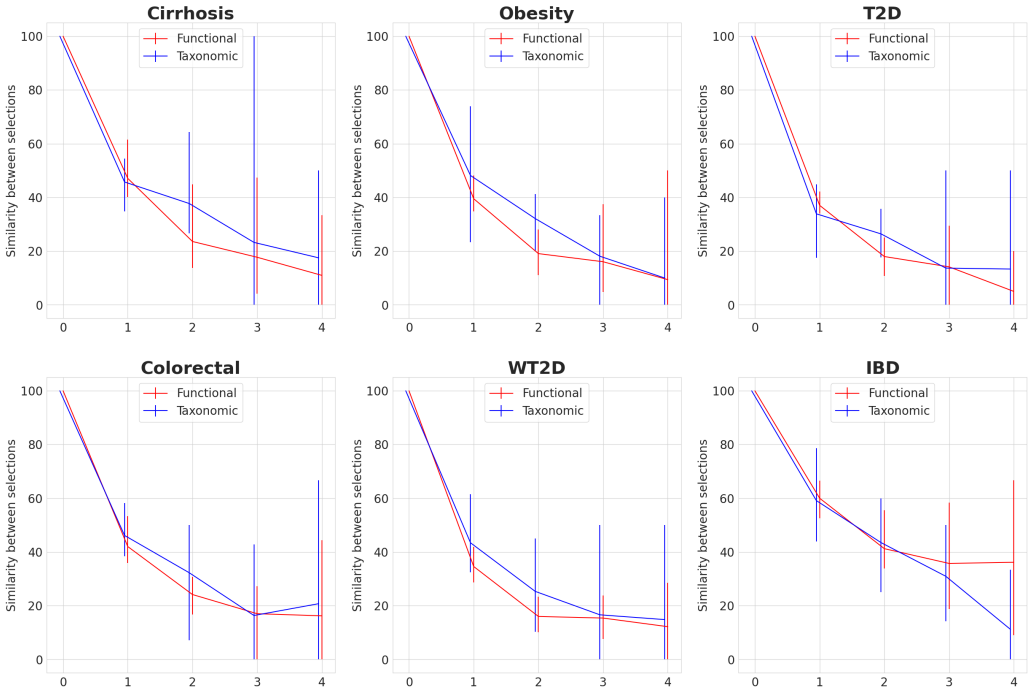


(b) Performances obtained on on functional profiles obtained via EsMeCaTa.

Figure 3.7 – **Classification performances obtained with RFs and automatic variable selection on all datasets (SPARTA), using RF-based selections based on Gini and SHAP, and using SVM classifiers on the full dataset and the best-performing selection in terms of classification for Gini-based RFs.** Similarly to Figure 3.4, the represented performances for the SPARTA (Gini, green for taxonomic and purple for functional, and SHAP, red) classifications are the median classification performances (AUC) for all types of profiles and each dataset, at the optimal level of selection over 10 full runs of the pipeline. SVM performances were obtained over a single run and were applied to the entire dataset (orange) or to the variable selections that correspond to the best performances for SPARTA Gini (blue). Performances obtained with SPARTA SHAP and SVMs were compared to those obtained with SPARTA Gini with a Mann-Whitney U-test. Those marked with a * showed a significant difference in distribution (p-value < 0.05). Consistent test and validation sets were used between all profiles for the classification tasks.

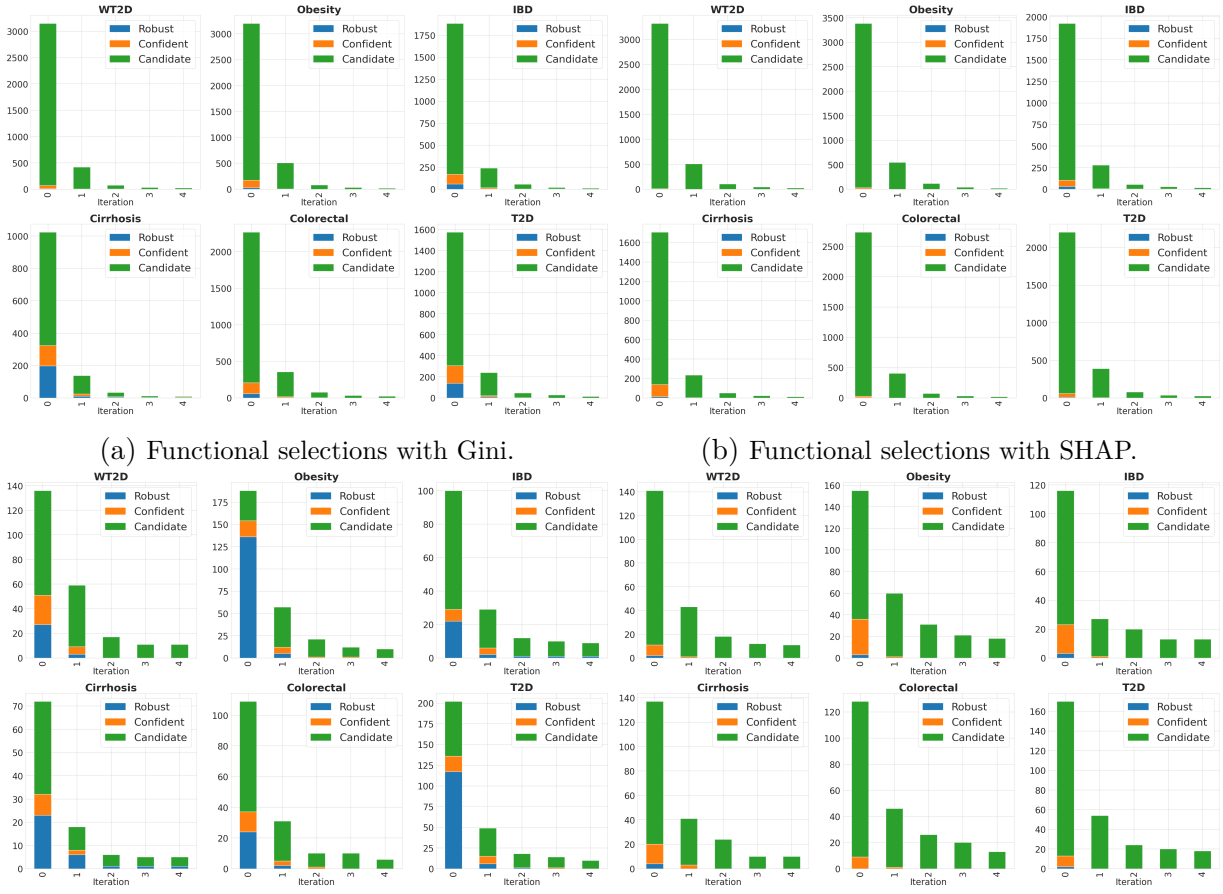


(a) Sizes of the functional and taxonomic selections obtained with Gini and SHAP over 10 runs with 5 selective iterations, for all datasets.



(b) Similarity percentage between the individual Gini and SHAP selections, for functional and taxonomic profiles.

Figure 3.8 – **Sizes and similarity of the individual Gini-based and SHAP-based selections.**



(a) Functional selections with Gini.

(b) Functional selections with SHAP.

(c) Taxonomic selections with Gini.

(d) taxonomic selections with SHAP.

Figure 3.9 – **Sizes of the robust, confident, and candidate selections obtained on each dataset over 5 iterations of our variable selection method, using Gini and SHAP.**

profile built from the raw reads of the IBD dataset with HUMAnN3 [34] which was previously described in Section 2.2, using the same parameters. During the process, sample V1.UC-19 could not be processed properly, resulting in a functional table devoid of this sample. As such, the performances obtained on this profile were compared to classification performances obtained on an IBD functional profile built by EsMeCaTa UniProt without the sample in question. The obtained results are presented in Figure 3.10 and show that median classification based on functional profiles built directly from the reads are on par with those obtained using EsMeCaTa, as the differences in performance are not significant based on a Mann-Whitney U-test (p-value = 0.45). Both functional profiles' performances are also non significantly different from the performance obtained on the IBD taxonomic dataset (p-value = 0.73 for HUMAnN and 0.36 for EsMeCaTa).

As such, in spite of our initial instincts, we found that EsMeCaTa and HUMAnN perform comparably as a basis for classification performance. However, processing patients samples with HUMAnN3 resulted in an over thirteen-fold increase in terms of computation time, and required handling inputs of 442 GB, compared to EsMeCaTa's 302 kB entry (see 2.1.4). Due to limitations in resources, we could only process one of the datasets' raw reads, namely IBD. In order to solidify these conclusions, this comparison should be applied to other datasets. However, these first results further illustrate how a reference-based method like EsMeCaTa is capable of performing on par with a state-of-the-art sequence-based approach like HUMAnN3. On the basis of these results, EsMeCaTa was chosen to be the default method of annotation for our next manipulations, as it was the fastest and lightest approach available when we already had taxonomic profiles at our disposal without compromising performance.

In summary

Alternatives to the scaling method (TF-IGM scaling or no), variable importance metric (Gini or SHAP) and classification algorithm (RF or SVM) were also tested. Results illustrated the benefits of the parameters chosen for our approach, in terms of performance and/or robustness.

Surprisingly, exploration of an alternative to our functional profiling method, namely HUMAnN3, also showed that the sequence-based approach performed on par with the functional profiles built from the reference-based EsMeCaTa approach.

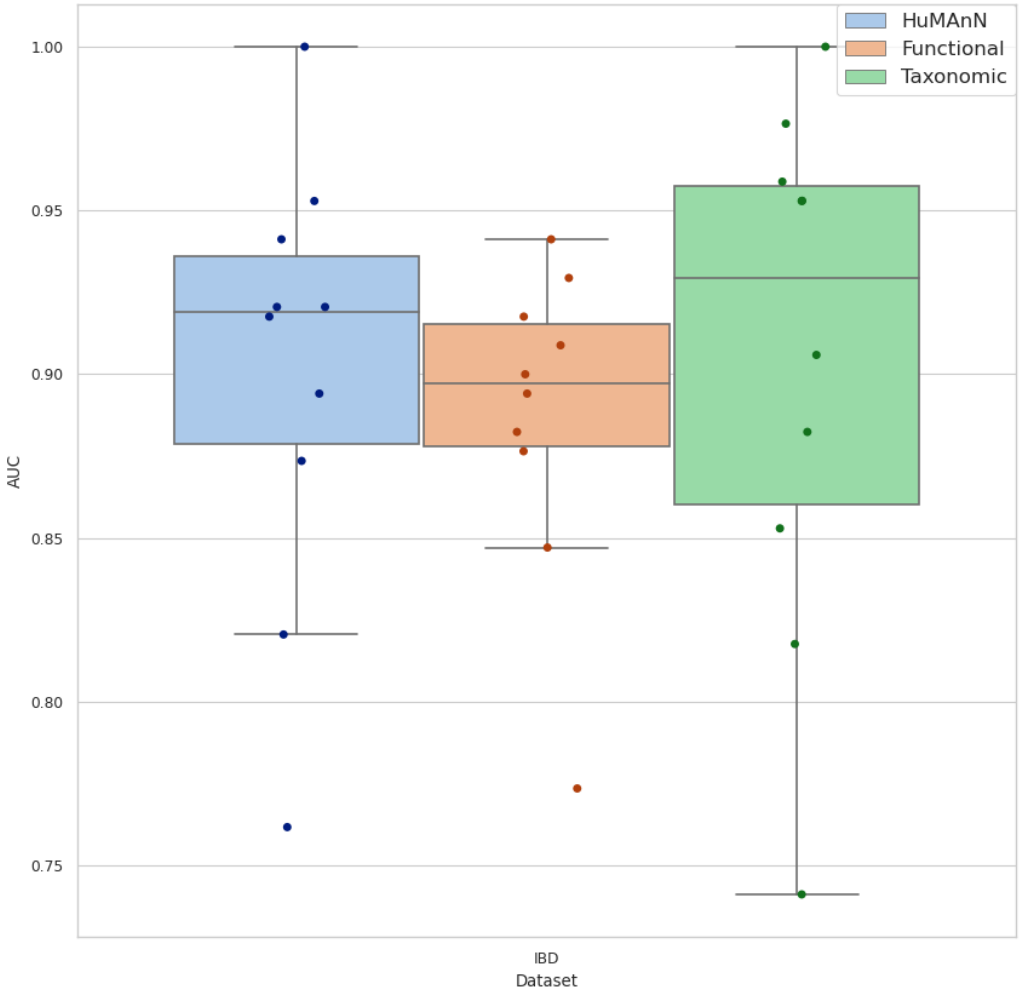


Figure 3.10 – **Classification performances obtained on the IBD dataset (minus sample V1.UC-19), annotated with EsMeCaTa (orange) and HUMAnN3 (blue), as well as on the taxonomic dataset (green).** Consistent test and validation sets were used for between all profiles for the classification tasks.

3.4 Exploring the advantages of a non-linear approach to variable selection.

3.4.1 Our non-linear approach is a more consistent selector of variables than limma.

To explore whether our approach to variable selection differs from classic linear approaches, we compared our approach with a standard method designed for continuous data [104] rather than for count data [103]. Specifically, selections obtained from direct pairwise comparison of the profiles using the limma tool [104] were used as basis for comparison. Variables were selected using a p-value threshold of 0.05, a classic threshold value exploited in several other studies that applied limma to metagenomic data [105–107]. Similarly to our previous manipulations, the selection process was iterated 10 times with variation induced from setting aside a subset of the samples, and variables were compiled into 'robust', 'confident', and 'candidate' categories depending on how often they were selected. The test sets put aside for limma were the same as those used to obtain the results of Section 3.2. Comparative results of this process are presented in Figure 3.11 and Table 3.4. For example, Figure 3.11 shows that, when applied 10 times to the Cirrhosis dataset, our approach selects a minimum of 6 annotations, and a maximum of 21, with a median of 11. In the same conditions, limma selects between 1,032 and 2,149 annotations, for a median of 1,642. These distributions are plotted, respectively, in purple and gray. Table 3.4 shows that with our selections, the Cirrhosis dataset outputs 4 robust annotations, 7 confidents, and 33 candidates, against a respective 878, 1,165 and 2,668 with limma. With these parameters, limma is the most stringent selector on all datasets aside from Cirrhosis. For the Colorectal, WT2D and Obesity datasets in particular, all selections are empty, leading to an empty candidate subset as described in Table 3.4. The IBD dataset also proves to be unsuitable for this approach, yielding empty robust and candidate subsets. Only the T2D and Cirrhosis datasets allow limma to yield a non-empty robust subset. On the other hand, our approach consistently yields non-empty robust and confident selections, both of which are reasonably sized for interpretation when compared to the candidate subsets, being close to 50 times smaller in the case of the WT2D dataset's confident and candidate subsets.

Among these datasets, Cirrhosis stands out as an outlier. Indeed, it is by far the dataset on which limma selects the most information: in Figure 3.11, we can see that

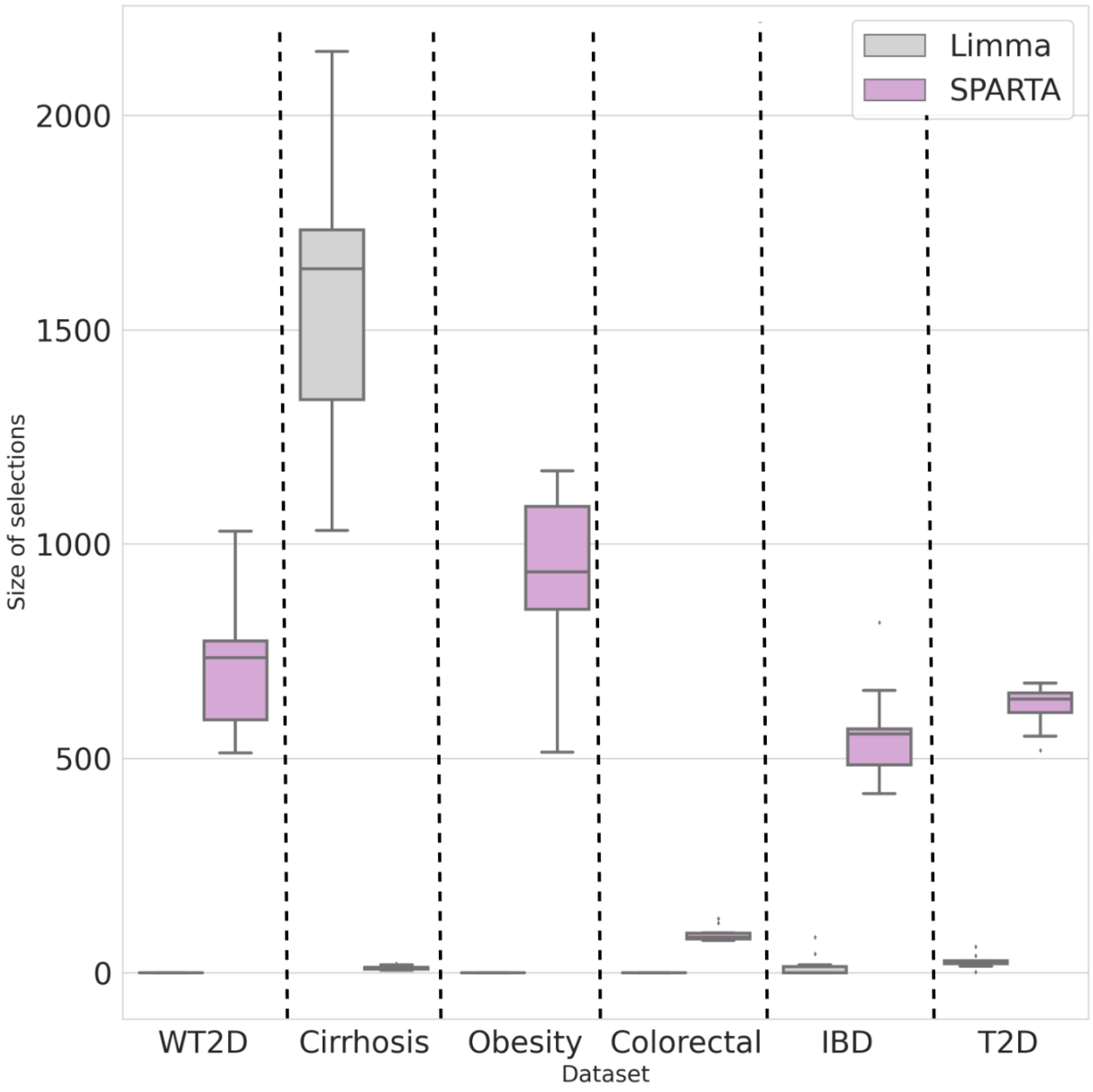


Figure 3.11 – **Number of important selected FAs for each run at best iteration for the six datasets** Amount of FAs selected by our approach (SPARTA) and limma, for all datasets. Limma selections were effectuated with an adjusted p-value threshold of 0.05. Both selection methods were repeated 10 times, with a common test subset set aside each time.

	Total size of the robust subset		Total size of the confident subset		Total size of the candidate subset	
	Iterative Gini-based selection	Limma	Iterative Gini-based selection	Limma	Iterative Gini-based selection	Limma
Cirrhosis	4	878	7	1,165	33	2,668
Colorectal	1	0	17	0	355	0
Obesity	26	0	169	0	3,199	0
WT2D	8	0	69	0	3,150	0
T2D	139	2	307	4	1,575	103
IBD	59	0	167	0	1,883	111

Table 3.4 – **Sizes of the selections obtained using limma and our iterative approach based on variable importance.** Limma was applied with an adjusted p-value threshold of 0.05. From left to right, the columns present, for each approach, the size of the robust, confident, and candidate subsets issued by the concerned selection method iterated 10 times with identical test subsets.

it selects 1550 annotations on average over 10 iterations, whereas the second highest amount, obtained with the T2D dataset, is only 26.1 on average. This also makes it the only case in which our variable selection approach proves to be the most stringent of the two, with an average of 12 selections per run, for a robust selection of size 4 against limma's 878 (see Table 3.4). The four annotations in question are: GO:0016984 (ribulose-bisphosphate carboxylase activity), GO:0003779 (actin binding), GO:0004081 (bis(5'-nucleosyl)-tetraphosphatase (asymmetrical) activity) and GO:0018112 (proline racemase activity). Actin binding (GO:0003779) signals the participation of the gut in the maintenance of the intestinal epithelia, which plays a role in the prevention of liver diseases such as Cirrhosis [123]. The activity of proline racemase (GO:0018112) is also indicative of proline metabolism in the gut, which has also been shown to be upregulated in cases of Cirrhosis [124]. The activity of the bis(5'-nucleosyl)-tetraphosphatase enzyme (GO:0004081) is involved in the metabolism of both purine and pyrimidine according to KEGG [56], which are disturbed in mice gut during the development of Cirrhosis [125]. Finally, ribulose-bisphosphate carboxylase (GO:0016984), though it is mostly known for its role in photosynthesis, can also be involved in the salvage of methionine [126], itself key in the development of liver disease [127].

As such, in the case of Cirrhosis, our approach robustly highlights a small subsection of biologically relevant annotations, themselves consistently highlighted by limma as linear indicators of the prevalence of the disease. This could illustrate a case in which the dataset is "too easy" to predict, due to an abundance of features that linearly differentiate the profiles, and a small sample of which is sufficient to be efficient in classification. This could lead to an over-selection from our performance-regulated approach, as even when relevant features are removed by the iterated selection, the remaining variables still allow for good

classification performance. In this case, it could be interesting to look at the selections from iterations before the optimum.

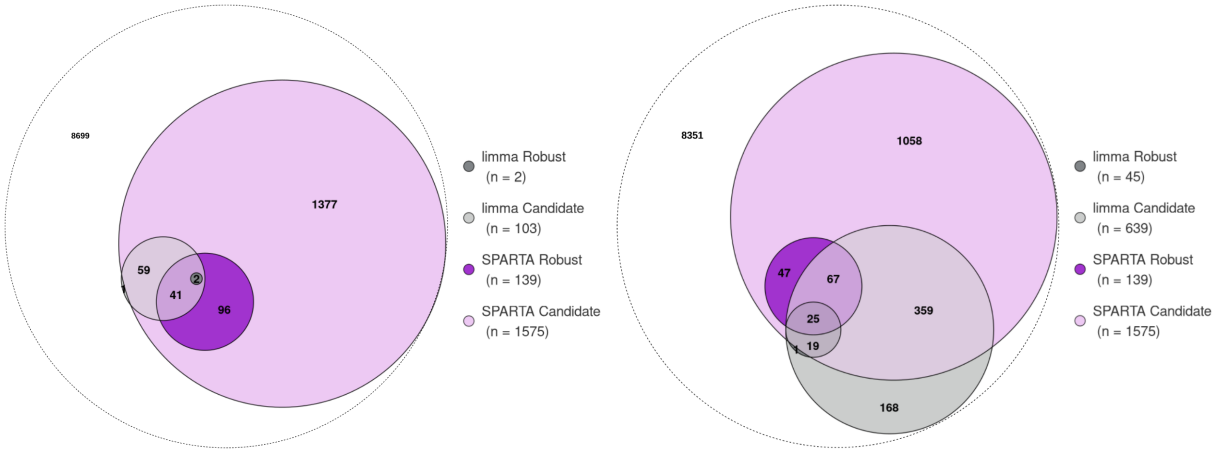
3.4.2 Non-linear approaches select linear factors, and more.

We then focused on the T2D dataset, which is the only other dataset on which limma i) extracts a non-empty robust selection with an adjusted p-value threshold of 0.05 (see Table 3.4) and ii) consistently provides non-empty FA selections. Figure 3.12a illustrates the overlap between the robust and candidate annotations selected by both approaches. T2D's limma selection is smaller than the one obtained through our method, englobing a total of 103 annotations in its candidate subset against 1,575 for the latter approach, as shown in Table 3.4. As shown by Figure 3.12a, all of these annotations aside from one are included in our approach's candidate selection. Similarly, limma's robust subset is entirely included in our approach's robust selection.

To put these results in perspective, there is no guarantee that a 0.05 p-value threshold yields an 'optimal' selection for this dataset when applying limma. This choice of threshold is, however, a required external input for the method, that is not required by our approach as it automates the choice of the selection's size. As such, the chosen threshold could arguably be too restrictive for the T2D dataset. As an illustration, a p-value threshold of 0.255, obtained to generate a limma candidate selection as close as possible to that of our approach's selection, was applied, as illustrated by Figure 3.12b. This much less restrictive threshold yields a limma selection that still largely overlaps its counterpart, as 74% of limma's annotations are included among those selected by our approach.

3.4.3 Linear selection is less effective as an enhancer of classification.

Finally, in cases where the limma functional selections were non-empty, they were compared to the functional selections obtained by our approach as basis for RF classification. Performances obtained on both selections are presented in Figure 3.13, and show that limma's selections perform beneath our approach's as basis for classification, as neither of the recorded performances obtained on limma's selections surpass their counterparts. The difference is only significant in the case of T2D (Mann-Whitney U-test p-value: 0.045), however these results show that the linear approach, in addition to being less consistent and less thorough as a selector, also never surpasses its RF-based counterpart as a means



(a) Limma alpha = 0.05 (b) Limma alpha = 0.255

Figure 3.12 – **Comparison between robust and candidate FAs for the T2D dataset obtained with limma and our approach (SPARTA).** The limma subsets were obtained using the classic threshold of 0.05, and an adjusted p-value threshold of 0.255, chosen to obtain comparably sized candidate sublists between both selections. Values indicate the number of annotations in each intersection and do not represent the size of a category as a whole. The white circle includes all annotations from the full dataset.

to enhance classification performance.

In summary

An application of limma, a linear approach for variable importance ranking, did not yield exploitable selections with a classic p-value threshold on four of our six datasets. The examination of the remaining two datasets allowed us to illustrate how our selection and limma behave comparatively in different situations. In T2D's situation, the limma selection was smaller and largely overlapped the one obtained through our approach, with limma's robust subset notably being entirely included in our approach's robust selection. For the Cirrhosis dataset, our approach's selection was the smallest of the two, however, it remained coherent with what limma selected, and yielded information that is coherent with the biological question at hand. RF classification performances obtained on both selections also showed that limma's selections performed beneath our approach's as basis for classification.

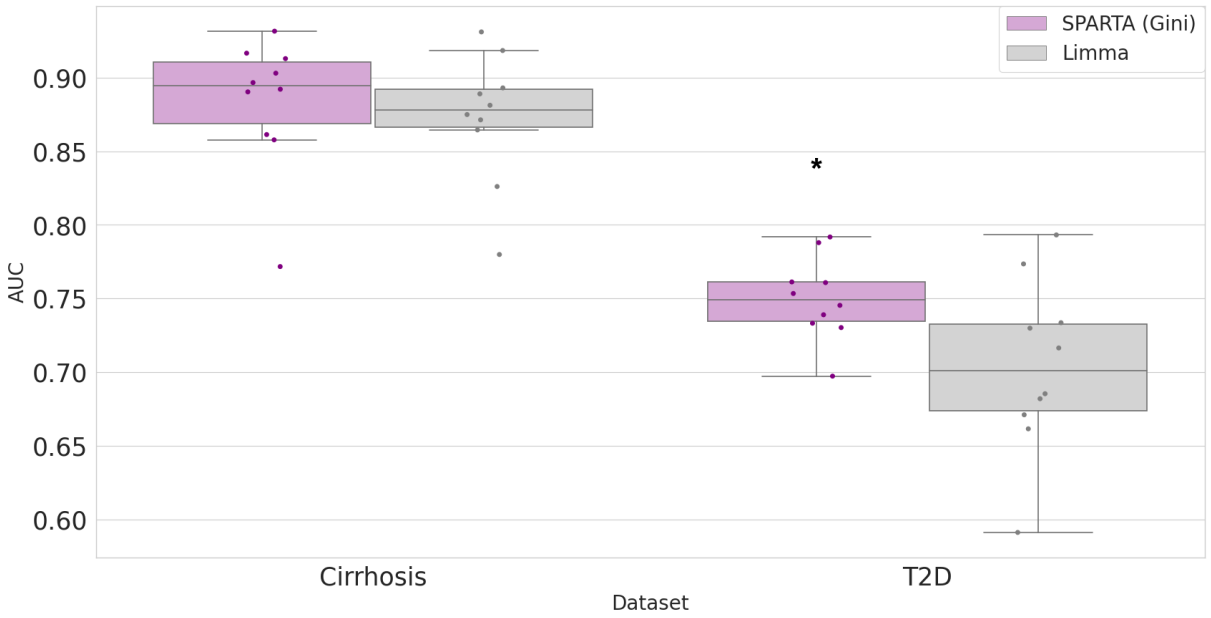


Figure 3.13 – **Classification performances obtained on the functional T2D and Cirrhosis datasets, based on our approach's best performing selection (SPARTA) and on a selection by limma (alpha = 0.05).** Performances per dataset were compared with a Mann-Whitney U-test. Those marked with a * showed a significant difference in distribution (p-value < 0.05).

3.5 Conclusion and discussion.

Through a novel method, we have explored the classification of individuals from their gut microbiota, described by both taxonomic and functional features. This approach differs from MetAML [87] and DeepMicro [94] by introducing repetitions of the training process, itself involving an iterated variable selection process on the basis of re-trained classifiers. The method introduces a test set to evaluate the final performance of the model for each run, and validation sets to induce variability in the training conditions of the RFs and derive a more robust variable ranking for selection.

Application of the method in question has allowed us to evaluate that the translation of the microbiota into functional profiles gives non-significantly different performances when compared to microbial profiles on 5 of 6 datasets. It has also shown that by adapting the dimensionality of the problem through an automated variable selection method, both types of profiles had similar potential as descriptors of an individual's health status, yielding comparable classification performances when used as basis for model training. In this context, variable selection as applied in the method was shown to boost performances, especially in the case of functional profiles. The method we developed also focused on

robustness and interpretability, with one of its main outputs being a shortlist of robustly discriminating variables per dataset. These results were found to be more reliable and robust than those obtained with state of the art tools such as limma, giving a measurable insight on the impact of a feature on solving a complex problem.

Optimal selection: looking beyond performance.

Our approach to identify the best level of variable selection was to use the performances of the classifiers as reference to judge the informative contents of the sublist. Though this constitutes a strong basis for a first approach, performance alone as a criterion has been shown to be potentially deceptive [113]. As such, the literature recommends investigating the significance of the evolution in performance, however an evaluation of the informative contents of the selected sublists, leveraging Semantic Web for example, could also be envisioned to complement the performance criterion when it comes to selecting an optimal selection of variables.

Limits of the approach as a basis for diagnosis.

Classification performances in the context of FAs have been reported to be on par or slightly inferior to classification performances based on taxa [98, 99]. This is also consistent with our observations. As a result, current FA-based approaches might not be best used for direct diagnostic prediction. The conditions in which a sample has been obtained, sequenced and processed most likely impacts classification performances, even for the same disease (see the differences in performance obtained on T2D and WT2D in Figure 3.4). The main advantage of current FA-based pipelines, including our approach, lie in the extraction of a robust list of important FAs related to a dataset of interest, rather than the production of a generic, directly reusable ML model.

Exploring performance based on other sequencing methods.

The results presented here were all obtained on MGS data, which was publicly available and allowed us to position ourselves in comparison with previous studies which had also used them for benchmarks. A transition to 16S data could, however, potentially have an impact on performances: a comparison of disease state classification based on MGS and 16S sequences from the same samples have shown 16S data to be an overall more efficient support for classification performance [99]. This was,however, done without the model re-training, variable selection iteration and process repetition that we implemented in our approach. Seeing as EsMeCaTa is compatible with both MGS and 16S entries, further exploration of applications of our approach to outputs of the latter technology could make for a strong complement to our results.

A robust shortlist of selected variables as a support for downstream biological interpretation.

This approach's main strength arguably resides in its capacity to strip down a massive amount of information to highlight variables that robustly characterize the issue at hand. As such, the obtained shortlists open up opportunities for biological interpretation and exploitation down the line. The example of the Cirrhosis dataset's robust selection, discussed in 3.4.1, constitutes a first illustration of this, but a more thorough biological discussion should be possible from this basis.

Chapter 4

An exploration of taxonomic and functional profiles' complementary biological significance.

The gut microbiota and its influence on host health has been a major subject for medical research in recent years, as the interactions between the microbial ecosystem of the gut and its host were found to play a significant role in several disorders and diseases, ranging from colorectal cancer to anxiety and depression [16, 128, 129]. These diseases can often be characterized by a dysbiosis, meaning an abnormality in the balance of the gut microbiota's composition. For example, a decreased diversity in the gut microbiota of IBD patients has been observed, due notably to a lack of bacteria from the *Firmicutes* group, and an overabundance of taxa from the *Bifidobacterium* group [16, 130]. From these observations, several taxa have been proven to be influential on host health, opening up perspectives for treatment or prevention through levers such as diet or probiotics [17].

As previously mentioned, voices within the medical community in recent years have called for a shift in paradigm when it comes to the analysis of the gut microbiota, arguing that knowledge of the microbiome on the functional scale is a prerequisite in order for more advanced therapeutic options to be developed [29]. Some metabolic pathways were discovered as vectors for the microbiota's influence on host health, mostly through the involvement of known important taxa. These pathways notably include the synthesis of short-chain fatty acids, bile acid metabolites, lipopolysaccharides or indoles and indole derivatives [16, 17]. These components, through interaction with epithelial receptors or transport by the bloodstream, are involved in biochemical pathways that regulate biological functions both locally and in other organs, such as inflammation, energy harvest and storage, or hormonal balance. As such, the functional expression of the gut microbiota has repercussions on general host health, and knowledge of potential deficiencies on this scale could open possibilities for therapies based on direct compensation through targeted

interventions, involving bio-engineered commensal bacteria or targeted drugs for example [17].

However, deriving metabolic pathways from influential taxa has its limits. Notably, as is mentioned by Heintz-Buschart and Wilmes [29], because there is a lot of functional redundancy within the gut microbiota. As such, taxa that are hardly noticed because they are interchangeably present within the gut populations of individuals that exhibit similar characteristics, could have an unsuspected cumulated influence through a metabolic trait that they have in common. As such, there is a need to study archetypes derived directly from functional descriptions of the microbiota.

The results of the methods described in the previous chapters allowed us to build traceable functional profilings of the gut microbiota from taxonomic abundances (Chapter 2), and build reduced lists of significant descriptors from each profile (Chapter 3). In this chapter, we will rely on this information to explore the validity and potential for novelty that can be extracted from these results, by exploring their implications in regard to the biological literature and interlinking both profiles.

4.1 Presenting the detailed robust shortlists of the IBD dataset.

In this section, we will describe and explore the biological coherence of the robust sublists obtained on the IBD dataset using the method described in Chapter 3, applied to the taxonomic profile as well as the functional profile annotated with EsMeCaTa annotation through UniProt. These results come from the pipeline's first iteration, which are the best performing selective iterations for both profiles (see Figure 3.4). The IBD dataset was chosen as an illustrative representative of our results, as it is an outlier in neither classification performance, being the third best performing dataset out of six, nor in the selection of variables by limma (see Section 3.4.1).

An important output of the pipeline is the shortlist of robust variables that are selected by the method, allowing for downstream interpretability. This comes in the form of tables of robustly significant annotations and taxa, as previously described. The annotation shortlist for the IBD dataset is given in Table 4.1. It contains 59 FAs, alongside extra information describing the annotations and their status in the analysis. For example, annotation GO:0006520, corresponding to the amino acid metabolic process, is first in the table because it has the highest average Gini importance score over all 200 forests

trained at this selection level, over 10 runs. It is on average 1.05 times as present in the diseased profiles as it is in the controls, the negative value of the 'Ponderated average ratio' meaning that the annotation is predominantly found in unhealthy samples. It is expressed by a total of 358 taxa over all samples, of which 20 were found to be robust. The subsequent bibliographic analysis of this list (see 4.2) graded its relevance to the disease as a 1, meaning that there is a known direct link between the annotation and IBD [131].

A similar selection of robustly discriminant taxa is also available as an output of the pipeline, with the IBD output given as an example in Table 4.2. 22 taxa are presented, along with the same information as the previous table aside from the bibliographic categories. For instance, *Alistipes finegoldii*, identified in our process as Organism 73, similarly ranks first because it has the highest Gini importance score on average over all trained RFs. Its differential expression shows that it is expressed on average 16 times as much in control profiles as it is in the unhealthy samples. As previously, we can establish which annotations are attached to each taxon, with *A.finegoldii* expressing a total 1,220 FAs, 15 of which are robustly significant.

In summary

By applying Chapter 3's method to the IBD dataset, we extracted two lists of robust important variables: one taxonomic, and one gunctional. The differential expression of these variables between unhealthy and control profiles gives a first insight on their influence on microbiota health. Thanks to the explicitation of the links between taxa and annotations by EsMeCaTa, the relationships between these taxonomic and functional signatures can be explored.

4.2 Methodology for the evaluation of a feature's biological relevance.

Having extracted robust significant information from the IBD dataset, the question that arises is whether this selection is coherent with what is already known of the disease's signatures at the level of the gut microbiota. In order to explore this question exhaustively, we conducted a thorough bibliographic exploration of the obtained robust functional selection.

ID	Names	Average RF importance	Ponderated average ratio (Control/Unhealthy)	Number of linked taxa	
				Total	Robust
GO:0006520	amino acid metabolic process	4.37E-03	-1.04801771712759	358	20
4.1.2.-	Aldehyde Lyases	4.01E-03	-1.83063815612961	28	2
GO:0102545	phosphatidyl phospholipase B activity	3.53E-03	-4.59984365662854	15	1
GO:0004122	cystathionine beta-synthase activity	3.42E-03	-3.75076174596704	8	1
GO:0008744	L-xylulokinase activity	3.24E-03	-5.44424049313829	4	1
GO:0047419	N-acetylgalactosamine-6-phosphate deacetylase activity	2.57E-03	-1.19907351364782	78	4
GO:0008788	alpha,alpha-phosphotrehalase activity	2.44E-03	-2.30364417355582	19	1
GO:0032440	2-alkenal reductase [NAD(P)+] activity	2.43E-03	3.17446593793351	5	1
GO:0001510	RNA methylation	2.40E-03	1.05457228463169	249	17
GO:0015444	P-type magnesium transporter activity	2.34E-03	-1.65841492481138	66	2
GO:0016832	aldehyde-lyase activity	2.24E-03	-1.12570888096648	200	12
GO:0047605	acetolactate decarboxylase activity	2.23E-03	-1.56525594318597	48	1
GO:1901135	carbohydrate derivative metabolic process	2.18E-03	-1.10067892552244	271	14
GO:0017065	single-strand selective uracil DNA N-glycosylase activity	2.14E-03	3.15494616303483	4	1
GO:0009346	ATP-independent citrate lyase complex	2.10E-03	-1.6037562809426	52	1
GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	2.05E-03	-1.25265006865793	162	4
GO:0008815	citrate (pro-3S)-lyase activity	2.03E-03	-1.63749881151623	53	1
GO:0042121	alginic acid biosynthetic process	1.94E-03	1.14990181028563	127	12
4.1.3.6	citrate (pro-3S)-lyase.	1.94E-03	-1.60221767316624	52	1
GO:0047395	glycerophosphoinositol glycerophosphodiesterase activity	1.93E-03	-5.70423027266411	2	1
GO:0008092	cytoskeletal protein binding	1.90E-03	3.03923451098608	3	1
GO:0045151	acetoin biosynthetic process	1.90E-03	-1.56525594318597	48	1
4.1.1.5	acetolactate decarboxylase.	1.85E-03	-1.56525594318597	48	1
GO:0033711	4-phosphoerythronate dehydrogenase activity	1.79E-03	1.21622800529026	99	6
GO:0043130	ubiquitin binding	1.79E-03	2.84452978123873	6	1
2.8.3.10	citrate CoA-transferase.	1.78E-03	-1.57616213702899	52	1
GO:0008910	kanamycin kinase activity	1.78E-03	-1.58851951224173	11	1
GO:0046537	2,3-bisphosphoglycerate-independent phosphoglycerate mutase activity	1.78E-03	1.07286170037927	185	17
GO:0047356	CDP-ribitol ribitolphosphotransferase activity	1.72E-03	-6.7139421245469	1	1
GO:0000310	xanthine phosphoribosyltransferase activity	1.69E-03	-1.08340423243633	201	10
GO:0008814	citrate CoA-transferase activity	1.68E-03	-1.57775123389852	52	1
GO:0005727	extrachromosomal circular DNA	1.68E-03	-1.83727037420844	13	0
GO:0004792	thiosulfate sulfurtransferase activity	1.67E-03	-1.17227801781067	82	2
GO:0008707	4-phytase activity	1.67E-03	3.14255076857911	1	1
GO:0019677	NAD catabolic process	1.64E-03	1.30706123906711	32	1
GO:0008610	lipid biosynthetic process	1.64E-03	-1.47728727133267	87	2
2.4.2.22	xanthine phosphoribosyltransferase.	1.64E-03	-1.08534297116337	199	10
GO:0047330	polyphosphate-glucose phosphotransferase activity	1.59E-03	-3.53503053492603	5	1
2.7.1.23	NAD(+) kinase.	1.56E-03	-1.04348206094609	325	16
GO:0016746	acyltransferase activity	1.54E-03	1.10289112410377	347	18
GO:0071702	obsolete organic substance transport	1.54E-03	-1.20335282764238	103	4
GO:0006741	NADP biosynthetic process	1.53E-03	-1.04692108483062	329	16
4.2.1.-	Hydro-Lyases	1.52E-03	-1.90522797851666	19	1
GO:0006144	purine nucleobase metabolic process	1.45E-03	-2.26707747018229	21	0
GO:0004135	amylo-alpha-1,6-glucosidase activity	1.45E-03	1.16316684039	73	7
GO:0032265	XMP salvage	1.40E-03	-1.08523959035346	199	10
GO:0008760	UDP-N-acetylglucosamine 1-carboxyvinyltransferase activity	1.40E-03	-1.06687124783542	361	17
2.1.1.195	cobalt-precorrin-5B (C(1))-methyltransferase.	1.33E-03	-1.11006222435079	89	5
3.5.3.6	arginine deiminase.	1.32E-03	-1.3947618868725	58	2
GO:0003953	NAD+ nucleosidase activity	1.31E-03	1.31930434489008	29	1
1.1.1.22	UDP-glucose 6-dehydrogenase.	1.30E-03	1.14909369105086	142	10
GO:0097056	obsolete selenocysteinyl-tRNA(Sec) biosynthetic process	1.29E-03	-1.23940120402784	214	5
GO:0016297	fatty acyl-[ACP] hydrolase activity	1.28E-03	1.09879874319015	122	11
GO:0006522	alanine metabolic process	1.24E-03	-2.03436740514923	17	1
GO:0008808	cardiolipin synthase activity	1.18E-03	1.0806848147406	239	15
GO:0009409	response to cold	1.13E-03	-1.9117942663824	35	0
GO:0008899	homoserine O-succinyltransferase activity	9.43E-04	-1.07852816252639	194	11
GO:0008276	protein methyltransferase activity	8.62E-04	-1.0433934172264	283	16
1.1.1.88	hydroxymethylglutaryl-CoA reductase.	6.52E-04	-1.48308851426957	50	0

Table 4.1 – **Robust subset of annotations from the IBD dataset.** Robust FAs of the IBD dataset, identified by their GO term or EC number, as well as their current name. Annotations are classified by decreasing average Gini importance score, over all 200 RFs trained at the optimal selection level (20 per run, 10 runs). Extra information include: the ratio between the average scores of the annotation in control and unhealthy profiles, ponderated by -1 if the annotation is most present in the unhealthy profiles, the amount of taxa attached to each FA, and the number of robust taxa within them.

ID	Names	Average RF importance	Ponderated average ratio (Control/Unhealthy)	Number of linked annotations	
				Total	Robust
Organism_73	Alistipes finegoldii	2.84E-02	16.3964097691144	1220	15
Organism_224	Akkermansia muciniphila	2.12E-02	3.23501956449667	1452	18
Organism_12	Bifidobacterium bifidum	2.03E-02	-11.3214966525224	1307	28
Organism_144	Lachnospiraceae bacterium 2 1 58FAA	1.91E-02	-18.4267002012075	355	5
Organism_169	Ruminococcus lactaris	1.90E-02	3.04069705100761	431	5
Organism_127	Beubacterium ventriosum	1.51E-02	2.6429359268965	1388	20
Organism_156	Oscillibacter unclassified	1.42E-02	-1.89778568117644	724	16
Organism_134	Butyrivibrio unclassified	1.39E-02	-1.65319948992604	916	10
Organism_54	Odoribacter splanchnicus	1.33E-02	1.85573337062113	1595	19
Organism_75	Alistipes onderdonkii	1.33E-02	2.48594496944176	1391	15
Organism_78	Alistipes shahii	1.30E-02	1.65146163415684	935	8
Organism_171	Subdoligranulum unclassified	1.27E-02	1.5560202207333	627	5
Organism_152	Roseburia hominis	1.18E-02	1.7903389716571	1500	20
Organism_138	Coprococcus sp ART55 1	1.16E-02	2.2748823646463	701	8
Organism_163	Ruminococcaceae bacterium D16	1.12E-02	-4.30319855302151	1390	30
Organism_162	Faecalibacterium prausnitzii	9.80E-03	-1.57257090414346	1220	18
Organism_53	Coprobacter fastidiosus	9.67E-03	6.06805781620637	1503	19
Organism_40	Bacteroides massiliensis	9.49E-03	1.53976594131914	1602	20
Organism_136	Coprococcus comes	9.19E-03	-1.68577511310286	116	1
Organism_74	Alistipes indistinctus	8.46E-03	1.2651644466561	1447	19
Organism_20	Collinsella aerofaciens	8.42E-03	-1.82725111812987	1349	20
Organism_123	Eubacterium hallii	7.81E-03	1.07627573371101	144	3

Table 4.2 – **Robust subset of taxa from the IBD dataset.** Robust taxa of the IBD dataset, identified by their internal identifier, as well as their current name. Taxa are classified by decreasing average Gini importance score, over all 200 Random Forests trained at the optimal selection level (20 per run, 10 runs). Extra information include: the ratio between the average abundances of the taxon in control and unhealthy profiles, ponderated by -1 if the taxon is most present in the unhealthy profiles, the amount of FAs attached to each taxon, and the amount of robust annotations within them.

4.2.1 Bibliographic exploration of an output shortlist.

The bibliographic examination was conducted on all of the robust annotations from the IBD dataset, as well as samples of 20 annotations that were present in 50% of the significant sublists obtained from the pipeline's runs, and 20 non-candidate annotations. The methodology was to research the name of the annotation alongside the name of the disease on Google Scholar(`https://scholar.google.com/`). If none of the research results provided conclusive information linking this annotation to IBD, be it in a host model or in the microbiota, the chemical products and eventual alternative names of the annotation were similarly tested, followed by related (parent or child) annotations, and finally the linked pathways listed in the BRENDA database [132]. From this exploration, the annotations were given a bibliographic relevance grade of 1 (most relevant to the disease) to 4 (least relevant to the disease) based on the following criteria:

Category 1: A direct link was established between the annotation, or a direct product metabolite, and IBD. This can come in the form of an explicitation of the metabolic

mechanisms involved, or simply in the form of measured differential presence between unhealthy and control individuals. To note: conclusions derived from other ML-based approaches were not considered to be sufficient evidence, as they could suffer from biases similar to our own approach.

Category 2: A direct link was established between a similar metabolic function and the disease. Were considered as similar: proteins or enzymes from the same family as the one involved in the annotation (i.e: ATP-dependent and ATP-independent citrate lyases), and parent and child annotations, signaling notably that the annotation is indeed relevant, but at the wrong scale.

Category 3: An indirect correlation was established between the annotation and the disease. This can mean that the annotation was not directly linked to IBD, but that it is involved in a larger pathway or expressed by a taxon that has significance.

Category 4: No leads were found, or the annotation was proven to be irrelevant.

The resulting bibliographic scores affiliated to the robust functional selection of the IBD dataset are detailed in Table 4.3. For exhaustive details on sourcing and grade justification, refer to Appendix B.

4.2.2 Our approach to variable selection is coherent with known expressions of the gut microbiota in context of the disease.

Among the robust annotations, several were found through bibliography to be relevant to the disease when expressed in the host organism as opposed to the microbiota. We considered both cases as a link found between the annotation and the disease, following the idea of permeability and interactions between the microbiota and its host [133].

When available, we also retrieved the group, namely unhealthy or control, most likely to express these annotations according to the bibliography. At the same time, we can measure which group most expresses each of these robust FAs on average. We confirmed these associations between FA and group with limma [104] as well, for better robustness. We found that bibliography predictions and prevalence in the IBD dataset patients were in agreement in 47% of cases. Functional annotations where disagreement exists between the bibliography and measured average differential expression might point towards a rescue of important functions in the host by the microbiota [134].

A complementary comparative analysis was conducted by the means of a Chi^2 contingency test [135] with a 5% p-value threshold between the prevalences of each bibliographic

ID	Name	Bibliographic category
GO:0006520	amino acid metabolic process	1
4.1.2.-	Aldehyde Lyases	2
GO:0102545	phosphatidyl phospholipase B activity	1
GO:0004122	cystathionine beta-synthase activity	1
GO:0008744	L-xylulokinase activity	3
GO:0047419	N-acetylgalactosamine-6-phosphate deacetylase activity	1
GO:0008788	alpha,alpha-phosphotrehalase activity	3
GO:0032440	2-alkenal reductase [NAD(P)+] activity	3
GO:0001510	RNA methylation	1
GO:0015444	P-type magnesium transporter activity	2
GO:0016832	aldehyde-lyase activity	2
GO:0047605	acetolactate decarboxylase activity	3
GO:1901135	carbohydrate derivative metabolic process	1
GO:0017065	single-strand selective uracil DNA N-glycosylase activity	1
GO:0009346	ATP-independent citrate lyase complex	2
GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	1
GO:0008815	citrate (pro-3S)-lyase activity	1
GO:0042121	alginic acid biosynthetic process	1
4.1.3.6	citrate (pro-3S)-lyase.	1
GO:0047395	glycerophosphoinositol glycerophosphodiesterase activity	1
GO:0008092	cytoskeletal protein binding	1
GO:0045151	acetoin biosynthetic process	3
4.1.1.5	acetolactate decarboxylase.	3
GO:0033711	4-phosphoerythronate dehydrogenase activity	3
GO:0043130	ubiquitin binding	1
2.8.3.10	citrate CoA-transferase.	1
GO:0008910	kanamycin kinase activity	1
GO:0046537	2,3-bisphosphoglycerate-independent phosphoglycerate mutase activity	3
GO:0047356	CDP-ribitol ribitolphosphotransferase activity	2
GO:0000310	xanthine phosphoribosyltransferase activity	3
GO:0008814	citrate CoA-transferase activity	1
GO:0005727	extrachromosomal circular DNA	1
GO:0004792	thiosulfate sulfurtransferase activity	1
GO:0008707	4-phytase activity	3
GO:0019677	NAD catabolic process	1
GO:0008610	lipid biosynthetic process	1
2.4.2.22	xanthine phosphoribosyltransferase.	3
GO:0047330	polyphosphate-glucose phosphotransferase activity	1
2.7.1.23	NAD(+) kinase.	1
GO:0016746	acyltransferase activity	2
GO:0071702	obsolete organic substance transport	3
GO:0006741	NADP biosynthetic process	1
4.2.1.-	Hydro-Lyases	2
GO:0006144	purine nucleobase metabolic process	1
GO:0004135	amylo-alpha-1,6-glucosidase activity	3
GO:0032265	XMP salvage	2
GO:0008760	UDP-N-acetylglucosamine 1-carboxyvinyltransferase activity	3
2.1.1.195	cobalt-precorrin-5B (C(1))-methyltransferase.	4
3.5.3.6	arginine deiminase.	1
GO:0003953	NAD+ nucleosidase activity	1
1.1.1.22	UDP-glucose 6-dehydrogenase.	1
GO:0097056	obsolete selenocysteinyl-tRNA(Sec) biosynthetic process	1
GO:0016297	fatty acyl-[ACP] hydrolase activity	3
GO:0006522	alanine metabolic process	1
GO:0008808	cardiolipin synthase activity	3
GO:0009409	response to cold	1
GO:0008899	homoserine O-succinyltransferase activity	3
GO:0008276	protein methyltransferase activity	2
1.1.1.88	hydroxymethylglutaryl-CoA reductase.	1

Table 4.3 – **Bibliographic scores of the robust subset of annotations from the IBD dataset.** Robust FAs of the IBD dataset, identified by their GO term or EC number, as well as their current name, and the bibliographic category assigned to each of them.

categories in the robust selection and those of randomly selected non-candidate annotations. Results are described in Table 4.4. This test established that the robust group significantly diverged from the non-candidate group. This significant difference is notably driven, as seen in Table 4.4, by a comparative increased proportion of Category 1, and decreased proportion of Category 4 annotations in the robust subset compared to the non-candidate selection.

These results support the notion that our pipeline allows for a relevant selector of information. Beyond this first observation, the question of this selection's potential for innovation arises, as we should explore whether it includes relevant factors beyond what is already well known.

	Category 1	**Category 2**	**Category 3**	**Category 4**	**Total sample size**	**Chi² contigency test p-value (vs robust)**
Robust	32	9	17	1	59	-
50% Candidates	5	4	10	1	20	0.13645
Non candidate	5	3	9	3	20	**0.027509**

Table 4.4 – **Counts of the different bibliographic categories per researched selection, and p-values of a Chi² contigency test compared to the robust subset.**

In summary

An in-depth exploration of the biological significance of all of the robust functional annotations derived from the IBD dataset was conducted, materialized by a grading of each annotation's relevance in regard to the disease according to existing research. A downstream analysis of these results revealed that the most relevant category of annotations was significantly more represented in our robust selection when compared to an excerpt of the list of non-selected annotations.

4.3 The interconnections between taxa and annotations expose cumulative metabolic signatures.

4.3.1 Exposing different types of dynamics between significant taxa and annotations from their interconnections.

The observed disparities and the non-redundancy between taxonomic and functional profilings (see Section 2.3) prompt the question of whether these profiles equally provide

valid descriptions of a subject's microbiota. A potential drawback of the taxonomic scale is the cumulation effect, wherein individual taxa may have little significance but contribute significantly to an essential metabolic process when grouped together. As a result, this collective impact might go unnoticed when focusing solely on individual taxa. The dynamics in terms of specificity between annotations and taxa are illustrated in Figure 4.1, which plots the amount of robust taxa associated to each annotation as a function of the total amount of associated taxa. For illustration purposes, the represented annotations were assigned into four profiles based on their number of associated taxa. We labeled the top 10% as "Ubiquitous" (5 annotations, top right in Figure 4.1), the bottom 10% as 'Specific' (18 annotations, bottom left of Figure 4.1), and all others were labeled 'In-Between' (32 annotations). Finally, a fourth category was drawn up, independently of the previous criteria, containing 4 annotations that have no link to robust taxa, which we labeled as 'Cumulative'. This representation shows that important annotations have differing relationships to their taxonomic counterparts, and that an annotation's importance can stem from the influence of several taxa, as is notably illustrated by the 'Cumulative' class.

4.3.2 Detailing the relationships between taxa and their functional annotations highlights the cumulative expression of functional signatures.

A detailed illustration of pairings between select robust annotations and taxa is proposed in Figure 4.2. The strength of the links is also represented, defined as the amount of proteins within a taxon's proteome that express a given annotation for Figure 4.2a, and following the following formula for Figure 4.2b:

$$\frac{\bar{n}_{t,i} \times x_{F,t}}{\sum_{t \in T} \bar{n}_{t,i} \times x_{F,t}}$$

where $x_{F,t}$ is the number of proteins within taxon t's proteome that are linked to the function F, $\bar{n}_{t,i}$ is the average of the abundances of a taxonomic affiliation within the dataset and T is the ensemble of all taxa associated with the annotation.

The represented annotations in Figure 4.2a were picked from each of the categories illustrated in Figure 4.1: GO:0006520 as representative of the 'Ubiquitous' class, 1.1.1.22 for the 'In-between' class, GO:0043130 for the 'Specific' class, and GO:0006144 as a 'Cumulative' example. From top to bottom, the first annotation (GO:0043130) is a case in

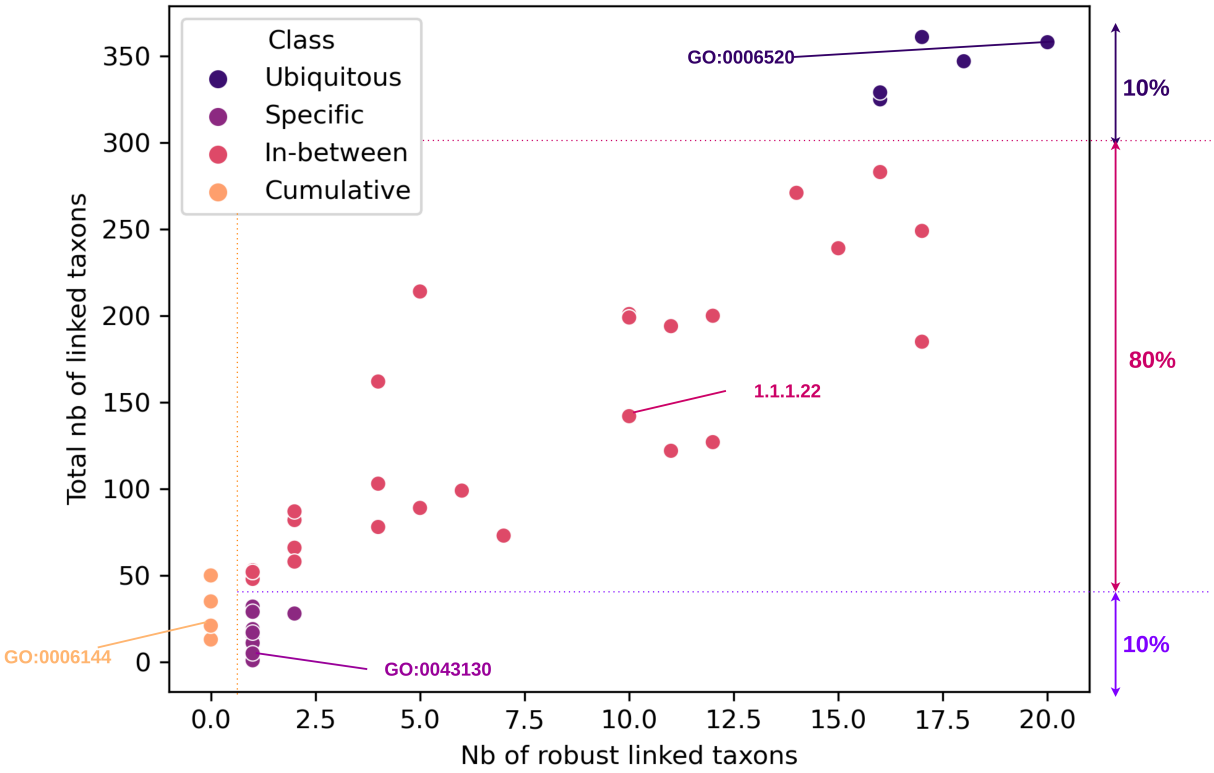
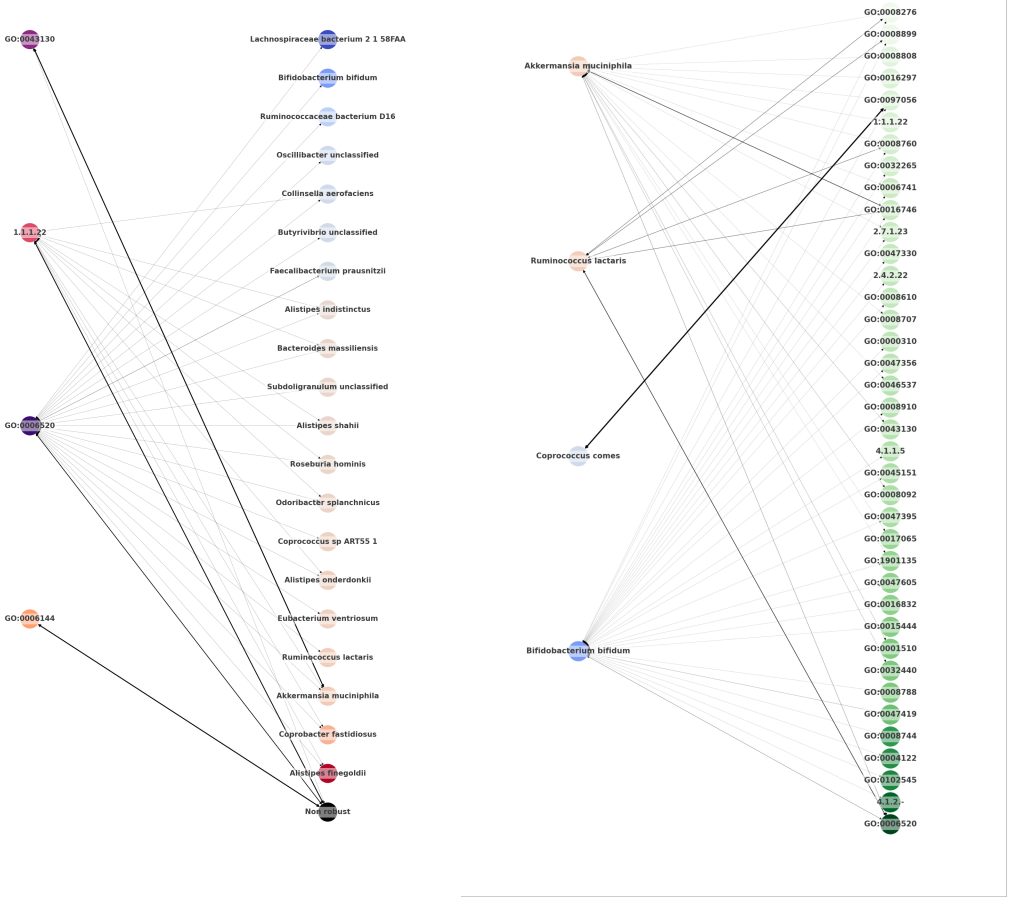


Figure 4.1 – **Number of taxa associated to each robust annotation, as a function of the number of associated robust taxa for the IBD dataset.** Four groups of annotations are represented, three of which were determined based on the total amount of taxa attached to the annotation: those within the top 10% of these values' scale were labeled 'Ubiquitous', those in the bottom 10% were labeled 'Specific', and the others were labeled 'In-between'. The final category corresponds to the robust significant annotations with no relationship to the robust significant taxa ('Cumulative'). The highlighted annotations are those used as illustrative examples in Figure 4.2.



(a) **Associations between robust functions and robust taxonomic counterparts, for the best iteration on the IBD dataset.**

(b) **Associations between robust taxa and robust functional counterparts, for the best iteration on the IBD dataset.**

Figure 4.2 – **Interassociations between robust annotations and taxa, from the IBD dataset.** **(a)** Depicted annotations were selected to be representative examples of the different categories highlighted in Figure 4.1, and are presented with the same color scheme. **(b)** Represented taxa were chosen to showcase control and healthy representatives with high and low numbers of connections to robust annotations. Relationships to non-robust annotations were not represented here for reasons pertaining to readability of the figure.
Taxa are colored on the basis of their normalized average differential expression between Control (red) and Unhealthy (blue) profiles. The width of the connections is proportional to the importance of the association. The arrow between a given function and the generic 'Non robust' node represents the contribution of non-robust taxa to the considered function.

which the feature's significance appears to be due to a strong association to a single robust significant taxon, namely *Akkermansia muciniphilia*. This taxon has an established impact on IBD remission, and is researched as a potential probiotic treatment of the disease [136]. This is also in accordance with the annotation's differential expression between profiles, as seen in Table 4.2, where the annotation is shown to be expressed in the control samples almost 3 times as frequently on average as it is in the sick samples. This kind of relationship could either indicate that this 'Specific' annotation derives its importance in our predictions from its strong and specific attachment to an important taxon, or that its impact on the disease is an important factor to explain this taxon's benefactory influence. GO:0043130 corresponds to ubiquitin binding, a mechanism which is known to regulate the inflammation process of intestines via different signalling pathways [137], and is categorized as a Category 1 annotation by our bibliographic research, showing that in the case of our example, the effects of the annotation and of its specifically associated robust taxon align. It should be noted that, as mentioned in our earlier discussion around our bibliographic work, the differential expression of a feature can be contradictory with its known effects, and should therefore be treated with caution. The second and third annotations (1.1.1.22 and GO:0006520), respectively from the 'In-between and 'Ubiquitous' groups, are very widespread among robust taxa, without any particularly strong link to any of them. In cases such as these, meaning metabolic functionalities commonly expressed within taxa, the issue of significance is shown to not be a purely binary question of expression or absence, as both annotations are consistently present in both profiles. Finally, the last annotation (GO:0006144) is exclusively linked to non robust taxa. All such annotations, from the 'Cumulative' group, are associated to several taxa (13 minimum), meaning that their importance results from the cumulated influence of multiple, individually non-significant taxa, that have a significant role when grouped functionally.

The reverse associations, plotted in Figure 4.2b, show that this form of cumulation is specific to FAs: the robust taxon with the least associations to robust annotations, *Coprococcus comes*, is represented and shown to still have a non-zero amount of correlations to robust annotations. As such, these results illustrate the notion that, while taxa will usually have a remarkable impact on host health through at least one important metabolic functionality, impactful functions can go unnoticed if we try to derive them from their taxonomic counterparts.

This further supports the importance of exploiting microbiota information at the functional level rather than at the taxonomic level. Annotation GO:0006144, which corre-

sponds to the purine metabolic process and is represented in orange in Figure 4.1, is a good illustration of this approach's advantages. This annotation was not correlated to any robust taxon, and therefore would be difficult to derive from an taxonomic approach. Indeed, the bibliography shows that this annotation was linked to IBD through oriented research following a first mechanistic study [138], whereas our approach was capable of identifying it efficiently and without any pre-orientation.

In summary

Looking at the details of the associations between robust taxa and annotations uncovers that the importance of functional signatures can be derived from:

- Being expressed by many taxa, of which many are robustly significant (**Ubiquitous**),
- Being expressed specifically by a few robustly significant taxa (**Specific**),
- Being expressed only by non-robustly significant taxa (**Cumulative**).

The existence of cumulative functional signatures in the gut microbiota confirms the benefits of using functional profiles for microbiota analysis instead of taxonomic abundances.

4.4 Conclusion and discussion.

Through bibliographic research, we have validated the relevance of our method's automatic selections on an example. However, it was through the exploitation of the interassociations between taxa and functional annotations, traced from building our functional profiles, that we arrived to this thesis' third contribution: confirmation of the existence of a functional cumulation effect within taxonomic profiles. While this effect was shown on only one example, meant primarily as proof of concept, it confirms the intuition that all relevant functional information in the gut microbiota cannot be derived from taxonomic signatures alone, and that there is knowledge to be gained from studying the gut microbiota directly at the functional scale.

Performance and selection sizes are indicators of a selection's quality.

The conclusions presented here were reached through the exploration of the contents of robust selections from the functional and taxonomic profiles of the IBD dataset. This was possible because in the case of this dataset, both selections were of a size that was

compatible with a deepened research, without being too small either. In the case of the Cirrhosis, Colorectal or WT2D datasets for example, the size of the robust functional selections could be deemed insufficient in size to contain relevant information. The exploration of the confident subset or the exploitation of a lower level of selection than the proposed optimum could be envisioned by a user if the content included in the recommended robust output is deemed insufficient. It should also be noted that the IBD dataset's good performances during classification (see Figure 3.4) are another element that support the relevance of the contents of its selection, which all datasets do not share. Users of this method should be mindful that the output list may not be as relevant if the classification performances are low.

Implementing the method for wider application.

Applied in succession, the manipulations described in Chapter 2, Chapter 3, and this chapter make up a method for functional analysis of the gut microbiota. In order to make this method available, and to make the reproduction of our results easier, we have implemented them as a pipeline software, meant for public use.

Chapter 5

Implementing the SPARTA pipeline.

The results presented in Chapter 4 were obtained through the application of the methods described in the previous sections: a computation of functional profiles from a taxonomic description of the gut microbiota, with the establishment of a traceable connection between taxa and their FAs, as described in Chapter 2, followed by a robust selection of variables derived from repeatedly iterating an interpretable classification and an automatic variable selection method, as described in Chapter 3. Our results highlight the need to perform at least one iteration and several repeated runs in order to reduce the dimensionality of notably the functional datasets, while maintaining the classification performance, and derive a list of robust variables. In order to facilitate the chaining of these manipulations and to make them reproducible, they were implemented in the form of a pipeline software: Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes (SPARTA).

Said implementation is available on GitHub, at the following URL: `https://github.com/baptisteruiz/SPARTA`. An initial version was made using bash and Python, but for performance and stability purposes, a later version coded entirely in Python was developed.

5.1 SPARTA overview: a Machine Learning-driven method for paired analysis of taxonomic assignations and functional annotations.

SPARTA (see Fig 5.1) requires two compulsory inputs. The first is a table describing the microbial relative abundances (i.e: taxonomic abundance tables) for each microbiota sample within the dataset, from which functional profiles will be computed. This profile can be computed on the basis of 16S or MGS sequences, as both can be handled by the pipeline. The other is a vector file indicating the groups according to which each sample

within the dataset should be classified, represented as green and red colors in Figure 5.1. Optionally, the pipeline can be run in two separate steps: `sparta esmecata` and `sparta classification`. The `sparta esmecata` step covers the formatting of the inputs and builds a functional profile based on EsMeCaTa. The `sparta classification` step runs the iterative classification and selection process, as well as the post-processing of the results. For further details, see Section 5.2. Complementary inputs and commands can also be used in certain circumstances, as described in Section 5.3.

SPARTA computes three major outputs. The first is a functional profile: by using the EsMeCaTa tool [67] to query the UniProt [68] database, we associate a representative proteome to each taxon from the original profiles, and link them to FAs (GO terms [58], EC numbers [117]) through UniProt once more [68] or using eggnog-mapper [70, 71]. The prevalence of each of the obtained annotations within the individual samples are then calculated as scores of FAs, as described in Chapter 2.

The second consists of classification performances: SPARTA trains RF [80] classifiers on the obtained functional profiles, and measures their performance in categorizing the samples. Classification performances for the best performing iterations are highlighted, however detailed results are also given for all iterations. This section of the code is based on a modified version of DeepMicro's [94] implementation.

Finally, SPARTA generates a list of features, both taxa and FAs, which are identified as significantly discriminating between the given sample groups on the basis of an automatically calculated selection threshold applied to their average importance scores (see Chapter 3). SPARTA provides the user with the list of important taxa and FAs for each iteration, with a focus on the best iteration after the first level of selection. The associations between taxa and annotations are also explicitated, allowing each feature to be linked notably to its significant counterparts.

Details of the output files are given in Section 5.3.

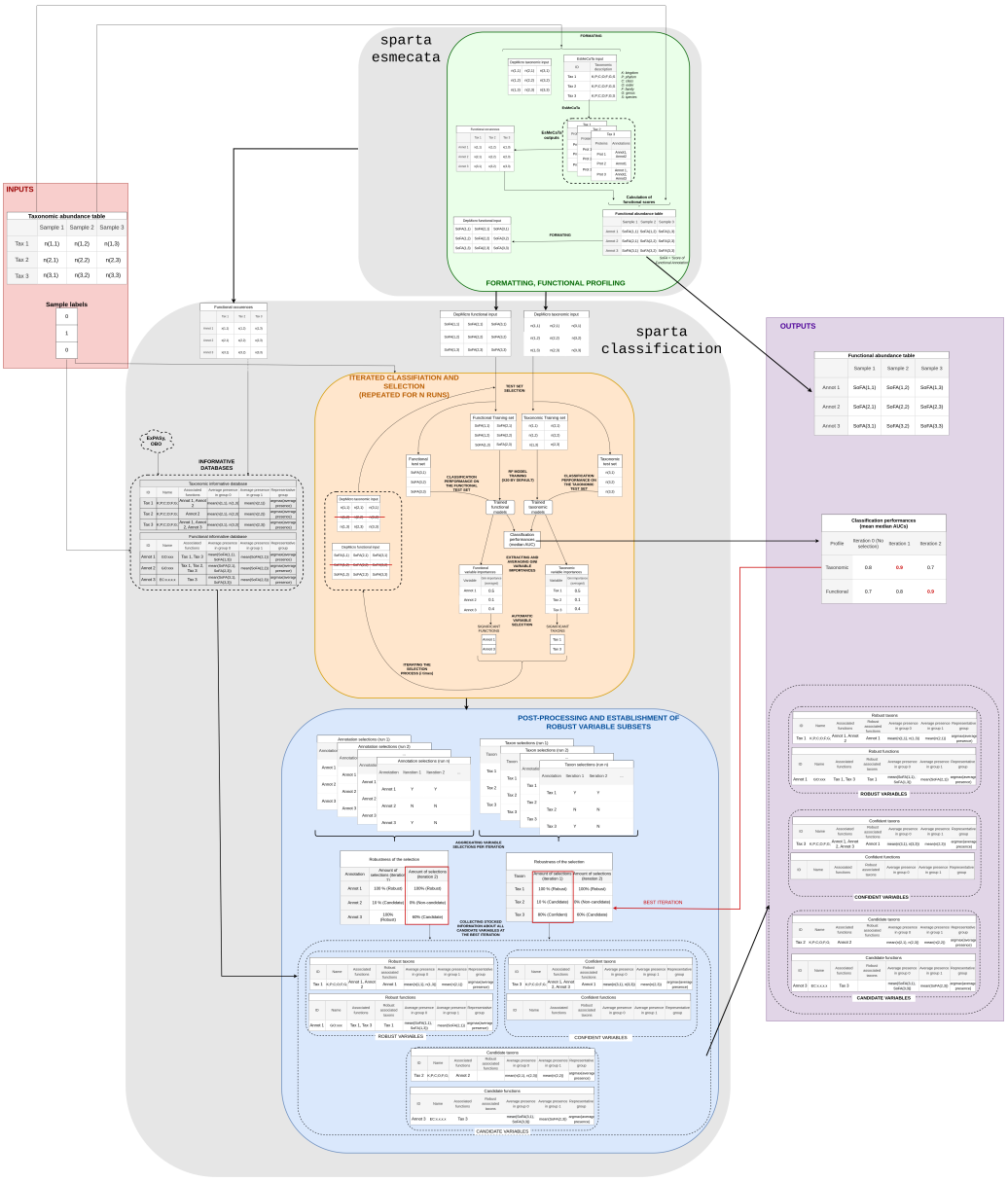


Figure 5.1 – **Visual representation of SPARTA's implementation.** The inputs (red block) include a taxonomic abundance table (tab delimited, txt format) and a vector of sample labels (csv file). SPARTA's implementation has three main blocks. The first block (green) formats the input abundance table and transforms it into a functional profile using the EsMeCaTa pipeline (see Sections 5.2.1, 5.2.2, 5.2.3). This functional abundance profile is given as output. The second block (orange) uses both functional and taxonomic profiles for repeated classification and variable selection, setting aside a test set for each run and training 20 Random Forest models per profile. Performance on test sets and automated variable selection results are given as output (see Section 5.2.4). The final block (blue) compiles selections from all runs to measure the robustness of each variable's importance at each iteration level. The list of robust variables is given as output for the best iteration level and all other levels in a separate folder (see Section 5.2.5). Zooms on each section are available in Appendix C. The first block can be run on its own through the `sparta esmecata` command, and the two following blocks can also be run directly using the `sparta classification` command. Otherwise, the entire process can be launched with the `sparta pipeline` command.

In summary

The SPARTA pipeline is implemented so as to be compute functional profiles, RF classification performances from taxonomic and functional profiles, and aggregated taxonomic and functional variable selections (robust, confident and candidates, see Section 3.1.3) from a taxonomic table and associated sample labels. The pipeline can be launched in full (`sparta pipeline`), but the steps for functional profiling (`sparta esmecata`) and ML-driven classification and variable selection (`sparta classification`) can also be decoupled from each other.

5.2 SPARTA's main functions.

The pipeline is executed in several steps, in order. The first, which can be called using the `sparta esmecata` command, is a formatting of the inputs for the creation and formalization of EsMeCaTa's input from the given data, followed by a run of the EsMeCaTa pipeline [67]. The functional profiles are then calculated, and a first review of the variables' information is operated. The two following steps are effectuated by the `sparta classification` command. This firstly involves the operation of the iterative classification and variable selection process, and repeated as many times as requested by the user. Once this step is over, the classification performances are plotted, and the robust, confident and candidate sub-lists are established. Figure 5.1 illustrates the implementation. The following sections will detail how each of these steps were executed.

5.2.1 Formatting the inputs and running EsMeCaTa (sparta esmecata command).

Before any operation is conducted on them, the pipeline's inputs are shaped into a form that is compatible with the next steps of the pipeline, notably those based on external tools (EsMeCaTa [67] and DeepMicro [94]), and removes eventual metadata from the input. Importation and handling of the data is made through the pandas [139] library. This step notably takes as input the full original abundance table, and outputs a metadata-less version of it, in the original format and transposed so as to be compatible with DeepMicro, and formatted as demanded by the user if applicable (conversion to

relative abundances, for example). A taxonomic description of all taxons, formatted so as to be compatible with the EsMeCaTa pipeline, is also created at this step.

This latter output is used to launch EsMeCaTa, which recovers the functional annotation of the taxonomic units given as input as described in Section 2.1.1.This step can be skipped entirely if results for a dataset with the same name are found, unless the "–esmecata_relaunch" flag is raised, to avoid redundant calculation.

Each of EsMeCaTa's steps is called individually, and checks are operated the outputs after the proteome and annotations step, during which issues with an HTTPS connexion error are most likely to occur.

The 'proteome' step, which downloads annotations associated to each taxonomic unit from the UniProt database [68], uses the Bioservices request option, as it has proven more stable. Other parameters are set to default. Following this operation, we check whether for each taxon given as input, we have a downloaded fasta file. If the verification fails, this step can be relaunched up to 20 times. If the amount of retries exceeds this number, the operation is aborted and an exception is raised.

The 'clustering' operation groups the obtained proteins into clusters based on identity using the Mmseqs2 package [69], and selects those that are representative of 80% or more of the species contained in the taxonomic unit. This step is called with EsMeCaTa's default parameters. Re-running this step over previous results can cause crashes, therefore if an existing incomplete output directory is found, it will be deleted and rewritten from scratch.

Finally, the 'annotation' step recovers the FAs associated with each cluster of proteins. The option is given here to use the eggnog-mapper tool [70, 71] to perform the annotation, otherwise it will default to interrogating UniProt again. The used parameters are EsMeCaTa's defaults. After this step, a check is operated to see if all taxa for which a non-empty protein consensus was found have a final output. Similarly to the 'proteomes' step, this process can be iterated up to 20 times.

The resulting annotations are resumed in a table of functional occurrences, which contains the amount of proteins that express each of the retrieved annotations in each taxon's proteome.

5.2.2 Calculation of functional scores (sparta esmecata command).

The second step is the calculation of the scores of the FAs obtained through the previous step, following the method described in Section 2.1.2 and using the table of functional occurrences, as well as the original microbial abundances. This ensures that the reference-based method EsMeCaTa provides a quantitative annotation-based description of the gut microbiota. Once the functional profile is obtained, its values can also be scaled using the TF-IGM normalization method, as described in Section 2.1.3. As with the profile given as input, the calculated profile is passed on in two formats: a table in the same format as the input, to be exported and written as an output, and a transposed version compatible as a DeepMicro input.

In summary

The `sparta esmecata` command runs the EsMeCaTa pipeline on all of the taxa contained in the input taxonomic table, to associate them with FAs from either the UniProt database, or using eggnog-mapper. Scores are then calculated for each gathered FA on the basis of these associations as well as the initial taxonomic abundances, following the method described in Section 2.1.2. The calculated functional profile serves as this step's main output.

5.2.3 Creating an informative database for the variables (sparta classification command).

Having all of the processed microbiota profiles at our disposal, descriptive tables are created for each variable in our profiles, both taxa and FAs. These descriptions notably measure the average presence of a feature in each labeled category as well as over the whole profile, and explicitates the category that expresses it the most on average. For taxa, the detailed taxonomy and the annotations it expresses are fetched from EsMeCaTa's inputs and outputs, in the form of a table of functional occurences for the latter. For each FA, the name of the annotation is fetched from a database (OBO PURLs for GO terms, hosted at `http://purl.obolibrary.org/obo/go/go-basic.obo`, and the ExPASy database for EC numbers, hosted at `https://ftp.expasy.org/databases/enzyme/enzyme.dat`), and a list of all taxa that express it is gathered. The datasets in

question are queried locally, and are downloaded by default the first time the pipeline is used. Interrogation of the databases is conducted using Biopython's [140] ExPASy.Enzyme module [1] for ExPASy, and goatools' [141] obo_parser function for GO terms [2].

5.2.4 Iterated classification and selection (sparta classification command).

The following steps are repeated for as many runs as demanded by the user through the "-r" argument of the command. The procedures for the selection of the test, validation and training set are seeded for reproducibility purposes. All of the seeds used in this process are generated using a master seed, which can be set by the user. The test sets for each run can also be directly specified by the user. Figure 3.2 illustrates this implementation. By default, SPARTA uses the same master seed as the one used to obtain the results presented in Chapter 3. For each run, the user may request that only a specified subset of the input profiles' variables be taken in account through the "–preselected-organisms" and "–preselected-annotations" arguments.

5.2.4.1 Setting aside a test set.

Individuals are selected to be set aside as a test set for the entirety of the current run. Selection is conducted using the scikit-learn library's [121] train_test_split function [3], with a test size parameter of 0.2, meaning that the test sample will be 20% the size of the full dataset. Selected individuals are removed from the functional and taxonomic profiles, and kept in a separate table, to be only used for measuring trained classifiers' performance.

This step is only effectuated once per run, whereas the following two are iterated in succession.

5.2.4.2 Training classifiers.

This step involves the training of several successive ML classifiers to sort individuals according to their associated labels, based on the relative abundance profiles of their

1. https://biopython.org/docs/1.75/api/Bio.ExPASy.html
2. https://github.com/tanghaibao/goatools
3. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

microbiota or on their calculated functional representation. The implementation of this step is taken and adapted from the DeepMicro tool [94], and is repeated as many times within an iteration as demanded by the '-c' argument. The user can choose, through the '-m' argument to train RF or SVM classifiers. In both cases, the models will be trained using their implementation by the scikit-learn library [121].

During training, the data not set aside as a test subset is randomly split into a training set and a validation set, with a respective 80% / 20% distribution. For the first iteration, this split is seeded to ensure reproducibility. For the following iterations of the same run, the obtained training and validation sets will be used again, as illustrated by Figure 3.2. In order to account for the disparity in representation between the unhealthy and control individuals within the datasets, both classes were given weights proportional to their frequency, as implemented by scikit-learn's 'balanced' class weight parameter [121]. The training involves a Grid Search, as implemented by scikit-learn [121], to optimize the estimator's parameters. For RFs, the optimized parameters are the number of estimators per forest, the number of leaves per estimator, and the amount of information to which each tree has access, and the split quality criterion is measured via the Gini Impurity metric. For SVMs, the optimized parameters are the regularization parameter, which tunes the impact of the loss function during training, and the classifier's kernel, which can be linear or Gaussian with Radial Basis, with a tuning of the gamma parameter (radius of each sample's area of influence) in the latter case.

Each time a model is trained, its performance metrics (ROC AUC score [86], accuracy, recall, precision and F1 score) on training, validation and test sets are recorded. The model with the highest AUC on the validation set during the iteration is exported using the joblib library [142]. If the user chose to train RF classifiers, the importance metrics of each variable for the trained model are also exported. These importances can be based on one of two metrics, depending on the user's input. The first option is based on the Gini Importance metric [80], calculating the mean accumulation of the impurity decrease within each tree, as implemented in the scikit-learn Python library [121]. The other option is the SHAP importance [85], which calculates each variable's contribution to a decision from the basis of a trained classifier. In our case, dealing with RFs, we relied on the SHAP package's [85] implementation of the TreeExplainer [143], which is an algorithm for calculation of SHAP values optimized for RF models. If multiple iterations of the classifier's training are made, the feature importances are averaged over all iterations. Features are then ranked based on this metric in decreasing order. In SHAP's case, this

ranking is made based on the absolute value of the importance scores. Predictions are made through the application of an optimal classification threshold, determined through the maximization of the Youden statistic [144], to probabilistic predictions by the model. The optimal thresholds found for each model are exported, alongside their performance metrics.

5.2.4.3 Variable selection.

This section is only applicable if the user chose to train RF classifiers. From the variables' importance scores averaged over all repetitions of the RF training process, an automatic variable selection is operated as described in Section 3.1.2. This consists in ranking the variables by decreasing importance score, then calculating the inflection point of the curve of decreasing importance scores, using kneebow's [122] Rotor and get_elbow_index methods [4], before cutting off all variables below the obtained index. A selection of the retained variables can then be operated on the taxonomic and functional profiles, to be used as input for the previous step as many times as required by the command's "-i" argument.

In summary

The first step of the `sparta classification` command involves the iteration of model training and variable selection, and its repetition with different test sets. The amount of re-trained classifiers, of variable selection iterations and of repetitions of the full process are the same as those used in our previous manipulations by default, but can also be changed by the user. The selection of test and validation sets, as well as the training of the classifiers, are fully seeded to ensure full reproducibility. Users can also specify their own test sets. This step's main outputs are the model classification performances.

5.2.5 Post-processing and establishment of robust variable subsets (sparta classification command).

After the previous iterative process has been repeated as many times as dictated by the "-r" argument, the classification performances obtained through the previous iterative

4. `https://github.com/georg-un/kneebow/tree/master`

steps are processed: for each run, the median AUC of each iteration is calculated. The median of these values is then calculated per iteration level over all runs, and the iteration level that maximizes this value is retained as optimal. If there was more than one iteration of variable selection, the first iteration, on which no variable selection was operated, is excluded. For each processed profile (taxonomic and functional), the best iteration's performances are plotted using the matplotlib [145] and seaborn [146] libraries. A statistical comparison of the median performances per run at optimal iteration for each profile is also made at this point, using scipy's [147] implementation of the Mann-Whitney U-test [5]. The test's p-value is indicated on the plot. The optimal selection levels found in this step are passed onward.

Finally, we compare the sublists obtained at each level of iteration for each run of the pipeline, and count the common occurrences. Unanimously selected variables are labeled as 'robust' (or 'core'), and those that appear at least once are labeled 'meta', which itself can be separated in: 'confident' if they are selected in 75% of the runs or more, or 'candidate' otherwise. The obtained lists are enriched with information, gathered previously in step 5.2.3 notably. A sublist of significant linked counterparts is established for each variable, referencing the linked counterparts (taxa that express an annotation, or annotations expressed by a taxon) that are listed as 'robust' at the same level of iteration.

The 'core' and 'meta' lists for all iterations, as well as all of the individual selections they are based on, are given as output, however the sublists obtained on the optimal iterations for taxonomic and functional profiles are saved in a separate folder, and reference each other when establishing significant linked counterparts. This means for example that the significant annotations linked to a taxon in the optimal sublist are deemed 'robust' based on the optimal iteration for the functional profile, rather than at the iteration level that is identical to the functional optimum, and vice-versa. If the best obtained median RF AUC is inferior to 0.6, a message warning that the selection may be unreliable will be passed to the user.

5. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`

In summary

The second step of the `sparta classification` command involves the post-processing of the results from the repeated classification and variable selection processes. This involves the identification of the optimal level of variable selection, and the calculation of the robust, confident and candidate selection subsets as defined in Section 3.1.3. These selection subsets are this step's main output; those of the optimal selection level are highlighted, but all of the obtained lists are made available.

5.3 Usage of the pipeline.

At time of submission of this thesis, the latest version of SPARTA is coded in Python, and requires the following Python packages to function: pandas [139], numpy [148], scikit-learn [121], scipy [147], matplotlib [145], joblib [142], seaborn [146], tqdm [149], goatools [141], Biopython [140], requests [150], shap [85] and kneebow [122]. An installation of the EsMeCaTa pipeline [67], along with its dependencies, is also required. The pipeline will also depend on an internet connection, in order to query and download from various databases.

The pipeline can be installed using pip, then launched from the command line. Three commands can be used to run SPARTA: `sparta esmecata`, `sparta classification` and `sparta pipeline`. Each command has optional arguments, with default values aligned with the parameters used in our previous manipulations. These options allow the user to adapt the parameters of the pipeline to their preference, facilitate the prospect of a reproduction of previous results, or allow them to adapt to specific use cases such as the application of the pipeline to a functional profile alone instead of a taxonomic input.

5.3.1 sparta esmecata

The `sparta esmecata` command requires the following arguments:

— "-p" (taxonomic abundance): a file indicating the abundance of the organisms, described according to their taxonomic affiliations, in the samples. This abundance profile should be in the form of a .txt file with tabular separation, describing the microbiota's composition per sample identificator. Metadata can be included in the table if the input is a taxonomic abundance file, it will not be taken account of. Taxon

names should be in the format: "k__kingdom|p__phylum|c__class|o__order|f__family|g__genus|s__species", a format derived from MetaPhlan's [33] output presentation, but that can also easily be built from QIIME [36] or FROGS [38] outputs, for example.

— "-o" (output): the name or path to the desired output folder.

The following optional arguments are also available:

— "-t", "–treatment": Data treatment for the functional table. Can be: 'tf_igm' (see 2.1.3), defaults to no treatment.

— "-s", "–scaling": Scaling method to apply to the taxonomic table. Can be: 'relative', default: no scaling.

— "–eggnog": Path to the eggnog database to be used by the EsMeCaTa pipeline. If not given, the pipeline will be launched with the 'UniProt' workflow by default.

— "–keep_temp": This option allows the user to keep temporary files at the end of the run.

— "–update_ncbi": This option allows the user to force an update of the local NCBI database. It is particularly recommended when running EsMeCaTa for the first time.

— "–esmecata_results": If a run of EsMeCaTa on the dataset has already been performed, it is possible to give the corresponding 'annotation_reference' folder with this option to avoid launching EsMeCaTa and directly compute the functional profile.

— "–esmecata_relaunch": This is a flag that allows the user to force a re-run of the EsMeCaTa pipeline over an already existing output. This is notably intended for cases where a previous run of the pipeline was botched at this step.

`sparta esmecata` will give the following main output:

— **Calculated functional profile:** The scores of functional annotations, calculated during the process, are made available with the other outputs.

It will also output the following files and folders, which can be used as input for the `sparta classification` command:

— functional_occurrence.tsv: a tsv file indicating the occurrence of functions in organisms.

— otu_table_stripped.tsv: a tsv file indicating the abundance of organisms in sam-

ples, after filtering of the input metadata.
— taxonomic_affiliations.tsv: a tsv file indicating the taxonomic affiliations of the organisms.
— EsMeCaTa_outputs: outputs of the EsMeCaTa pipeline, only the results of the 'annotation' step are kept for storage efficiency.

5.3.2 sparta classification

The `sparta classification` command works with the following compulsory inputs:
— "-l" (label): a csv file indicating the label of each sample to make the classification. Headers should indicate the name of each sample, and the associated label should be given underneath.
— "-o" (output): the name or path to the desired output folder.

In complement, at least one of the following inputs must also be given:
— "-fp" (functional profile): a csv file indicating the abundance of functions in samples. This can be built using the `sparta esmecata` command, or on the basis of another functional profiling approach (i.e. HUMAnN or PiCRUSt, for example).
— "-tp" (taxonomic profile): a csv file indicating the abundance of organisms in samples. This profile must not include metadata. The 'otu_table_stripped.tsv' output of the `sparta esmecata` command can be used for this purpose.

It is possible to apply this command only to a taxonomic or functional profile. In this case, post-processing steps that involve correlating these two profiles to each other (listing associated counterparts, calculating the strength of associations between taxa and annotations) will be skipped. Functional profiles can be built using `sparta esmecata`, but can also be formatted on the basis of functional profilings built by other tools. Depending on the inputs given at this stage, the following files can also be required:
— "-fo" (functional occurence): a tsv file measuring the expression of functions by an organisms. Using EsMeCaTa, this can be the number of proteins within an organism's proteome that express a given function. Using HUMAnN, this can be the intermediary functional scores of each annotation per species. This input is required if both taxonomic and functional profiles are given as input. The 'functional_occurrence.tsv' output of the `sparta esmecata` command can be used here.
— "-ta" (taxonomic affiliations): a tsv file indicating the taxonomic affiliations of the

organisms. The 'taxonomic_affiliations.tsv' output of the `sparta esmecata` command can be used for this purpose. This input is required if a taxonomic profile is given as input.

The following optional arguments are also available:

— "-r", "–runs": amount of pipeline runs. Defaults to 10 runs.
— "-i", "–iterations": number of iterations of the method. Defaults to 5 iterations.
— "-c", "–classifiers": amount of trained classifiers per iteration of the command. Defaults to 20 classifiers.
— "-m", "–method": classifying method to be run. Default to RF ('rf' value), but can also handle SVMs ('svm' value)
— "-v", "–variable_ranking": if the value of "-m" is "rf", this indicates the method for RF variable importance ranking. The default is Gini importance ('gini' value), but it can also handle SHAP ('shap' value)
— "–reference_test_sets": path to reference test sets (csv file) allowing the user to give their own test sets to be used during classification.
— "–preselected-organisms": if a taxonomic profile was given, the user can use this argument to specify organisms to be preselected at each run (should link to a csv file containing this information).
— "–preselected-annotations": if a functional profile was given, the user can use this argument to specify annotations to be preselected at each run (should link to a csv file containing this information).
— "–seed": the classification process is seeded, as explained in Section 5.2.4. The master seed to define the randomness of these steps, for reproducibility, can be defined here. Default value: 42.

`sparta classification` will give the following main outputs:

— **Classification performances:** The performances of the functional and taxonomic profiles at their respective best selective iteration are plotted. Detailed performance metrics of each RF model trained during each run and iteration on the training, validation and test subsets of the taxonomic and functional data, and their optimal found parameters are also made available.
— **Selected variables:** All variables selected at each iteration of each run of the pipeline are listed and available as output. A compilation of the robust, confident and candidate variables are also available for all iteration levels, with the results of

each profile's optimal iteration set aside in a separate folder. Information concerning each variable (name, expression per label, associated counterparts) are also given.

The following files and folders are also given, for purposes of visualization and transparency of the results:

— median_OTU_vs_SoFA_(best_vs_best).png: graphical representation of the classification performances (median ROC AUC per run) at the optimal selective iteration for both taxonomic and functional profiles. Both performance distributions are compared statistically by a Mann-Whitney U-test, the p-value of which is given in the figure's title. The optimal selection levels for both profiles are also given.

— Test_sets.csv: sample IDs used as test sets for each run of the pipeline. This file can be re-used as such on a later run of SPARTA on the same dataset to ensure that the same test sets are used.

5.3.3 sparta pipeline

The `sparta pipeline` command runs `sparta esmecata` and `sparta classification` in succession. It takes in the compulsory inputs of each of these commands, and gives all of their outputs. It can be tuned with the optional arguments of both commands. It will run the `sparta classification` command with both a taxonomic and a functional profile, using the outputs of `sparta esmecata` directly.

5.4 Conclusion and discussion.

In order to make this thesis' works reproducible and accessible, we have implemented a software that automates the entire process. This implementation covers the calculation of a functional profile from a taxonomic table, a repeated process of iterative RF-based classification and automatic selection, and a post-processing step to measure the robustness of the selected variables. The required inputs are a taxonomic abundance table and labels associated to each sample, and the outputs include the calculated functional profiling, classification performances and the robust, confident and candidate variables from the dataset, complemented by information about their nomenclature, presence in each profile, and associations to their robust counterparts. Parameters such as the amount of runs and iterations to be performed, or the amount of forests trained per iteration, can be modified to the user's convenience. Quality of life options have also been implemented, to

allow the user to choose their own test subsets or run the pipeline directly and exclusively on an input functional profile. The process is fully seeded, to ensure reproducibility both at the level of classifier training, and of the selection of test and validation sets.

This implementation makes it possible to run a complete analysis of the gut microbiota on the taxonomic and functional scale, and retains information concerning the interconnections between the two, which allows us to make the manipulations shown in Chapters 2, 3 and 4 fully reproducible. More generally, SPARTA automatizes and makes accessible a novel method for gut microbiota analysis, that requires a minimal amount of computing resources and time to be effectuated compared to other state of the art methods. Generally, it remains an open question to choose the right trade-off between computation time, classification performance and interpretability when handling microbiota data. The modular implementation of SPARTA, allowing the user to directly specify functional profiles, aims at providing the corresponding flexibility to adjust the pipeline to the type of raw data (MGS or 16S data) or the phenotype of interest.

Pipeline availability, and material for result reproduction.

A full implementation of SPARTA is available at `https://github.com/baptisteruiz/SPARTA`. The repository also includes inputs and instructions for the reproduction of the results of Chapter 3 ('article_data' folder). This is made possible by the sharing of the intermediary results obtained during our manipulations. Indeed, a full relaunch of the SPARTA pipeline could potentially give different results, as updates to the UniProt database notably could result in different functional profilings. As such, the adaptability of SPARTA's implementation, which allows it to refer to previous intermediary results, makes it possible to perpetrate the reproduction of any previous results, while still giving the opportunity to refer to the latest versions of the external resources if the user desires.

Chapter 6

Conclusion

6.1 Conclusion.

During the course of this thesis, we have presented several significant contributions to the field of gut microbiota analysis, aiming to expand upon previous explorations to propose an accessible method for robust ML-based classification and interpretation both on the taxonomic and functional scale. Overall, we have offered advances in the methodologies for gut microbiota functional representation and classification, offering robust, interpretable, and comprehensive tools for health-related research.

This methodological development led to the implementation of an open-source software: the SPARTA pipeline. Special attention was put into the seeding of the process, to make our results fully reproducible. Broader options, allowing for the compatibility of the software with other functional profiling tools for example, were also implemented. This implementation facilitates the broader application and verification of our findings, but also provides opportunity for future research and development in the field of gut microbiota analysis, as it requires little in terms of computing resources and only takes a few days to complete the entirety of its process. SPARTA groups all of this thesis' other contributions.

This analysis remains accessible, both in terms of the software being open and of its requirements for functionality, and, if applied to a wider array of data, could open new perspectives in understanding how the microbiota expresses its most impactful metabolic signatures.

6.1.1 Functional profiles of the gut microbiota from taxonomic profilings.

The gut microbiota has most commonly been explored at the taxonomic level, in accordance with the initial intuition that gut health is conditioned by the balance of the

microbial species that compose it. While this approach has allowed the characterization of several taxa of interest in the context of gut health, and has contributed largely to the development of probiotic-based therapies, the stakes are now shifting towards the understanding of the gut microbiota on a functional scale. Indeed, in order to progress our understanding of the gut microbiota further, we need to better understand *what* gut microbes do, more so than *who* they are.

As such, tools for the functional profiling of microbial communities are at the center of important methodological stakes, as they are the entry point to this new paradigm. Several processes exist to build functional profiles from genome sequencing data, but these existing tools are also very demanding in terms of computational resources. Previous studies have, however, made available taxonomic profiles, built from sequences. These profiles could make for a lighter entry point for functional exploration of the gut microbiota, as they would spare downstream users the need to process heavy sequence data.

As a first contribution for this thesis, we have developed a novel method for functional representation of the gut microbiota, which distinguishes itself from existing approaches such as PiCRUSt and HUMAnN by enabling conversion solely from taxonomic profiles. This method is versatile, applicable to profiles derived from both 16S rRNA and MGS sequencing. Our comparative analysis demonstrated that the EsMeCaTa pipeline, upon which our method is based, extracts more comprehensive information than HUMAnN and avoids redundancy between functional and taxonomic profiles. All of this is also done at a lesser computational cost, as this new process takes a much lighter input and runs much faster than its sequence-based counterpart. EsMeCaTa only generates associations between taxa and FAs; as such, a novel approach was proposed to translate these associations into a quantification of each annotation's expression within the gut microbiota.

Thanks to its reliance on a reference-based approach like EsMeCaTa instead of a sequence-based approach, this new method makes gut microbiota analysis more efficient, and therefore more accessible. The required input in the form of a taxonomic table is much lighter than the MGS sequences that are required to run HUMAnN. The difference in run time between both methods, to retrieve comparable functional information, also contributes to making this new approach more practical, in cases where a taxonomic profile is already available for the studied dataset.

6.1.2 Classification through automated variable selection.

The gut microbiota has been largely explored as a basis for ML classification, which has allowed for the identification of approaches best suited for this task. Among these, RFs have shown particular potential, for both classification performance, as well as their inherent capacity for variable ranking. This property of the classifier has been exploited by previous works to conduct variable selection, in order to allow for biological interpretation of the results, as well as improving classification performances through the correction of the data's initially unfavorable dimensionality.

Though these approaches for variable selection have shown their benefits, they still had limitations. The first was the bias induced by user-defined parameters, which introduced empirical decisions into the process and, as such, limited the adaptability of the approach when compared to a fully automated method. The second was the lack of insight into the robustness of the selections, with little regard into how a variation of the training conditions could impact its results.

As a second contribution for this thesis, we introduced a new method for classifying individuals based on their gut microbiota profiles, incorporating both taxonomic and functional features. This approach adapts the problem's dimensionality through a novel method for automated variable selection based on RF variable importance, showing that both profile types have comparable potential for describing an individual's health status. Our method notably enhances performance, especially for functional profiles, and prioritizes robustness and interpretability through an extension of the state-of-the-art approaches in three dimensions, involving re-training of the classifiers, iteration of the variable selection, and repetition of the entire training and selection process.

This approach's efficiency and innovation are based on two factors: the novel approach to variable selection that it applies, and its repetition which allows us to evaluate the robustness of the obtained variable selections, ranking each selected feature based on how resilient it is to the variation of the training context. The resulting shortlist of robustly significant variables offers deeper insights than traditional tools like limma, paving the way for further biological interpretation and application, notably taking account of the intercorrelations between taxa and the metabolic functions they express, which is a seldom explored aspect of the data.

6.1.3 The functional scale for interpretability.

Previous forays into the interpretation of the biological signal at the top of RF variable importance rankings had been made in the context of the gut microbiota. This usually amounted to presenting a list of the classifier's top features, the dimensions of which were empirically chosen, and identifying a few variables that are coherent with the problem at hand among them. These analyses were usually a point of discussion for these studies, and therefore weren't exhaustive because they were not meant to be their main focus. Also, even in cases where both the taxonomic and functional scales of the gut microbiota were studied, the relationships between taxa and their FAs were never explored, despite there being an important incentive to understanding these dynamics.

We addressed these interrogations through our third contribution: a validation of our method's robust functional selection by means of a thorough bibliographic research, which confirmed the relevance of the robust selection in regard to the state-of-the-art biological knowledge. We also pushed this further by explicitating the links between selected taxons and annotations, allowing us to confirm the presence of a functional cumulation effect within taxonomic profiles.

Although demonstrated on a single example as proof of concept, this finding underscores the importance of studying the gut microbiota at the functional level, beyond taxonomic signatures alone. Thanks to its reliance on the reference-based EsMeCaTa approach, linking both descriptions to one another was also made easier, which allowed us to illustrate the gains in insight that can be generated from exploring the taxonomic and functional profiles in conjunction.

6.2 Hints to guide perspectives : preliminary studies on a real case study.

In context of this thesis, explorations were conducted surrounding the application of SPARTA to a dataset of gut microbiota samples gathered from Crohn's disease patients by our partners in CHU Rennes, and the interpretation of its conclusions in light of their expertise. The cohort in question consists of 567 samples, taken from 383 individuals diagnosed with Crohn's disease. Each individual's microbiota was sampled at CHU Rennes between 1 and 6 times, and descriptions of the patient's heath status and records were taken in complement (age, sex, body mass index, surgical and antibiotic antecedents,

disease severity). Samples were sequenced using the 16S rRNA approach.

This work was not presented in this manuscript because definitive results could not be obtained in time for submission. The explorations that could be conducted did however clarify some practical aspects of SPARTA's application to data outside of our benchmark datasets, and put forward methodological challenges that, though they could not be tackled in the time frame of this thesis, should be addressed in its continuity.

6.2.1 A confirmation of SPARTA's compatibility with 16S data.

A first application of SPARTA to an earlier version of the dataset on which some of the samples were yet to be added, differentiated patients in the 'remission' category from the others, with moderate success (AUC around 0.65 for both taxonomic and functional profiles). This initial manipulation did however confirm that SPARTA was applicable to 16S data, as EsMeCaTa was capable of functionally profiling the inputs without issues.

This first approach also opened the door for a test of SPARTA's adaptability in light of the medical staff's demands and expertise, as the question onward became not to differentiate remission patients from others, but rather to analyze whether a functional annotation could explain the betterment or worsening of an individual's condition.

6.2.2 A first test of SPARTA's compatibility with problems integrating temporality.

Further explorations were made with the final version of the data, with an adaptation of the problem to more precise medical questions. Notably, in order to evaluate the impact of all variables on the progression of the disease, we focused on individuals who had given more than one sample, and only looked at the second sample onward. Each sample was graded depending on the evolution of the patient's health status in comparison to their previous visit: 0 if the health state did not change, 1 if an aggravation of the disease was observed, or 2 if it had receded. This selection reduced the analyzed sample pool to 184 from the initial 567.

The ensuing application to this data confirmed SPARTA's compatibility with multi-label problems. Classification performances remained underwhelming however (AUC around 0.6 for both functional and taxonomic profiles), and the application did not produce any robust annotations from the 8,410 FAs that were recovered by EsMeCaTa. A confident selection, containing 30 annotations, was however obtained, and the analysis of its contents

revealed interesting and coherent functional information, as validated by our partners' expertise.

These disappointing results could be due to the problem at hand being too complex. Notably, we believe that using a static information (the composition of a patient's microbiota at a moment t) to predict a dynamic label (the evolution of the disease state between t-1 and t) may not have been the correct approach. Due to the more nuanced definitions of the new labels, which describe a dynamic rather than a control/unhealthy split, we could also expect the difference between individuals to be more subtle or variable, which could explain both the under-performance of the model and its lack of a robust subset. Redundancy among the annotations obtained in the functional profile may also be a hindering factor. As such, further tuning of the inputs was envisioned to adapt their relevance to the problem.

6.2.3 Lessons and unexplored ideas.

From these experiments, more manipulations had been envisaged to correct the limitations that were still observable in our results. Notably, a better integration of the problem's temporal aspect was looked into. An idea was to look at the difference in expression between samples, in an effort to represent the evolution of the microbiota, which was to be correlated with that of the patient's health state. The integration of the available metadata, notably surrounding treatment history, was also looked into.

These leads, though they could not be applied to completion in context of this thesis, remain natural extensions to the work presented in this manuscript. The applications described in the previous paragraphs have allowed us to confirm SPARTA's adaptability when it comes to the input data (compatibility with 16S), and its compatibility with multi-label classification. Their pursuit would allow for a more complete showcase of SPARTA's potential for analysis in context of a real medical situation.

6.3 Perspectives

The presented contributions constitute a first step towards a better integration of functional data into the analysis of gut microbiota data. The resulting approach remains, however, perfectible in several ways, and opens many perspectives for further work.

6.3.1 A compromise to be found between performance and information.

SPARTA's criterion for optimal variable selection is to retain the subset which generates the best classification metric. Though this constitutes a strong basis for a first approach, previous works have also warned of it being potentially deceptive, and encouraged to investigate the significance of the evolution in performance measurements [113]. As such, a more nuanced vision of our selection approach should be envisioned. As is, we would recommend that, in cases where the information contained in the 'robust' shortlists appears insufficient to a SPARTA user, they also look at the 'confident' and potentially 'candidate' selections. For future work, a fine-tuning of our selection approach could be implemented, with the establishment of a new criterion to identify the optimal iterative selection level of our approach. Said criterion should take account of model performance, but also combine it with a measurement the output annotation lists' redundancy and informative content. Metrics akin to Semantic Similarity and Information Content [151] could be leveraged to approach these latter issues.

6.3.2 Expanding the functional information through the semantic web.

The FAs produced by our method are derived from various levels within their respective ontologies. While the structure of EC numbers explicitates their hierarchy, the relationships between GO terms are much less obvious. As such, the combination of SPARTA's outputs with a visualization method adapted for both of the employed nomenclatures would be a complement to our outputs, allowing for a more intuitive exploration of their biological ramifications. However, this dual hierarchy can also lead to redundancy within the functional profiles, as information about a particular GO term may overlap with its child terms. To address this, leveraging ontologies to group related annotations (those sharing a common ancestor or within the same pathway) could enhance our process in several ways.

Firstly, by representing all levels of annotation ancestry within a dataset and then performing variable selection based on importance, we could gain insights into which level of description most effectively characterizes the gut microbiota. While our current method focuses on the annotations identified by EsMeCaTa, it is plausible that higher-level descriptions could be more efficient. Grouping annotations at a higher level might reduce

the dimensionality of the input data, thereby improving classification performance. Additionally, higher-level descriptions could enhance interpretability. Currently, the specificity of SPARTA's outputs necessitates recontextualization to fully understand their impact. More generalized annotations might provide clearer insights and facilitate a more straightforward interpretation of the data.

Secondly, just as we demonstrated that taxons might cumulatively influence the microbiota's metabolism by occupying the same functional niche, overly specific descriptions could dilute significant signals. For instance, several annotations within the same category may individually have little influence, but collectively they might represent a strong signal. An approach that utilizes ontologies could identify whether such dynamics are present and significant. Grouping related annotations could reveal hidden patterns and interactions within the data, leading to a more robust understanding of the gut microbiota's functional landscape.

As such, integrating ontological information to consolidate related annotations offers a promising avenue for improving both the accuracy and interpretability of our functional profiles. This approach could streamline the input data, enhance classification performance, and provide deeper insights into the functional dynamics of the gut microbiota. Future research should explore this potential, aiming to refine our methods and broaden our understanding of the microbiota's role in health and disease.

6.3.3 Integration of further medical metadata.

Beyond FAs, there are numerous additional data sources that can be integrated when exploiting a taxonomic table to enhance the understanding and classification of gut microbiota profiles. For instance, the phylogenetic relationships among taxa present in a sample can provide supplementary information at the taxonomic level. Integrating phylogenetic data with information about taxa's metabolic activity has been explored in several studies, generally employing neural network approaches. This work has demonstrated that such integration can significantly improve classification performance [152]. Given these promising results, developing a method to incorporate phylogenetic relationships into a RF framework could similarly enhance the performance of these classifiers. This approach could provide a more comprehensive view of the evolutionary and functional context of the taxa, leading to more accurate and robust predictions.

Furthermore, clinical datasets often come with a wealth of medical metadata, which includes detailed information about each individual's characteristics, such as age, weight,

and medication history. Leveraging this clinical metadata can offer substantial benefits.The integration of this information could help classification, especially when they lead to a rapid and significant change in microbiota composition. For instance, the menstrual cycle [153], diet [154], or antibiotic treatments [155] could be recognized and accounted for by the models. More importantly, this integration can help in identifying signals that are highly relevant to specific biological questions. For example, understanding whether a medical treatment positively impacts a patient's health through its effects on the gut microbiota could be significantly enhanced by incorporating such metadata. This could lead to more personalized and effective treatment strategies based on individual patient profiles.

To fully realize the potential of these additional data sources, future research should focus on developing robust methods for their integration. The integration of heterogeneous biological data for ML is a complex field of research itself, and several approaches have been explored in this context [156–158]. For the question at hand, approaches based on Multi-Kernel Learning [159], through which each description of the micriobiota is used to compute distance matrices which can then be combined through a weighted sum, appear as one of the most promising leads.

6.3.4 Exploring applications beyond the gut microbiota.

Though it was developed and tested on gut microbiota data, SPARTA's genericity is such that it can be applied to any data describing a microbial community. Health subjects related to other microbiotas, of the skin, mouth or nasal cavity for example, could also be explored through this approach. Beyond human health, SPARTA could also contribute insights in the domain of plant biology, with applications to the algae microbiome for instance, or in bioengineering, for example to better understand the functional intricacies within the microbial community of a methanizer. The stakes of integrating functional information when exploring microbial communities range beyond the applications presented in this thesis, and further contributions to this approach could prove useful in fields beyond ours.

Bibliography

1. Ruiz, B. *et al.,* SPARTA: Interpretable functional classification of microbiomes and detection of hidden cumulative effects, *PLOS Computational Biology* **20,** 1–36, https://doi.org/10.1371/journal.pcbi.1012577 (Nov. 2024).
2. Ozen, M. & Dinleyici, E., The history of probiotics: the untold story, *Beneficial Microbes* **6,** 159–165, ISSN: 1876-2891, http://dx.doi.org/10.3920/BM2014.0103 (Jan. 2015).
3. Pinart, M. *et al.,* Gut Microbiome Composition in Obese and Non-Obese Persons: A Systematic Review and Meta-Analysis, *Nutrients* **14,** ISSN: 2072-6643, https://www.mdpi.com/2072-6643/14/1/12 (2022).
4. Cao, S.-Y. *et al.,* Dietary plants, gut microbiota, and obesity: Effects and mechanisms, *Trends in Food Science & Technology* **92,** 194–204, ISSN: 0924-2244, https://www.sciencedirect.com/science/article/pii/S0924224419300226 (2019).
5. Aldars-García, L., Chaparro, M. & Gisbert, J. P., Systematic Review: The Gut Microbiome and Its Potential Clinical Application in Inflammatory Bowel Disease, *Microorganisms* **9,** ISSN: 2076-2607, https://www.mdpi.com/2076-2607/9/5/977 (2021).
6. Rebersek, M., Gut microbiome and its role in colorectal cancer, *BMC Cancer* **21,** 1325, ISSN: 1471-2407, https://doi.org/10.1186/s12885-021-09054-2 (Dec. 2021).
7. Fujita, K. *et al.,* Gut microbiome and prostate cancer, *International Journal of Urology* **29,** 793–798, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iju.14894, https://onlinelibrary.wiley.com/doi/abs/10.1111/iju.14894 (2022).
8. Ruo, S. W. *et al.,* Role of gut Microbiota dysbiosis in breast cancer and novel approaches in prevention, diagnosis, and treatment, *Cureus* **13,** e17472 (Aug. 2021).
9. Forsythe, P., Bienenstock, J. & Kunze, W. A., Vagal pathways for microbiome-brain-gut axis communication, en, *Adv. Exp. Med. Biol.* **817,** 115–133 (2014).

10. Evrensel, A. & Ceylan, M. E., The gut-brain axis: The missing link in depression, en, *Clin. Psychopharmacol. Neurosci.* **13,** 239–244 (Dec. 2015).

11. Nocera, A. & Nasrallah, H. A., The Association of the Gut Microbiota with Clinical Features in Schizophrenia, *Behavioral Sciences* **12,** ISSN: 2076-328X, https://www.mdpi.com/2076-328X/12/4/89 (2022).

12. Bostancikliоğlu, M., The role of gut microbiota in pathogenesis of Alzheimer's disease, *Journal of Applied Microbiology* **127,** 954–967, ISSN: 1364-5072, eprint: https://academic.oup.com/jambio/article-pdf/127/4/954/48996325/jambio0954.pdf, https://doi.org/10.1111/jam.14264 (Oct. 2019).

13. Yang, D. *et al.,* The role of the gut Microbiota in the pathogenesis of Parkinson's disease, en, *Front. Neurol.* **10,** 1155 (Nov. 2019).

14. Hoption Cann, S. A., van Netten, J. P. & van Netten, C., Dr William Coley and tumour regression: a place in history or in the future, en, *Postgrad. Med. J.* **79,** 672–680 (Dec. 2003).

15. Gebrayel, P. *et al.,* Microbiota medicine: towards clinical revolution, en, *J. Transl. Med.* **20,** 111 (Mar. 2022).

16. Afzaal, M. *et al.,* Human gut microbiota in health and disease: Unveiling the relationship, *Frontiers in Microbiology* **13,** ISSN: 1664-302X, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.999001 (2022).

17. Fan, Y. & Pedersen, O., Gut microbiota in human metabolic health and disease, *Nature Reviews Microbiology* **19,** 55–71, ISSN: 1740-1534, https://doi.org/10.1038/s41579-020-0433-9 (Jan. 2021).

18. Morales, E. F. & Escalante, H. J., *in Biosignal Processing and Classification Using Computational Learning and Intelligence* (eds Torres-García, A. A., Reyes-García, C. A., Villaseñor-Pineda, L. & Mendoza-Montoya, O.) 111–129 (Academic Press, 2022), ISBN: 978-0-12-820125-1, https://www.sciencedirect.com/science/article/pii/B9780128201251000178.

19. Nadella, G. S., Satish, S., Meduri, K. & Meduri, S. S., A Systematic Literature Review of Advancements, Challenges and Future Directions of AI And ML in Healthcare, *International Journal of Machine Learning for Sustainable Development* **5,** 115–130, https://www.ijsdcs.com/index.php/IJMLSD/article/view/519 (2023).

20. Shen, D., Wu, G. & Suk, H.-I., Deep learning in medical image analysis, en, *Annu. Rev. Biomed. Eng.* **19,** 221–248 (June 2017).

21. Daugaard Jørgensen, M., Antulov, R., Hess, S. & Lysdahlgaard, S., Convolutional neural network performance compared to radiologists in detecting intracranial hemorrhage from brain computed tomography: A systematic review and meta-analysis, en, *Eur. J. Radiol.* **146,** 110073 (Jan. 2022).

22. Rajkomar, A., Dean, J. & Kohane, I., Machine learning in medicine, en, *N. Engl. J. Med.* **380,** 1347–1358 (Apr. 2019).

23. Sutton, R. T. *et al.,* An overview of clinical decision support systems: benefits, risks, and strategies for success, en, *NPJ Digit. Med.* **3,** 17 (Feb. 2020).

24. Mehta, R., Jain, R. K. & Badve, S., Personalized medicine: the road ahead, en, *Clin. Breast Cancer* **11,** 20–26 (Mar. 2011).

25. Marcos-Zambrano, L. J. *et al.,* Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment, *Frontiers in Microbiology* **12,** ISSN: 1664-302X, `https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.634511` (2021).

26. Hughes, G., On the mean accuracy of statistical pattern recognizers, *IEEE Transactions on Information Theory* **14,** 55–63 (1968).

27. Statnikov, A. *et al.,* A comprehensive evaluation of multicategory classification methods for microbiomic data, *Microbiome* **1,** 11, ISSN: 2049-2618, `https://doi.org/10.1186/2049-2618-1-11` (Apr. 2013).

28. Basu, S., Kumbier, K., Brown, J. B. & Yu, B., Iterative random forests to discover predictive and stable high-order interactions, *Proceedings of the National Academy of Sciences* **115,** 1943–1948, eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.1711236115`, `https://www.pnas.org/doi/abs/10.1073/pnas.1711236115` (2018).

29. Heintz-Buschart, A. & Wilmes, P., Human Gut Microbiome: Function Matters, *Trends in Microbiology* **26,** 563–574, ISSN: 0966-842X, `https://www.sciencedirect.com/science/article/pii/S0966842X17302512` (2018).

30. Inkpen, S. A. *et al.,* The coupling of taxonomy and function in microbiomes, en, *Biol. Philos.* **32,** 1225–1243 (Dec. 2017).

31. Fraher, M. H., O'Toole, P. W. & Quigley, E. M. M., Techniques used to characterize the gut microbiota: a guide for the clinician, *Nature Reviews Gastroenterology & Hepatology* **9,** 312–322, ISSN: 1759-5053, https://doi.org/10.1038/nrgastro.2012.44 (June 2012).

32. Yi, X. *et al.,* Unravelling the enigma of the human microbiome: Evolution and selection of sequencing technologies, *Microbial Biotechnology* **17,** e14364 MICROBIO-2023-468, e14364, eprint: https://enviromicro-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/1751-7915.14364, https://enviromicro-journals.onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.14364 (2024).

33. Segata, N. *et al.,* Metagenomic microbial community profiling using unique clade-specific marker genes, *Nature Methods* **9,** 811–814, ISSN: 1548-7105, https://doi.org/10.1038/nmeth.2066 (Aug. 2012).

34. Beghini, F. *et al.,* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3, *eLife* **10** (eds Turnbaugh, P., Franco, E. & Brown, C. T.) e65088, ISSN: 2050-084X, https://doi.org/10.7554/eLife.65088 (May 2021).

35. Blanco-Míguez, A. *et al.,* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4, *Nature Biotechnology* **41,** 1633–1644 (Nov. 2023).

36. Bolyen, E. *et al.,* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, *Nature Biotechnology* **37,** 852–857 (Aug. 2019).

37. Schloss, P. D. *et al.,* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities, *Applied and Environmental Microbiology* **75,** 7537–7541, ISSN: 1098-5336, http://dx.doi.org/10.1128/AEM.01541-09 (Dec. 2009).

38. Escudié, F. *et al.,* FROGS: Find, Rapidly, OTUs with Galaxy Solution, *Bioinformatics* **34,** 1287–1294, ISSN: 1367-4803, https://doi.org/10.1093/bioinformatics/btx791 (Apr. 2018).

39. Langmead, B. & Salzberg, S. L., Fast gapped-read alignment with Bowtie 2, en, *Nat. Methods* **9,** 357–359 (Mar. 2012).

40. DeSantis, T. Z. *et al.,* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, en, *Appl. Environ. Microbiol.* **72,** 5069–5072 (July 2006).

41. Quast, C. *et al.,* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, *Nucleic Acids Research* **41,** D590–D596, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/41/D1/D590/3690367/gks1219.pdf, https://doi.org/10.1093/nar/gks1219 (Nov. 2012).

42. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F., VSEARCH: a versatile open source tool for metagenomics, en, *PeerJ* **4,** e2584 (Oct. 2016).

43. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M., Swarm: robust and fast clustering method for amplicon-based studies, *PeerJ* **2,** e593, ISSN: 2167-8359, http://dx.doi.org/10.7717/peerj.593 (Sept. 2014).

44. CaMaCho, C. & CoulouriS, G., v. avaGYaN, N. Ma, J. papadopouloS, k. Bealer, aNd tl MaddeN. 2009. BLAST+: Architecture and applications, *BMC Bioinformatics* **10,** 421.

45. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R., Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Applied and environmental microbiology* **73,** 5261–5267 (2007).

46. Chorlton, S. D., Ten common issues with reference sequence databases and how to mitigate them, *Frontiers in Bioinformatics* **4,** ISSN: 2673-7647, https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2024.1278228 (2024).

47. Abubucker, S. *et al.,* Metabolic reconstruction for metagenomic data and its application to the human microbiome, en, *PLoS Comput. Biol.* **8,** e1002358 (June 2012).

48. Franzosa, E. A. *et al.,* Species-level functional profiling of metagenomes and metatranscriptomes, *Nature Methods* **15,** 962–968, ISSN: 1548-7105, https://doi.org/10.1038/s41592-018-0176-y (Nov. 2018).

49. Langille, M. G. I. *et al.,* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences, *Nature Biotechnology* **31,** 814–821 (Sept. 2013).

50. Douglas, G. M. *et al.,* PICRUSt2 for prediction of metagenome functions, *Nature Biotechnology* **38,** 685–688, ISSN: 1546-1696, https://doi.org/10.1038/s41587-020-0548-6 (June 2020).

51. Sayers, E. W. *et al.,* Database resources of the national center for biotechnology information, *Nucleic Acids Research* **50,** D20–D26, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/50/D1/D20/42058080/gkab1112.pdf, https://doi.org/10.1093/nar/gkab1112 (Dec. 2021).

52. Suzek, B. E. *et al.,* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics* **31,** 926–932, ISSN: 1367-4803, eprint: https://academic.oup.com/bioinformatics/article-pdf/31/6/926/49011550/bioinformatics_31_6_926.pdf, https://doi.org/10.1093/bioinformatics/btu739 (Nov. 2014).

53. Buchfink, B., Xie, C. & Huson, D. H., Fast and sensitive protein alignment using DIAMOND, *Nature Methods* **12,** 59–60, ISSN: 1548-7105, https://doi.org/10.1038/nmeth.3176 (Jan. 2015).

54. Tatusov, R. L., Koonin, E. V. & Lipman, D. J., A Genomic Perspective on Protein Families, *Science* **278,** 631–637, eprint: https://www.science.org/doi/pdf/10.1126/science.278.5338.631, https://www.science.org/doi/abs/10.1126/science.278.5338.631 (1997).

55. Mistry, J. *et al.,* Pfam: The protein families database in 2021, *Nucleic Acids Research* **49,** D412–D419, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/49/D1/D412/35363969/gkaa913.pdf, https://doi.org/10.1093/nar/gkaa913 (Oct. 2020).

56. Kanehisa, M. & Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research* **28,** 27–30, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/28/1/27/9895154/280027.pdf, https://doi.org/10.1093/nar/28.1.27 (Jan. 2000).

57. Ashburner, M. *et al.,* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, eng, *Nature genetics* **25,** 10802651[pmid], 25–29, ISSN: 1061-4036, https://pubmed.ncbi.nlm.nih.gov/10802651 (May 2000).

58. Consortium, T. G. O., The Gene Ontology resource: enriching a GOld mine, *Nucleic Acids Research* **49,** D325–D334, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/49/D1/D325/35364517/gkaa1113.pdf, https://doi.org/10.1093/nar/gkaa1113 (Dec. 2020).

59. Caspi, R., Dreher, K. & Karp, P. D., The challenge of constructing, classifying, and representing metabolic pathways, en, *FEMS Microbiol. Lett.* **345,** 85–93 (Aug. 2013).

60. Caspi, R. *et al.,* The MetaCyc database of metabolic pathways and enzymes - a 2019 update, en, *Nucleic Acids Res.* **48,** D445–D453 (Jan. 2020).

61. Ye, Y. & Doak, T. G., A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes, *PLOS Computational Biology* **5,** 1–8, https://doi.org/10.1371/journal.pcbi.1000465 (Aug. 2009).

62. Markowitz, V. M. *et al.,* IMG: the Integrated Microbial Genomes database and comparative analysis system, en, *Nucleic Acids Res.* **40,** D115–22, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/40/D1/D115/9475546/gkr1044.pdf, https://doi.org/10.1093/nar/gkr1044 (Jan. 2012).

63. Finn, R. D., Clements, J. & Eddy, S. R., HMMER web server: interactive sequence similarity searching, *Nucleic Acids Research* **39,** W29–W37, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/39/suppl_2/W29/7628106/gkr367.pdf, https://doi.org/10.1093/nar/gkr367 (May 2011).

64. Barbera, P. *et al.,* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences, *Systematic Biology* **68,** 365–369, ISSN: 1063-5157, eprint: https://academic.oup.com/sysbio/article-pdf/68/2/365/27739127/syy054.pdf, https://doi.org/10.1093/sysbio/syy054 (Sept. 2018).

65. MIRARAB, S., NGUYEN, N. & WARNOW, T., *in Biocomputing 2012* 247–258 (), eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789814366496_0024, https://www.worldscientific.com/doi/abs/10.1142/9789814366496_0024.

66. Zaneveld, J. R. R. & Thurber, R. L. V., Hidden state prediction: a modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses, *Frontiers in Microbiology* **5,** ISSN: 1664-302X, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2014.00431 (2014).

67. Belcour, A., Ruiz, B., Frioux, C., Blanquart, S. & Siegel, A., EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations, *bioRxiv,* eprint: https://www.biorxiv.org/content/early/2022/03/18/2022.03.16.484574.full.pdf, https://www.biorxiv.org/content/early/2022/03/18/2022.03.16.484574 (2022).

68. Consortium, T. U., UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research* **49,** D480–D489, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf, https://doi.org/10.1093/nar/gkaa1100 (Nov. 2020).

69. Steinegger, M. & Söding, J., MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, *Nature Biotechnology* **35** (Oct. 2017).

70. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J., eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale, *Molecular Biology and Evolution* **38,** 5825–5829, ISSN: 1537-1719, eprint: https://academic.oup.com/mbe/article-pdf/38/12/5825/41714617/msab293.pdf, https://doi.org/10.1093/molbev/msab293 (Oct. 2021).

71. Huerta-Cepas, J. *et al.,* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, *Nucleic Acids Research* **47,** D309–D314, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/47/D1/D309/27437484/gky1085.pdf, https://doi.org/10.1093/nar/gky1085 (Nov. 2018).

72. Knights, D., Costello, E. K. & Knight, R., Supervised classification of human microbiota, en, *FEMS Microbiol. Rev.* **35,** 343–359 (Mar. 2011).

73. Zou, H. & Hastie, T., Regularization and Variable Selection Via the Elastic Net, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67,** 301–320, ISSN: 1369-7412, eprint: https://academic.oup.com/jrsssb/article-pdf/67/2/301/49795094/jrsssb_67_2_301.pdf, https://doi.org/10.1111/j.1467-9868.2005.00503.x (Mar. 2005).

74. Costello, E. K. *et al.,* Bacterial community variation in human body habitats across space and time, en, *Science* **326,** 1694–1697 (Dec. 2009).

75. Fierer, N. *et al.,* Forensic identification using skin bacterial communities, en, *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6477–6481 (Apr. 2010).

76. Alekseyenko, A. V. *et al.,* Community differentiation of the cutaneous microbiota in psoriasis, *Microbiome* **1,** 31, ISSN: 2049-2618, https://doi.org/10.1186/2049-2618-1-31 (Dec. 2013).

77. Nossa, C. *et al.,* Foregut microbiome in development of esophageal adenocarcinoma, *Nature Precedings,* ISSN: 1756-0357, https://doi.org/10.1038/npre.2010.5026.1 (Oct. 2010).

78. Cortes, C. & Vapnik, V., Support-vector networks, *Machine learning* **20,** 273–297 (1995).

79. Vovk, V., *in Empirical inference* 105–116 (Springer, 2013).

80. Breiman, L., Random Forests, *Machine Learning* **45,** 5–32, ISSN: 1573-0565, https://doi.org/10.1023/A:1010933404324 (Oct. 2001).

81. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M., *in Data Mining in Agriculture* 83–106 (Springer New York, New York, NY, 2009), ISBN: 978-0-387-88615-2, https://doi.org/10.1007/978-0-387-88615-2_4.

82. Specht, D. F., Probabilistic neural networks, *Neural Networks* **3,** 109–118, ISSN: 0893-6080, https://www.sciencedirect.com/science/article/pii/089360809090049Q (1990).

83. Gini, C., *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.]* https://books.google.fr/books?id=fqjaBPMxB9kC (Tipogr. di P. Cuppini, 1912).

84. Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P., *Understanding variable importances in Forests of randomized trees in*, **26** (Dec. 2013).

85. Lundberg, S. M. & Lee, S.-I., *A Unified Approach to Interpreting Model Predictions in Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

86. Hanley, J. A. & McNeil, B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** PMID: 7063747, 29–36, eprint: https://doi.org/10.1148/radiology.143.1.7063747, https://doi.org/10.1148/radiology.143.1.7063747 (1982).

87. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N., Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights, *PLOS Computational Biology* **12,** 1–26, https://doi.org/10.1371/journal.pcbi.1004977 (July 2016).

88. Zeller, G. *et al.,* Potential of fecal microbiota for early-stage detection of colorectal cancer, *Molecular Systems Biology* **10,** 766, eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20145645, https://www.embopress.org/doi/abs/10.15252/msb.20145645 (2014).

89. Qin, N. *et al.,* Alterations of the human gut microbiome in liver cirrhosis, *Nature* **513,** 59–64, ISSN: 1476-4687, https://doi.org/10.1038/nature13568 (Sept. 2014).

90. Le Chatelier, E. *et al.,* Richness of human gut microbiome correlates with metabolic markers, *Nature* **500,** 541–546, ISSN: 1476-4687, https://doi.org/10.1038/nature12506 (Aug. 2013).

91. Qin, J. *et al.,* A human gut microbial gene catalogue established by metagenomic sequencing, *Nature* **464,** 59–65, ISSN: 1476-4687, https://doi.org/10.1038/nature08821 (Mar. 2010).

92. Qin, J. *et al.,* A metagenome-wide association study of gut microbiota in type 2 diabetes, *Nature* **490,** 55–60, ISSN: 1476-4687, https://doi.org/10.1038/nature11450 (Oct. 2012).

93. Karlsson, F. H. *et al.,* Gut metagenome in European women with normal, impaired and diabetic glucose control, *Nature* **498,** 99–103, ISSN: 1476-4687, https://doi.org/10.1038/nature12198 (June 2013).

94. Oh, M. & Zhang, L., DeepMicro: deep representation learning for disease prediction based on microbiome data, *Scientific Reports* **10,** 6026, ISSN: 2045-2322, https://doi.org/10.1038/s41598-020-63159-5 (Apr. 2020).

95. LaValle, S. M., Branicky, M. S. & Lindemann, S. R., On the relationship between classical grid search and probabilistic roadmaps, *The International Journal of Robotics Research* **23,** 673–692 (2004).

96. F.R.S., K. P., LIII. On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2,** 559–572 (1901).

97. Rumelhart, D. E., Hinton, G. E. & Williams, R. J., *in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* 318–362 (MIT Press, Cambridge, MA, USA, 1986), ISBN: 026268053X.

98. Jones, C. M. A. *et al.,* Bacterial Taxa and Functions Are Predictive of Sustained Remission Following Exclusive Enteral Nutrition in Pediatric Crohn's Disease, *Inflammatory Bowel Diseases* **26,** 1026–1037, ISSN: 1078-0998, eprint: https://academic.oup.com/ibdjournal/article-pdf/26/7/1026/33402521/izaa001.pdf, https://doi.org/10.1093/ibd/izaa001 (Jan. 2020).

99. Douglas, G. M. *et al.,* Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease, *Microbiome* **6,** 13, ISSN: 2049-2618, https://doi.org/10.1186/s40168-018-0398-3 (Jan. 2018).

100. Efron, B., *in The jackknife, the bootstrap and other resampling: Plans* (Society for Industrial and Applied Mathematics, 1982).

101. *in*, *Encyclopedia of Machine Learning* (eds Sammut, C. & Webb, G. I.) 600–601 (Springer US, Boston, MA, 2010), ISBN: 978-0-387-30164-8, https://doi.org/10.1007/978-0-387-30164-8_469.

102. Hansen, R. *et al.,* The Microaerophilic Microbiota of De-Novo Paediatric Inflammatory Bowel Disease: The BISCUIT Study, *PLOS ONE* **8,** 1–10, https://doi.org/10.1371/journal.pone.0058825 (Mar. 2013).

103. Love, M. I., Huber, W. & Anders, S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology* **15,** 550, ISSN: 1474-760X, https://doi.org/10.1186/s13059-014-0550-8 (Dec. 2014).

104. Ritchie, M. E. *et al.,* limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research* **43,** e47–e47, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf, https://doi.org/10.1093/nar/gkv007 (Jan. 2015).

105. Chen, P. *et al.,* Interaction between endometrial microbiota and host gene regulation in recurrent implantation failure, *Journal of Assisted Reproduction and Genetics* **39,** 2169–2178, ISSN: 1573-7330, https://doi.org/10.1007/s10815-022-02573-2 (Sept. 2022).

106. Wipperman, M. F. *et al.,* Gastrointestinal microbiota composition predicts peripheral inflammatory state during treatment of human tuberculosis, *Nature Communications* **12,** 1141, ISSN: 2041-1723, https://doi.org/10.1038/s41467-021-21475-y (Feb. 2021).

107. Zhang, B. *et al.,* Integrated multi-omics identified the novel intratumor microbiome-derived subtypes and signature to predict the outcome, tumor microenvironment heterogeneity, and immunotherapy response for pancreatic cancer patients, *Frontiers in Pharmacology* **14,** ISSN: 1663-9812, https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1244752 (2023).

108. Qi, Y., *in Ensemble Machine Learning: Methods and Applications* (eds Zhang, C. & Ma, Y.) 307–323 (Springer New York, New York, NY, 2012), ISBN: 978-1-4419-9326-7, https://doi.org/10.1007/978-1-4419-9326-7_11.

109. Svetnik, V., Liaw, A., Tong, C. & Wang, T., *Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules in Multiple Classifier Systems* (eds Roli, F., Kittler, J. & Windeatt, T.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004), 334–343, ISBN: 978-3-540-25966-4.

110. Soneson, C. & Robinson, M. D., Bias, robustness and scalability in single-cell differential expression analysis, *Nature Methods* **15,** 255–261, ISSN: 1548-7105, https://doi.org/10.1038/nmeth.4612 (Apr. 2018).

111. Shahjaman, M., Manir Hossain Mollah, M., Rezanur Rahman, M., Islam, S. S. & Nurul Haque Mollah, M., Robust identification of differentially expressed genes from RNA-seq data, *Genomics* **112,** 2000–2010, ISSN: 0888-7543, https://www.sciencedirect.com/science/article/pii/S0888754319305750 (2020).

112. Stupnikov, A. *et al.,* Robustness of differential gene expression analysis of RNA-seq, *Computational and Structural Biotechnology Journal* **19,** 3470–3481, ISSN: 2001-0370, https://www.sciencedirect.com/science/article/pii/S200103702100221X (2021).

113. Kursa, M. B., Robustness of Random Forest-based gene selection methods, *BMC Bioinformatics* **15,** 8, ISSN: 1471-2105, https://doi.org/10.1186/1471-2105-15-8 (Jan. 2014).

114. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T., REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms, *PLOS ONE* **6,** 1–9, https://doi.org/10.1371/journal.pone.0021800 (July 2011).

115. Kanehisa, M., Sato, Y. & Kawashima, M., KEGG mapping tools for uncovering hidden features in biological data, *Protein Science* **31,** 47–53, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4172, https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4172 (2022).

116. Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J. & Churpek, M. M., Comparison of variable selection methods for clinical predictive modeling, *International Journal of Medical Informatics* **116,** 10–17, ISSN: 1386-5056, https://www.sciencedirect.com/science/article/pii/S1386505618305811 (2018).

117. Bairoch, A., The ENZYME database in 2000, *Nucleic Acids Research* **28,** 304–305, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/28/1/304/9895166/280304.pdf, https://doi.org/10.1093/nar/28.1.304 (Jan. 2000).

118. Chen, K., Zhang, Z., Long, J. & Zhang, H., Turning from TF-IDF to TF-IGM for term weighting in text classification, *Expert Systems with Applications* **66,** 245–260, ISSN: 0957-4174, https://www.sciencedirect.com/science/article/pii/S0957417416304870 (2016).

119. Huttenhower, C. *et al.,* Structure, function and diversity of the healthy human microbiome, *Nature* **486,** 207–214, ISSN: 1476-4687, https://doi.org/10.1038/nature11234 (June 2012).

120. Madhikermi, M., Kubler, S., Robert, J., Buda, A. & Främling, K., *in*, 145–164 (2016), https://www.sciencedirect.com/science/article/pii/S095741741630330X.

121. Pedregosa, F. *et al.,* Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

122. Georg, U., *kneebow: Knee or elbow detection for curves* 2019, https://github.com/georg-un/kneebow.

123. Plaza-Díaz, J. *et al.,* The Gut Barrier, Intestinal Microbiota, and Liver Disease: Molecular Mechanisms and Strategies to Manage, *International Journal of Molecular Sciences* **21,** ISSN: 1422-0067, https://www.mdpi.com/1422-0067/21/21/8351 (2020).

124. Wei, X. *et al.,* Abnormal Gut Microbiota Metabolism Specific for Liver Cirrhosis, *Frontiers in Microbiology* **9,** ISSN: 1664-302X, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.03051 (2018).

125. Xiong, Y. *et al.,* Dynamic Alterations of the Gut Microbial Pyrimidine and Purine Metabolism in the Development of Liver Cirrhosis, *Frontiers in Molecular Biosciences* **8,** ISSN: 2296-889X, https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2021.811399 (2022).

126. Miller, A. R., North, J. A., Wildenthal, J. A. & Tabita, F. R., Two distinct aerobic methionine salvage pathways generate volatile methanethiol in Rhodopseudomonas palustris, *MBio* **9** (Apr. 2018).

127. Li, Z. *et al.,* Methionine metabolism in chronic liver diseases: an update on molecular mechanism and therapeutic implication, *Signal Transduction and Targeted Therapy* **5,** 280, ISSN: 2059-3635, https://doi.org/10.1038/s41392-020-00349-7 (Dec. 2020).

128. Lynch, S. V. & Pedersen, O., The human intestinal microbiome in health and disease, en, *N. Engl. J. Med.* **375,** 2369–2379 (Dec. 2016).

129. Thomas, S. *et al.,* The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists, *Cancer Research* **77,** 1783–1812, ISSN: 0008-5472, eprint: https://aacrjournals.org/cancerres/article-pdf/77/8/1783/2764278/1783.pdf, https://doi.org/10.1158/0008-5472.CAN-16-2929 (Apr. 2017).

130. Wang, W. *et al.,* Increased Proportions of Bifidobacterium and the Lactobacillus Group and Loss of Butyrate-Producing Bacteria in Inflammatory Bowel Disease, *Journal of Clinical Microbiology* **52** (ed Forbes, B. A.) 398–406, ISSN: 1098-660X, http://dx.doi.org/10.1128/jcm.01500-13 (Feb. 2014).

131. Sugihara, K., Morhardt, T. L. & Kamada, N., The Role of Dietary Nutrients in Inflammatory Bowel Disease, *Frontiers in Immunology* **9,** ISSN: 1664-3224, https://www.frontiersin.org/articles/10.3389/fimmu.2018.03183 (2019).

132. Chang, A. *et al.,* BRENDA, the ELIXIR core data resource in 2021: new developments and updates, *Nucleic Acids Research* **49,** D498–D508, ISSN: 0305-1048, eprint: https://academic.oup.com/nar/article-pdf/49/D1/D498/35364686/gkaa1025.pdf, https://doi.org/10.1093/nar/gkaa1025 (Nov. 2020).

133. Chakaroun, R. M., Massier, L. & Kovacs, P., Gut Microbiome, Intestinal Permeability, and Tissue Bacteria in Metabolic Disease: Perpetrators or Bystanders?, *Nutrients* **12,** ISSN: 2072-6643, https://www.mdpi.com/2072-6643/12/4/1082 (2020).

134. Karen P. Scott Antoine Jean-Michel, T. M. & van Hemert, S., Manipulating the gut microbiota to maintain health and treat disease, *Microbial Ecology in Health and Disease* **26,** PMID: 25651995, 25877, eprint: https://www.tandfonline.com/doi/pdf/10.3402/mehd.v26.25877, https://www.tandfonline.com/doi/abs/10.3402/mehd.v26.25877 (2015).

135. Howell, D. C., *in International Encyclopedia of Statistical Science* (ed Lovric, M.) 250–252 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), ISBN: 978-3-642-04898-2, https://doi.org/10.1007/978-3-642-04898-2_174.

136. Zhang, T., Ji, X., Lu, G. & Zhang, F., The potential of Akkermansia muciniphila in inflammatory bowel disease, *Applied Microbiology and Biotechnology* **105,** 5785–5794, ISSN: 1432-0614, https://doi.org/10.1007/s00253-021-11453-1 (Aug. 2021).

137. Xiao, Y., Huang, Q., Wu, Z. & Chen, W., Roles of protein ubiquitination in inflammatory bowel disease, *Immunobiology* **225,** 152026, ISSN: 0171-2985, https://www.sciencedirect.com/science/article/pii/S0171298520305489 (2020).

138. Lee, J. S. *et al.,* Microbiota-Sourced Purines Support Wound Healing and Mucous Barrier Function, *iScience* **23,** 101226, ISSN: 2589-0042, https://www.sciencedirect.com/science/article/pii/S2589004220304119 (2020).

139. McKinney, W., *Data Structures for Statistical Computing in Python in Proceedings of the 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) (2010), 56–61.

140. Cock, P. J. *et al.,* Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Bioinformatics* **25,** 1422–1423 (2009).

141. Klopfenstein, D. V. *et al.,* GOATOOLS: A Python library for Gene Ontology analyses, *Scientific Reports* **8,** 10872, ISSN: 2045-2322, https://doi.org/10.1038/s41598-018-28948-z (July 2018).

142. Joblib Development Team, *Joblib: running Python functions as pipeline jobs* 2020, https://joblib.readthedocs.io/.

143. Lundberg, S. M. *et al.,* From local explanations to global understanding with explainable AI for trees, *Nature Machine Intelligence* **2,** 2522–5839 (2020).

144. Schisterman, E. F., Faraggi, D., Reiser, B. & Hu, J., Youden Index and the optimal threshold for markers with mass at zero, *Statistics in Medicine* **27,** 297–315, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2993, https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2993 (2008).

145. Hunter, J. D., Matplotlib: A 2D graphics environment, *Computing in Science & Engineering* **9,** 90–95 (2007).

146. Waskom, M. L., seaborn: statistical data visualization, *Journal of Open Source Software* **6,** 3021, https://doi.org/10.21105/joss.03021 (2021).

147. Virtanen, P. *et al.,* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods* **17,** 261–272 (2020).

148. Harris, C. R. *et al.,* Array programming with NumPy, *Nature* **585,** 357–362, https://doi.org/10.1038/s41586-020-2649-2 (Sept. 2020).

149. Da Costa-Luis, C. *et al., tqdm: A fast, Extensible Progress Bar for Python and CLI* version v4.66.4, May 2024, https://doi.org/10.5281/zenodo.11107065.

150. Requests Development Team, *Requests: HTTP for Humans* 2023, https://requests.readthedocs.io/en/latest/.

151. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A., Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics* **19,** 1275–1283, ISSN: 1367-4803, eprint: https://academic.oup.com/bioinformatics/article-pdf/19/10/1275/48903852/bioinformatics_19_10_1275.pdf, https://doi.org/10.1093/bioinformatics/btg153 (July 2003).

152. Monshizadeh, M. & Ye, Y., Incorporating metabolic activity, taxonomy and community structure to improve microbiome-based predictive models for host phenotype prediction, *Gut Microbes* **16,** PMID: 38214657, 2302076, eprint: https://doi.org/10.1080/19490976.2024.2302076, https://doi.org/10.1080/19490976.2024.2302076 (2024).

153. Schieren, A., Koch, S., Pecht, T. & Simon, M.-C., Impact of Physiological Fluctuations of Sex Hormones During the Menstrual Cycle on Glucose Metabolism and the Gut Microbiota, *Experimental and clinical endocrinology & diabetes : official journal, German Society of Endocrinology [and] German Diabetes Association* **0** (Feb. 2024).

154. David, L. *et al.,* Diet rapidly and reproducibly alters the gut microbiome, *Nature* **505** (Dec. 2013).

155. Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R. & Stanton, C., Impact of antibiotics on the human microbiome and consequences for host health, *MicrobiologyOpen* **11,** e1260, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mbo3.1260, https://onlinelibrary.wiley.com/doi/abs/10.1002/mbo3.1260 (2022).

156. Gligorijević, V. & Pržulj, N., Methods for biological data integration: perspectives and challenges, *Journal of The Royal Society Interface* **12,** 20150571, eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2015.0571, https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2015.0571 (2015).

157. Li, Y., Wu, F.-X. & Ngom, A., A review on machine learning principles for multi-view biological data integration, *Briefings in Bioinformatics* **19,** 325–340, ISSN: 1477-4054, eprint: https://academic.oup.com/bib/article-pdf/19/2/325/25524236/bbw113.pdf, https://doi.org/10.1093/bib/bbw113 (Dec. 2016).

158. Li, P., Luo, H., Ji, B. & Nielsen, J., Machine learning for data integration in human gut microbiome, *Microbial Cell Factories* **21,** 241, ISSN: 1475-2859, https://doi.org/10.1186/s12934-022-01973-4 (Nov. 2022).

159. Gönen, M. & Alpaydın, E., Multiple kernel learning algorithms, *The Journal of Machine Learning Research* **12,** 2211–2268 (2011).

List of acronyms

ASV: Amplicon Sequence Variant

AUC: Area Under the Receiver Operating Characteristic Curve

CDSS: Clinical Decision Support Systems

EC: Enzyme Commission

EHR: Electronic Health Records

FA: functional annotation

GO: Gene Ontology

IBD: Inflammatory Bowel Disease

IMG: Integrated Microbial Genomes

KEGG: Kyoto Encyclopedia of Genes and Genomes

LOOCV: Leave One Out Cross-Validation

MGS: Shotgun Metagenomic Sequencing

ML: Machine Learning

OTU: Operational Taxonomic Unit

PCA: Principal Component Analysis

RF: Random Forests

RFVS: RF-based backward elimination procedure

SCFA: Short-Chain Fatty Acid

SoFA: Score of Functional Annotation

SPARTA: Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes

SVM: Support Vector Machines

T2D: Type 2 Diabetes (Chinese cohort)

WT2D: Type 2 Diabetes (European women cohort)

Appendices

Appendix A

This appendix contains the full details of the Jaccard similarity heatmap presented in Chapter 2 (see Section 2.3.2 and Figure 2.4).

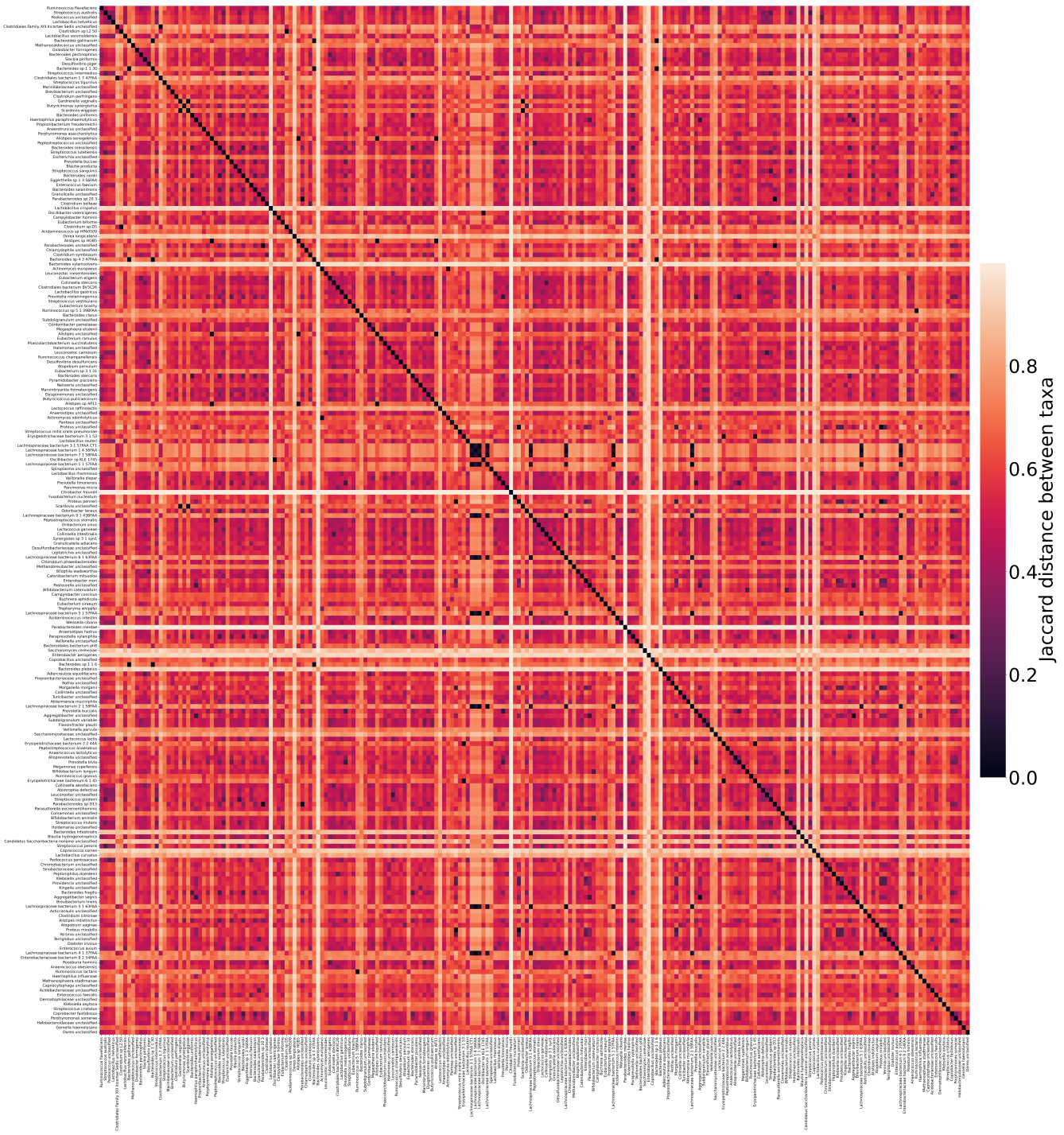


Figure S1 – **Jaccard distances calculated between the functional profiles of all taxa from the IBD dataset, annotated by EsMeCaTa with UniProt: upper left quarter of the heatmap.** This figure is complementary with Supplementary Figures S2, S3 and S4

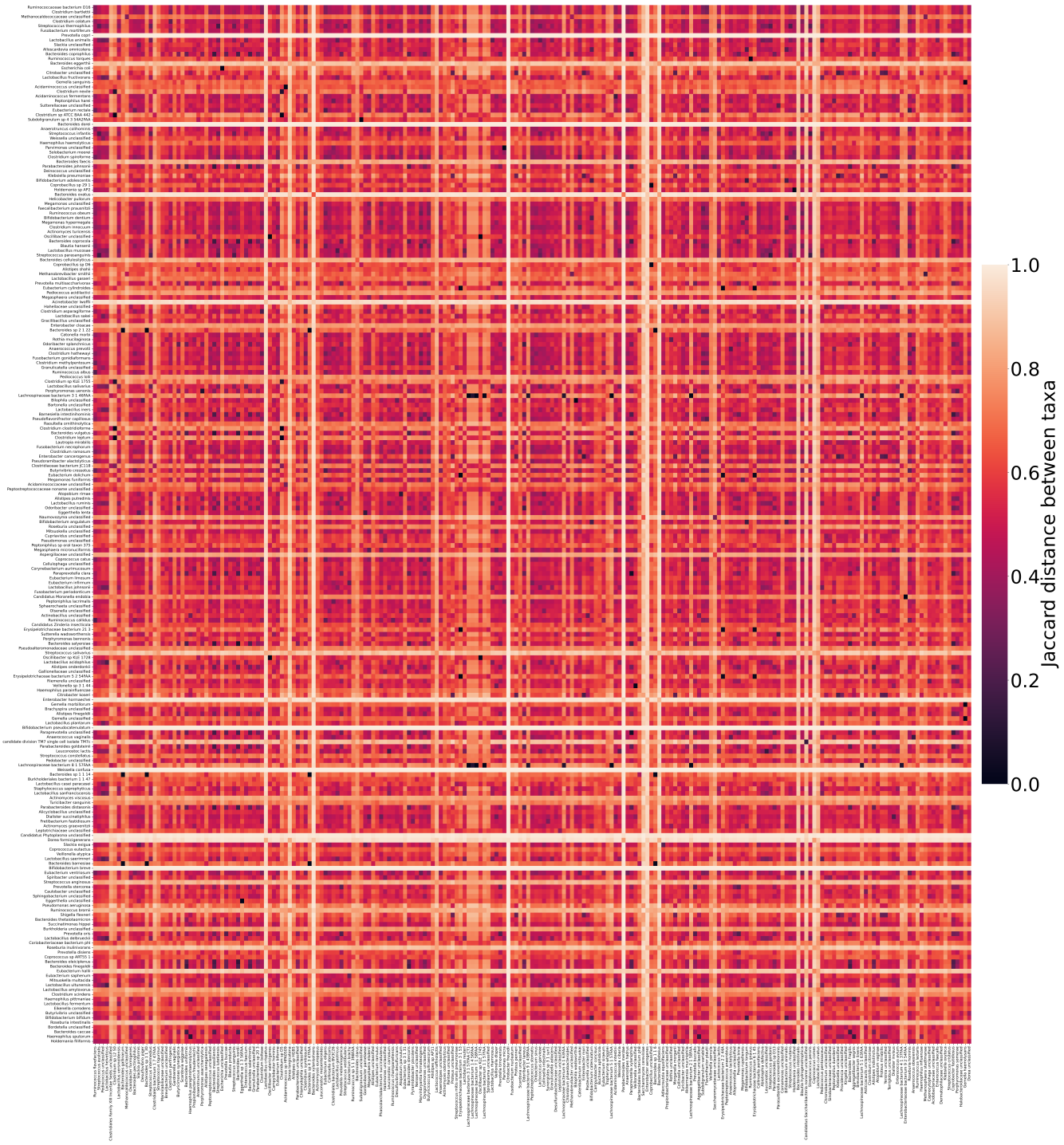


Figure S2 – **Jaccard distances calculated between the functional profiles of all taxa from the IBD dataset, annotated by EsMeCaTa with UniProt: lower left quarter of the heatmap.** This figure is complementary with Supplementary Figures S1, S3 and S4

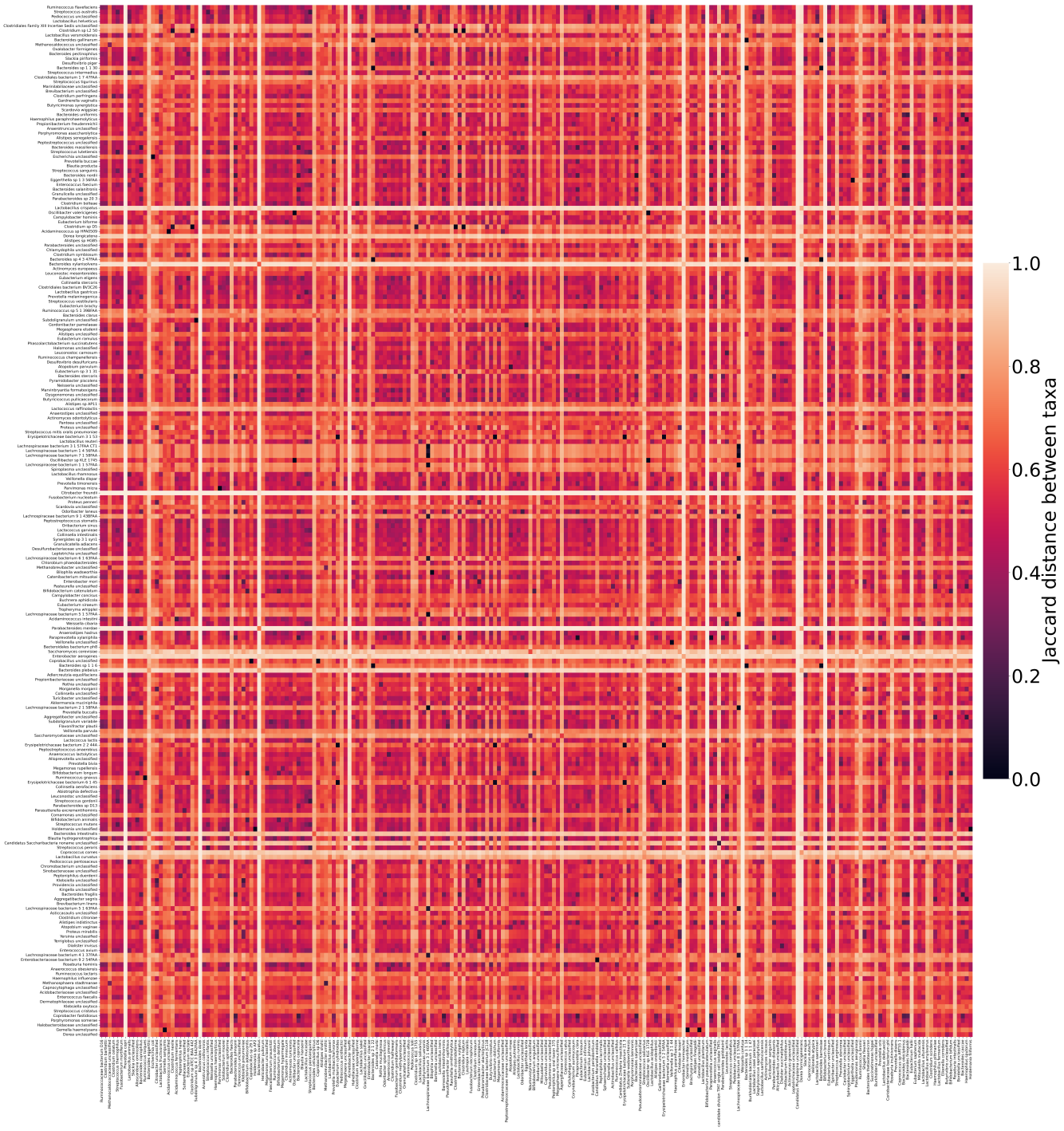


Figure S3 – **Jaccard distances calculated between the functional profiles of all taxa from the IBD dataset, annotated by EsMeCaTa with UniProt: upper right quarter of the heatmap.** This figure is complementary with Supplementary Figures S1, S2 and S4

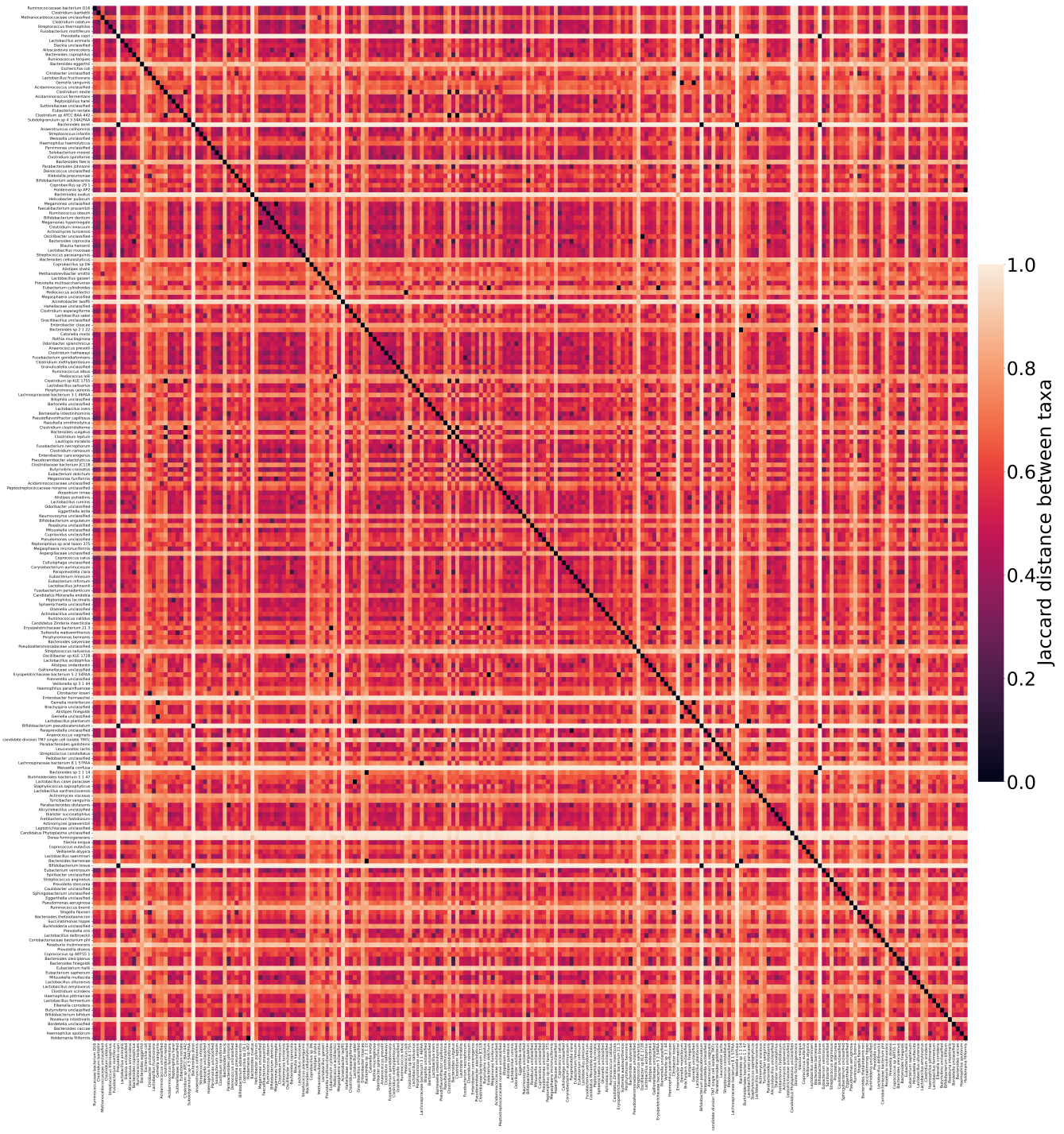


Figure S4 – **Jaccard distances calculated between the functional profiles of all taxa from the IBD dataset, annotated by EsMeCaTa with UniProt: liwer right quarter of the heatmap.** This figure is complementary with Supplementary Figures S1, S2 and S3

Appendix B

This appendix contains the full details of the bibliographic exploration conducted in Chapter 4 (see Section 4.2).

Table S1 – **Bibliographic research on the IBD dataset's Robust (Core-significant) functional annotations.**

ID	Names	Family	Bibliography category	Bibliography	Quotes	Bibliographic grade justification	Impacted organism	Family according to bibliography
GO:0102545	phosphatidylphospholipase B activity	1	1	https://doi.org/10.1016/j.bbrc.2008.03.087	'We have shown that LPC can induce recruitment of monocytes and pro-inflammatory cytokine production at nM concentrations.'	Lysophosphatidylcholine (LPC) is a product of the hydrolysis that Phosholipase B catalyses. It is also known as a pro-inflammatry agent, making it directly relevant to IBD (category 1).	Human (macrophages)	1
GO:0008744	L-xylulokinase activity	1	3	https://doi.org/10.3390/microorganisms10010167 (1), https://doi.org/10.1371/journal.pone.0178426 (2)	'various artificial sweeteners such as acesulfame potassium and aspartame can exacerbate the impairment of the intestinal mucus layer observed in CD' (ref 1) + 'L-xylulokinase, D-xylonolactonase and alpha-amylase, were also decreased in female animals administered Ace-K (acesulfame potassium).' (ref 2)	L-xyluokinase is known to be underexpressed in the gut microbiome of organisms that consume Acesulfame-potassium (Ace-K or ACK), an artificial sweetener, as shown by the second linked study. The other study reviews the effect of sweeteners in the diet on host health in the case of IBD, and flags Ace-K as an exacerbator of Crohn's and Colitis. As such, L-xyluokinase in the microbiota can be correlated to the presence of a known factor, but has no direct impact on the disease itself (category 3).	Microbiota	0
4.1.2.-	Not Found (Aldehyde Lyases)	1	2	https://doi.org/10.1155/2017/7685142 (1), https://doi.org/10.1007/s00535-016-1220-2 (2), https://doi.org/10.1096/fj.201800076RR (3)	'SPL (sphinganine-1-phosphate lyase, EC 4.1.2.27) deficiency in gut epithelial cells promotes colitis and colitis-associated carcinogenesis (CAC)' (ref 1) , indole synthase (EC 4.1.2.8): 'In other words, different approaches, such as the intake of foods that favor indole-inducing bacteria, microbiota that produce indoles from tryptophan, and indole compounds themselves, should be attempted for the prevention of disease as well as its maintenance and remission.' (ref 2), propioin synthase (EC 4.1.2.35): 'Decreased bacterial diversity characterizes the altered gut microbiome present in IBD [...] One potential biomarker is propanal (N.B: from propioin synthase) . It was markedly increased in abundance in samples from both acute and chronic colitis' (ref 3)	Though the annotation itself proved too generic to be correlated to the disease itself, several of its children (see citations) were found to be relevant to IBD (category 2). NOTE: some of these enzymes are benefic, and others are markers of the disease.	Microbiota (+fecal matter) + Human (epithelial cells)	N.A

GO:0006520	amino acid metabolic process	1	1	https://doi.org/10.3389/fimmu.2018.03183	'Intestinal inflammation affects several metabolic pathways and disturbances in amino acid metabolism are observed in IBD patients.', 'Metagenomic studies have revealed that amino acid biosynthesis genes are downregulated and amino acid transporter genes are upregulated in the gut microbiome of IBD patients, indicating that the gut microbiota lessens the production of amino acids and increases the rate of their utilization '	A direct bibliographic link between the annotation and IBD was established (category 1)	Microbiota	1
GO:0001510	RNA methylation	0	1	https://doi.org/10.3390/epigenomes4030016	'Our analysis resulted in five candidate genes corresponding to two of the major IBD subtypes: UBE2L3 and SLC22A4 for Crohn's Disease and TCF19, C6orf47 and SNAPC4 for Ulcerative Colitis. Further analysis using in silico predictions and co-expression analyses in combination with in vitro functional studies showed that our candidate genes seem to be regulated by m6A-dependent mechanisms. These findings provide the first indication of the implication of RNA methylation events in IBD pathogenesis.'	A direct bibliographic link between the annotation and IBD was established (category 1) NOTE: colitis involves more methylation on some genes, and less on others, therefore it is difficult to establish a direct link between methylation itself and the control/unhealthy status	Human (immune response)	N.A
GO:0008092	cytoskeletal protein binding	0	1	https://doi.org/10.1016/S0002-9440(10)63308-1	'Another important protein associated with enterocytic differentiation state is villin, an actin-binding cytoskeletal protein located within the microvilli of intestinal epithelial cells. It is well established that the expression of villin along the crypt-villus axis increases as cells migrate from the crypt to the villus surface concomitant with an increased rate of synthesis during enterocytic differentiation and considerable posttranslational stability. Decreased villin levels in CD and UC relative to HCs with the lowest concentration in RACE cells our data point to a disturbance of differentiation and maturation processes in RACE cells. This could also explain why antigen uptake is increased in these cells, because immature enterocytes are known to be more susceptible to antigen uptake than fully differentiated cells.'	A direct bibliographic link between the annotation and IBD was established (category 1) NOTE: (This is a little surprising to me: Do the actin-binding molecules of the bacteriae impact our own epithelial cells?)	Human (epithelial cells)	0

GO:0016746	acyltransferase activity	0	2	https://doi.org/10.1002/cbin.11362 (1), https://doi.org/10.1096/fj.14-255208 (2)	'GOAT (Ghrelin-O-acyltransferase) overexpression significantly enhanced the induction of colitis' (ref 1), 'N-acylethanolamines (NAEs) [...]are produced on demand from membrane phospholipids by the sequential actions of an N-acyltransferase and an NAPE-preferring phospholipase D (NAPE-PLD). [...][Results] would suggest a dysregulation in the production of NAEs during colon inflammation, with potentially reduced production.' (ref 2)	Though the annotation itself proved too generic to be correlated to the disease itself, several of its children (see citations) were found to be relevant to IBD (category 2). NOTE: some of these enzymes are benefic, and others are markers of the disease.	Human (intestinal mucosa, colon)	N.A
GO:0017065	single-strand selective uracil DNA N-glycosylase activity	0	1	https://doi.org/10.1172/JCI63338	Definition of Goterm: 'Single-strand selective monofunctional uracil DNA glycosylase (SMUG1) recognizes breaks in the genome and initiates repair.' + 'Increased levels of DNA base lesions have been documented in patients with chronic inflammatory conditions, including IBD ' (ref)	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (genome)	0
GO:0003953	NAD+ nucleosidase activity	0	1	https://doi.org/10.3390/antiox12061230	'In the case of IBD, the maintenance of intestinal homeostasis relies on a delicate balance between NAD+ biosynthesis and consumption. Consequently, therapeutics designed to target the NAD+ pathway are promising for the management of IBD.'+'NAD+ depletion in UC may result from increased NAD+ catabolism '	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (intestines)	1
GO:0016832	aldehyde-lyase activity	1	2	https://doi.org/10.1155/2017/7685142 (1), https://doi.org/10.1007/s00535-016-1220-2 (2), https://doi.org/10.1096/fj.201800076RR (3)	'SPL (sphinganine-1-phosphate lyase, EC 4.1.2.27) deficiency in gut epithelial cells promotes colitis and colitis-associated carcinogenesis (CAC)' (ref 1) , indole synthase (EC 4.1.2.8): 'In other words, different approaches, such as the intake of foods that favor indole-inducing bacteria, microbiota that produce indoles from tryptophan, and indole compounds themselves, should be attempted for the prevention of disease as well as its maintenance and remission.' (ref 2), propioin synthase (EC 4.1.2.35): 'Decreased bacterial diversity characterizes the altered gut microbiome present in IBD [...] One potential biomarker is propanal (N.B: from propioin synthase) . It was markedly increased in abundance in samples from both acute and chronic colitis' (ref 3)	Though the annotation itself proved too generic to be correlated to the disease itself, several of its children (see citations) were found to be relevant to IBD (category 2). NOTE: some of these enzymes are benefic, and others are markers of the disease.	Microbiota (+fecal matter) + Human (epithelial cells)	N.A

GO:0019677	NAD catabolic process	0	1	https://doi.org/10.3390/antiox12061230	'In the case of IBD, the maintenance of intestinal homeostasis relies on a delicate balance between NAD+ biosynthesis and consumption. Consequently, therapeutics designed to target the NAD+ pathway are promising for the management of IBD.'+'NAD+ depletion in UC may result from increased NAD+ catabolism '	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (intestines)	1
GO:0005727	extrachromosomal circular DNA	1	1	https://doi.org/10.3390/cells12151953	'These studies suggest that self-derived circulating DNA (both non-circular cfDNA in nucleosomes or naked, and eccDNA (Extrachromosomal circular DNA)) are able to induce and sustain the inflammation machinery'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (epithelial cells)	1
GO:0000310	xanthine phosphoribosyltransferase activity	1	3	https://doi.org/10.1186/s12866-023-02932-8 (1), https://doi.org/10.1016/j.isci.2020.101226 (2)	'Hypoxanthine phosphoribosyltransferase (Hpt), Adenine phosphoribosyltransferase (apt), and xanthine phosphoribosyltransferase (xpt) are the enzymes crucial to the purine salvage pathway.' (1st citation) + 'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis. MSPs support energy balance, proliferation, and mucous barrier function' (2nd citation)	The annotation itself doesn't have a direct link to IBD in the bibliography. It is however involved in the purine pathway, which is a known factor in intestinal mucosa repair. The annotation is therefore involved in a larger relevant mechanism (category 3)	Human (epithelial cells)	0
GO:0016297	acyl-[acyl-carrier-protein] hydrolase activity	0	3	https://doi.org/10.1111/1541-4337.12926	GO term synonym: acyl-ACP thioesterase activity (https://www.ebi.ac.uk/QuickGO/term/GO:0016297). 'Genes associated with both fatty acid profile determination and assembly of triglycerides were identified successfully as CnDGAT1, CnPDAT, CnFATB3 which code for the enzymes diacylglycerol acyltransferase, phospholipid:diacylglycerol acyltransferase, and acyl-ACP thioesterase class B, respectively, and these can further be utilized for producing crops with high MCT yield.' + 'MCTs were first introduced in 1950, especially for the treatment of malabsorption disorders. They have been beneficial in treating disorders, such as Crohn's and coeliac diseases, which are majorly associated with the inflammatory response of the intestines.'	When expressed in crops, acyl-ACP hydrolases are a precursor for Medium chain triglycerides (MCTs), which themselves are used as treatment for Crohn's disease. As such, an indirect link can be found between this annotation and the disease (category 3)	Human	0

GO:0004122	cystathionine beta-synthase activity	1	1	https://doi.org/10.1093/ecco-jcc/jjz027	'Decreased expression of CBS (cystathionine beta-synthase) propagates the pathogenesis of UC by exacerbating inflammation-induced intestinal barrier injury.'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (epithelial cells)	0
GO:0008788	alpha,alpha-phosphotrehalase activity	1	3	https://doi.org/10.1080/19490976.2020.1750273	'Clinical manifestations of CD (Clostridioides difficile) infection include diarrhea, pseudomembranous colitis, and in extreme cases, death.' + 'Mechanisms by which these variants utilize trehalose include hypersensitive de-repression of the phosphotrehalase, TreA gene, and more efficient trehalose transport from the extracellular space via the PtsT transporter.'	This annotation does not directly link to IBD, however it is known to be expressed by taxon Clostridioides difficile, which is a known aggressive pathogen responsible for colitis symptoms (category 3)	Microbiota	1
GO:0047605	acetolactate decarboxylase activity	1	3	PMID: 32509162 (1), https://doi.org/10.1002/mnfr.202300337 (2)	'Lactobacillus reuteri is a normal resident species of the healthy gut microflora that can prevent IBD by altering the intestinal micro-environment and the immune system' (ref 1), 'This study identifies the coding gene (aldB) of acetolactate decarboxylase (ALDC) as an important regulatory gene of the intracellular pH in Lactobacillus reuteri (L. reuteri)' (ref 2)	Though it does not appear to be directly linked to the disease in metabolic terms, this annotation is expressed by a taxon known to be beneficial, and used as a probiotic tratment for IBD, Lactobacillus reuteri (category 3)	Microbiota	0
GO:0043130	ubiquitin binding	0	1	https://doi.org/10.1016/j.imbio.2020.152026	'In the past few years, accumulative evidence illustrated that six E3 ubiquitin ligases, namely, ring finger protein (RNF) 183, RNF 20, A20, Pellino 3, TRIM62 and Itch, exhibited clear mechanisms in the development of IBD. They regulate the intestinal inflammation by facilitating the ubiquitination of targeted proteins which participate in different inflammatory signaling pathways.'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (signalling pathways)	0
GO:0047419	N-acetylgalactosamine-6-phosphate deacetylase activity	1	1	https://doi.org/10.3748/wjg.v13.i20.2826	This enzyme catalyzes the production of acetate (https://www.ebi.ac.uk/QuickGO/term/GO:0047419). 'Our findings suggest that propionate and acetate, in addition to butyrate, could be useful in the treatment of inflammatory disorders, including IBD.', 'it is also clearly demonstrated that acetate and propionate ameliorate an ongoing inflammatory response at the cellular level'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0

1.1.1.22	UDP-glucose 6-dehydrogenase.	0	1	https://doi.org/10.1136/gutjnl-2015-310705	'Energy metabolism was also identified herein to be altered in IBD, which includes the candidate biomarkers inorganic pyro-phosphatase, visfatin and UDP-glucose 6-dehydrogenase.'; Table 2 does not mark UDP-glucose 6-dehydrogenase as 'elevated in patients with IBD'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (colon)	0
GO:0004135	amylo-alpha-1,6-glucosidase activity	0	3	https://doi.org/10.3390/ijms22094381	'[GSDIII] diagnosis is made by identifying biallelic pathogenic variants in glycogen debranching enzyme, amylo-alpha-1,6-glucosidase, or 4-alpha-glucanotransferase (AGL)' + 'Populations with GSDI suffer from Crohn's-like inflammatory bowel disease (C-IBD)'	Amylo-alpha-1,6-glucosidase deficiency is known to cause GSDIII, a form of glycogen storage deficiency. Another form of the same pathology, GSDI, has shown to be a cause of IBD. Though it is far from direct, a potential link to the disease can be traced (category 3)	Human	0
GO:0008707	4-phytase activity	0	3	https://doi.org/10.3390/molecules26010031	4-phytase participates in the inositol phosphate metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=3.1.3.26) 'Inositol and its derivatives, as natural compounds, have shown a significant effect on inhibiting inflammation and carcinogenesis'	The annotation is implicated in a metabolic pathway, of which the product is an inflammation inhibitor (category 3)	Human (signalling pathways)	0
GO:0046537	2,3-bisphosphoglycerate-independent phosphoglycerate mutase activity	0	3	https://doi.org/10.1016/j.ygeno.2018.05.022	'Disease preventing property of different strains of B. coagulans demonstrated in separate studies include [...] reduction of symptoms associated with Clostridium difficile-induced colitis in mice' + Table 2: the enzyme is part of the genes detected in the probiotic	Though it does not appear to be directly linked to the disease in metabolic terms, this annotation is expressed by a taxon known to be beneficial against colitis, B. coagulans (category 3)	Microbiota	0
GO:0008610	lipid biosynthetic process	1	1	https://doi.org/10.1097/MIB.0000000000000394	'We demonstrate that a number of ether lipids (alkylphospholipid and plasmalogens) are significantly and negatively associated with CD. These alterations of lipid profiles particularly plasmalogens may contribute to the pathogenesis of IBD.'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (plasma)	0
GO:0042121	alginic acid biosynthetic process	0	1	https://doi.org/10.1111/j.1365-3083.2005.01571.x	'Our data suggest that LVA (low-viscosity sodium alginate, a salt of alginic acid) could potentially be a novel therapeutic option for inflammatory bowel disease.'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0

2.4.2.22	xanthine phosphoribosyltransferase.	1	3	https://doi.org/10.1186/s12866-023-02932-8 (1), https://doi.org/10.1016/j.isci.2020.101226 (2)	'Hypoxanthine phosphoribosyltransferase (Hpt), Adenine phosphoribosyltransferase (apt), and xanthine phosphoribosyltransferase (xpt) are the enzymes crucial to the purine salvage pathway.' (1st citation) + 'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis.MSPs support energy balance, proliferation, and mucous barrier function' (2nd citation)	The annotation itself doesn't have a direct link to IBD in the bibliography. It is however involved in the purine pathway, which is a known factor in intestinal mucosa repair. The annotation is therefore involved in a larger relevant mechanism (category 3)	Human (epithelial cells)	0
GO:0071702	organic substance transport	1	3	https://doi.org/10.1016/j.crohns.2011.08.003	'The IBD5 locus on the 5th chromosome was identified as conferring a 2 to 6 fold risk to develop CD. This region codes for two genes (SLC22A4 and SLC22A5) encoding the organic cation/carnitine transporters (OCTN) 1 and 2.'	A direct correlation between the annotation and the disease wasn't established, but the overexpression of genes coding for transporters in this family were found to be a risk factor for Crohn's (category 3)	Human (genetic)	1
GO:0032440	2-alkenal reductase [NAD(P)+] activity	0	3	https://doi.org/10.1016/j.freeradbiomed.2016.11.033	GO term synonym: NAD(P)H-dependent alkenal/one oxidoreductase activity (https://www.ebi.ac.uk/QuickGO/term/GO:0032440). 'Another route for the metabolic detoxification of HNE (4-hydroxy-2-nonenal) involves the reduction of the C2-C3 double bond by NAD(P)H-dependent alkenal/one oxidoreductase'+'HNE is converted to the epoxynonanal, which reacts with the amino groups of guanosine, adenosine and cytidine, and after cyclization, forms the etheno-DNA adducts 1,N2-etheno-2'-deoxyguanosine (?dG) 1,N6-etheno-2'-eoxyadenosine (?dA), and 3,N4-etheno-2'-deoxycytidine (?dC).' + 'Elevated etheno-DNA adduct levels have been found in the injured tissues of subjects with chronic pancreatitis, ulcerative colitis, and Crohn's disease'	This enzyme is involved in the detoxification of HNE (4-hydroxy-2-nonenal) which, at an elevated level in the cell, interacts with DNA coding sequences in a way that generates etheno-DNA, the presence of which has been correlated with IBD (category 3)	Human (cellular)	0

GO:0047356	CDP-ribitol ribitolphosphotransferase activity	1	2	https://doi.org/10.1073/pnas.0504084102	GO term synonym: teichoic-acid synthase activity (https://www.ebi.ac.uk/QuickGO/term/GO:0047356). 'Teichoic acids (TAs), and especially lipoteichoic acids (LTAs), are one of the main immunostimulatory components of pathogenic Gram-positive bacteria.' + 'Taken together, these results emphasize the importance of LTA composition in the proinflammatory or antiinflammatory properties of Lactobacillus cells, but also point to the potential for use of specific Lactobacillus cell wall mutants for treatment of IBD.'	The expression of this specific molecule in Lactobacillus plantarum has been proven to affect the host immunoresponse, making it capable of enhancing or downregulating inflammatory responses depending on their form (most notably, it has been shown to reduce inflammation when it is in a shape that inclused less d-Ala). (category 2)	Microbiota	N.A
GO:0008910	kanamycin kinase activity	1	1	https://doi.org/10.1016/S0016-5085(85)80015-9	"More recently, it has been demonstrated that a variety of antibiotic compounds can depress intestinal neuroeffector transmission in vitro and that those best able to accomplish this are those most often associated with antibiotic-associated colitis. Included in this category are [...] kanamycin"	A direct bibliographic link between the annotation's product and IBD was established (category 1) (when given as an antibiotic, kanamycin appears to favor IBD)	Human (intestinal neuroeffector)	1
2.8.3.10	citrate CoA-transferase.	1	1	https://doi.org/10.3748/wjg.v13.i20.2826	This enzyme catalyzes the production of acetate (https://enzyme.expasy.org/EC/2.8.3.10). 'Our findings suggest that propionate and acetate, in addition to butyrate, could be useful in the treatment of inflammatory disorders, including IBD.', 'it is also clearly demonstrated that acetate and propionate ameliorate an ongoing inflammatory response at the cellular level'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0

4.2.1.-	Not Found (Hydro-Lyases)	1	2	https://doi.org/10.1002/mnfr.202000461 (1), https://doi.org/10.1093/ecco-jcc/jjz027 (2), https://doi.org/10.1038/nrgastro.2017.104 (3)	Tryptophan synthase (4.2.1.20): 'Chronic colitis is accompanied by a decrease in the serum Trp level.' + 'Trp supplementation ameliorated DSS-induced colitis through AhR' (citation 1), cystathionine beta-synthase (4.2.1.22): ''Decreased expression of CBS (cystathionine beta-synthase) propagates the pathogenesis of UC by exacerbating inflammation-induced intestinal barrier injury.' (citation 2), pseudouridylate synthase (4.2.1.70) : 'Heterogeneity in the phenotype of Crohn's disease, including development of perianal disease, has fostered the study of genetic predispositions. The gene PUS10 (coding for pseudouridylate synthase 10) has a substantial protective effect against the development of perianal disease'	Though the annotation itself proved too generic to be correlated to the disease itself, several of its children (see citations) were found to be relevant to IBD (category 2).	Human (signaling pathways)	0
GO:0006741	NADP biosynthetic process	1	1	https://doi.org/10.1002/1873-3468.14528	'In dextran sulphate sodium salt (DSS)-induced colitis, activated NADK produces NADP from NAD to mount an oxidative burst and increased infiltration of neutrophils, resulting in increased inflammation and immune dysregulation'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (biosynthetic pathways)	1
4.1.1.5	acetolactate decarboxylase.	1	3	PMID: 32509162 (1), https://doi.org/10.1002/mnfr.202300337 (2)	'Lactobacillus reuteri is a normal resident species of the healthy gut microflora that can prevent IBD by altering the intestinal micro-environment and the immune system' (ref 1), 'This study identifies the coding gene (aldB) of acetolactate decarboxylase (ALDC) as an important regulatory gene of the intracellular pH in Lactobacillus reuteri (L. reuteri)' (ref 2)	Though it does not appear to be directly linked to the disease in metabolic terms, this annotation is expressed by a taxon known to be beneficial, and used as a probiotic tratment for IBD, Lactobacillus reuteri (category 3)	Microbiota	0
GO:0032265	XMP salvage	1	2	https://doi.org/10.1016/j.isci.2020.101226	GO:0032261 - purine nucleotide salvage is a parent class of this GO term (https://www.ebi.ac.uk/QuickGO/term/GO:0032265); 'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis.MSPs support energy balance, proliferation, and mucous barrier function'	A direct bibliographic link between the annotation's prarent and IBD was established (category 2)	Human (epithelial cells)	0

GO:0045151	acetoin biosynthetic process	1	3	https://doi.org/10.1111/j.1365-2036.2011.04799.x	"This supports our previous hypothesis that the mechanism causing diarrhoea involves cell signalling, analogous to the diarrhoea in severe gut infections, and is not simply an osmotic effect of the lactose, as has been previously thought. Important metabolic toxins are methylglyoxal, acetoin, diacetyl, butan 2,3 diol and related aldehydes and ketones."	Link is not quite established, but this function is implicated in aspects of lactose sensitivity that are analogous to IBD. Could be a simple correlation, as "Sensitivity to lactose has been reported in Crohn's disease, but its true role in inflammatory bowel disease (IBD) is unclear." (category 3)	Human (signalling pathways)	1
3.5.3.6	arginine deiminase.	1	1	https://doi.org/10.1016/j.intimp.2020.106583	"The results showed enhanced expression of [...] PAD4 (Protein Arginine Deiminase-4) in TNBS-induced colitis mice"	A direct bibliographic link between the annotation and IBD was established (category 1) (the expression of this protein increases the production of Neutrophin Extracellular Traps, which induce inflammation)	Human (neutrophils)	1
GO:0009346	ATP-independent citrate lyase complex	1	2	https://doi.org/10.1016/j.mucimm.2022.11.001	"An increase in glucose uptake through GLUT3 supported the generation of acetyl-CoA and increased levels of citrate—pharmacological and genetic suppression of ATP-citrate lyase-dependent (ACLY) acetyl-CoA generation prevented histone acetylation at inflammatory gene loci and reduced cytokine responses"	We astablished that the expression of a similar protein, the ATP dependent citrate lyase, was correlated to inflammation (category 2)	Human (signalling pathways)	1
4.1.3.6	citrate (pro-3S)-lyase.	1	1	doi:10.3748/wjg.v13.i20.2826	This enzyme catalyzes the prouction of acetate (https://enzyme.expasy.org/EC/4.1.3.6). 'Our findings suggest that propionate and acetate, in addition to butyrate, could be useful in the treatment of inflammatory disorders, including IBD.', 'it is also clearly demonstrated that acetate and propionate ameliorate an ongoing inflammatory response at the cellular level'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0
GO:0008760	UDP-N-acetylglucosamine 1-carboxyvinyltransferase activity	1	3	https://doi.org/10.1136/gut.2010.232918	This enzyme is involved in the biosynthesis of peptidoglycan (https://www.brenda-enzymes.org/enzyme.php?ecno=2.5.1.7) 'PGN (peptidoglycan) purified from Ls33 rescued mice from colitis in an IL-10-dependent manner and favoured the development of CD103+ DCs and CD4+ Foxp3 + regulatory T cells.'	This annotation is involved in a pathway, the product of which is known to be beneficial in IBD when expressed in certain taxa	Microbiota (membrane)	0
GO:0006144	purine nucleobase metabolic process	1	1	https://doi.org/10.1111/imcb.12167	"The purine metabolic pathway is involved in various inflammatory processes including IBD."	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (signalling pathways)	1

GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	1	1	https://doi.org/10.2527/jas.54010	"We conclude that decreases in the small intestinal apical activities of these examined hydrolases likely contribute to overgrowth of pathogenic bacterial populations in the distal small intestine and the colon, leading to the pathogenesis of IBD."	A direct bibliographic link between the annotation and IBD was established (category 1)	Human + Microbiota	1
GO:0009409	response to cold	1	1	https://doi.org/10.1016/S0140-6736(03)15024-6	"All findings point to refrigeration as a potential risk factor for Crohn's disease."	A direct bibliographic link between the annotation and IBD was established (category 1)	Human + microbiota (environmental variable)	1
GO:1901135	carbohydrate derivative metabolic process	1	1	https://doi.org/10.3390/ijms23031105 (1), https://doi.org/10.1097/MIB.0000000000000116 (2)	"Short chain fatty acids (SCFAs) are among the important class of gut microbiota bio-products, produced mainly from fermentation of non-digestible carbohydrates, including dietary fiber, that become available to the gut microbiota" (citation 1), "In sum, increased intake of fermentable dietary fiber, or SCFAs, protects against colonic inflammation and therefore seems to be clinically beneficial in the treatment of GI disorders, such as colitis." (citation 2)		Human (signalling pathways)	0
GO:0008814	citrate CoA-transferase activity	1	1	https://doi.org/10.3748/wjg.v13.i20.2826	This enzyme catalyzes the prouction of acetate (https://enzyme.expasy.org/EC/2.8.3.10). 'Our findings suggest that propionate and acetate, in addition to butyrate, could be useful in the treatment of inflammatory disorders, including IBD.', 'it is also clearly demonstrated that acetate and propionate ameliorate an ongoing inflammatory response at the cellular level'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0
GO:0047330	polyphosphate-glucose phosphotransferase activity	1	1	https://doi.org/10.1186/1750-1172-6-27	Product of catalyzed reaction: glucose-6-phosphate (https://www.ebi.ac.uk/QuickGO/term/GO:0047330). "Glucose-6-phosphatase deficiency (G6P deficiency) [is] responsible for tendency towards infections, relapsing aphtous gingivostomatitis, and inflammatory bowel disease."	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human	0
GO:0015444	P-type magnesium transporter activity	1	2	doi:10.3390/nu13124188	"Nutritional deficiencies are common in inflammatory bowel diseases (IBD). In patients, magnesium (Mg) deficiency is associated with disease severity, while in murine models, dietary Mg supplementation contributes to restoring mucosal function."	Though the transporters themselves are not directly referenced, a lack of Magnesium in the diet is correlated to IBD severity, creating a direct link to a relevant molecule (category 2)	Human	0

2.7.1.23	NAD(+) kinase.	1	1	https://doi.org/10.1002/1873-3468.14528	'In dextran sulphate sodium salt (DSS)-induced colitis, activated NADK (NAD kinase) produces NADP from NAD to mount an oxidative burst and increased infiltration of neutrophils, resulting in increased inflammation and immune dysregulation'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (biosynthetic pathways)	1
GO:0008808	cardiolipin synthase activity	0	3	https://doi.org/10.3389/fimmu.2022.1028953 (1), https://doi.org/10.1023/A:1026646816672 (2)	"NLRP3 inflammasome can also be activated by direct interaction with mitochondrial cardiolipin, in a mtROS-independent manner. Cardiolipin is a phospholipid located exclusively in the inner mitochondrial membrane (IMM). It can translocate to the outer mitochondrial membrane (OMM) by mtROS production, PAMPs (e.g., LPS), or pro-apoptotic stimuli, thus promoting mitophagy. This suggests that cardiolipin could have a role in IBD pathogenesis. However, there are still no studies linking cardiolipin to IBD." (ref 2) + anti-cardiolipin antibodies are known to be more prevalent in IBD patients and are a risk factor for thrombosis (ref 2)	No direct link is established in the litterature. However, correlations between cardiolipin antibodies and IBD are known, and the molecule itself is cited as a promising lead (cat 3)	Human (mitochondria)	1
GO:0097056	selenocysteinyl-tRNA(Sec) biosynthetic process	1	1	https://doi.org/10.2147/JIR.S288412	Goterm synonym: selenocysteine biosynthesis; "Selenocysteine and selenocystine significantly attenuated IBD-related symptoms, including preventing weight loss, decreasing disease activity index (DAI) scores, and increasing colon length."	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (colon)	0
GO:0008815	citrate (pro-3S)-lyase activity	1	1	https://doi.org/10.3748/wjg.v13.i20.2826	This enzyme catalyzes the prouction of acetate (https://enzyme.expasy.org/EC/4.1.3.6). 'Our findings suggest that propionate and acetate, in addition to butyrate, could be useful in the treatment of inflammatory disorders, including IBD.', 'it is also clearly demonstrated that acetate and propionate ameliorate an ongoing inflammatory response at the cellular level'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0
GO:0047395	glycerophosphodiester phosphodiesterase activity	inositol	1	https://doi.org/10.3748/wjg.v23.i28.5115	This enzyme catalyzes the prouction of myo-inositol: 'In mice, pβ-cateninS552 staining faithfully reported the effects of myo-inositol in reducing inflammation and intestinal stem cell activation.'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (signalling pathways)	0

GO:0033711	4-phosphoerythronate dehydrogenase activity	0	3	https://doi.org/10.1016/S0002-9270(02)05828-8	This enzyme is involved in the metabolism of Vitamin B6 (https://www.brenda-enzymes.org/enzyme.php?ecno=1.1.1.290), 'Median vitamin B6 levels were significantly lower in IBD patients (22.0 pmol/L, range 3.6–231.0) than in controls (31.1 pmol/L, 3.7–363.4; $p < 0.01$). '	The annotation itself hasn't been correlated to IBD. It is however involved in the metabolism of Vitamin B6, which is known to be deficient in IBD patients (category 3)	Human (biosynthetic pathways)	0
GO:0004792	thiosulfate sulfurtransferase activity	1	1	https://doi.org/10.1016/j.bbadis.2020.165716	'Expression of TST (thiosulfate sulfurtransferase) in colon mucosa is often markedly reduced in patients with ulcerative colitis and colon cancer when compared to normal mucosa, although the evidence is not completely consistent [72]. This decrease in TST activity corresponds with the development of colitis, and is followed by an elevation of TST activity in erythrocytes [[71], [72], [73]].'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (signalling pathways)	0
2.1.1.195	cobalt-precorrin-5B (C(1))-methyltransferase.	1	4	https://doi.org/10.3389/fimmu.2023.1139799	This enzyme is involved in the metabolism of vitamin B12 (https://www.brenda-enzymes.org/enzyme.php?ecno=2.1.1.195) 'Further studies are warranted to elucidate the possible association between vitamin B12 and risk of IBD.'	The most relevant aspect of this annotation here is its implication in the metabolism of vitamin B12, which is heavily researched in the context of IBD. However, recent studies on the subject are still not conclusive in linking vitamin B12 levels to the disease. Therefore, though this promising lead has to be noted, there is no established correlation between this annotation and IBD (cat 4)	N.A	N.A
GO:0006522	alanine metabolic process	1	1	https://doi.org/10.3390/ph14111190	'Using 1H-NMR spectroscopy, Balasubramanian et al. studied the metabolism of the colonic mucosa of CD and UC patients with active and quiescent disease, as well as control subjects. During active phase, significantly lower concentrations of amino acids (isoleucine, leucine, valine, alanine, glutamate, and glutamine), membrane components (choline, glycerophosphorylcholine (GPC), and myo-inositol), lactate, and succinate were observed compared to control subjects, whereas, in remission, their concentrations were similar.'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human (mucosa)	0

GO:0008899	homoserine O-succinyltransferase activity	1	3	https://doi.org/10.3390/nu9090920	This enzyme is involved in the metabolism of methionine (https://www.brenda-enzymes.org/enzyme.php?ecno=2.3.1.46):'rats fed with methionine-restricted diet were found to have higher transepithelial electrical resistance and claudin-3 protein expression, and decreased severity of epithelial injury in an ulcerative colitis model induced by DSS'	The annotation itself hasn't been correlated to IBD. It is however involved in the metabolism of methionine, the overabundance of which has been correlated to colitis in mice models (category 3)	Human	1
GO:0008276	protein methyltransferase activity	1	2	https://doi.org/10.3390/life11080817	'Rare missense variants of SETDB1 (SET Domain Bifurcated Histone Lysine Methyltransferase 1, related to annotation) have been identified in Inflammatory Bowel Disease (IBD) patients and are associated with its pathogenesis.'	A protein related to the annotation was found to be relevant to the pathogenesis of IBD (cat 2)	Human (epithelial cells)	1
1.1.1.88	hydroxymethylglutaryl-CoA reductase.	1	1	https://doi.org/10.1016/j.ejim.2020.02.017	'Hydroxymethylglutaryl-CoA reductase inhibitors (statins) are the most commonly prescribed drugs worldwide', 'Some evidence suggests that statins may have an impact on IBD activity'	A direct bibliographic link between the annotation and IBD was established (category 1)	Human	1

Table S2 – **Bibliographic research on a subset of the IBD dataset's Meta-50 functional annotations.**

ID	Names	Family	Bibliography category	Bibliography	Quotes	Bibliographic grade justification	Impacted organism	Family according to bibliography
GO:0007165	signal transduction	0	1	https://doi.org/10.3390/cancers14153821	'Here, we show that pattern recognition receptors not only recognize pathogens and initiate inflammatory signal transduction to induce immune responses, but also influence the composition of intestinal microorganisms, thus affecting the development of intestinal inflammation and cancer through various mechanisms.'	A direct link is established between the annotation and the disease (cat 1)	Human + Microbiota	1
4.2.1.10	3-dehydroquinate dehydratase.	1	3	https://doi.org/10.1002/mnfr.202000461	This annotation is involved in the biosynthesis of tryptophan (https://www.brenda-enzymes.org/enzyme.php?ecno=4.2.1.10) 'Chronic colitis is accompanied by a decrease in the serum Trp level.' + 'Trp supplementation ameliorated DSS-induced colitis through AhR'	This annotation is included in a biological pathway relevant to IBD (cat 3)	Human (signaling pathways)	0

GO:0030254	protein secretion by the type III secretion system	1	3	https://doi.org/10.1128/cmr.00013-07 (1), https://doi.org/10.3389/fcimb.2018.00336 (2)	'Infections with many Gram-negative pathogens, including Escherichia coli, Salmonella, Shigella, and Yersinia, rely on the injection of effectors via type III secretion systems (T3SSs)' (1) , 'Yersinia have been implicated in Crohn's Disease (CD, an inflammatory bowel disease)' (2)	This annotation does not directly link to IBD, however it is known to be expressed by bacteria such as the Yersinia strain, which is a known pathogen of IBD (category 3)	Microbiota	1
GO:0016226	iron-sulfur cluster assembly	1	3	https://doi.org/10.1016/j.jinorgbio.2017.02.005 (1), https://doi.org/10.1097/mog.0000000000000949 (2)	'The human pathogen Clostridium difficile infection (CDI) is one of the most important healthcare-associated infections. The Wood-Ljungdahl pathway, which is responsible for Acetyl-CoA biosynthesis, is essential for the survival of the pathogen and is absent in humans. The key proteins and enzymes involved in the pathway are attractive targets for the treatment of CDI. Corrinoid iron-sulfur protein (CoFeSP) is a key protein and acts as a methyl transformer in the Wood-Ljungdahl pathway.' (1) + '' CDI remains common in IBD with complications including flares in disease activity, recurrent CDI episodes, and prolonged hospital stays. .' (2)	This annotation does not directly link to IBD, however it is known to be expressed by Clostridia Dificile, which is a known pathogen of IBD (category 3)	Microbiota	1
4.2.1.119	enoyl-CoA hydratase 2.	1	2	https://doi.org/10.1155/2019/1426954	'Principal component analysis (PCA) grouped noninflamed samples separately from the inflamed samples suggesting a distinctive proteomic signature of the colon mucosa in acute UC. A total of 43 individual protein spots were identified corresponding to 33 individual proteins. These proteins included those involved in energy metabolism (triosephosphate isomerase, glycerol-3-phosphate dehydrogenase, alpha enolase and L-lactate dehydrogenase B-chain, isocitrate dehydrogenase, inorganic pyrophosphatase, and enoyl-CoA hydratase)'	A direct link is established between the annotation's wider class and the disease (cat 2)	Human (mucosa)	1
GO:0004077	biotin-[acetyl-CoA-carboxylase] ligase activity	0	1	https://doi.org/10.1038/ncomms7592	'Here we identify a virulence-regulating pathway in which the biotin protein ligase BirA signals to the global regulator Fur, which in turn activates LEE (locus of enterocyte effacement) genes to promote EHEC adherence in the low-biotin large intestine.', 'HEC induces much severer symptoms, producing diarrhoea complicated by haemorrhagic colitis'	A direct link is established between the annotation and the disease (cat 1)	Microbiota	0

1.3.98.1	dihydroorotate oxidase (fumarate).	1	1	https://doi.org/10.3748/wjg.v20.i1.163	This enzyme catalyzes succinate production (https://www.brenda-enzymes.org/enzyme.php?ecno=1.3.98.1): "The most significant differences in urine were found between the group of patients with active IBD and the group with IBD in remission, and between the group of patients with active IBD and healthy control subjects. In the first case a lower concentration of acetoacetate and a higher concentration of glycine were detected, while in the second case citrate, hippurate, trigonelline, taurine, succinate, 2-hydroxyisobutyrate (downregulated) and an unknown metabolite with 4-hydroxyphenyl group and signal at δ 6.85 ppm (up-regulated) were found to be the strongest biomarker candidates (Table 5)."	This enzyme is involved in the generation within bacteriae of citrate and succinate, both of which are negatively correlated to IBD (category 3)	Microbiota	0
4.2.1.70	pseudouridylate synthase.	1	2	https://doi.org/10.1038/nrgastro.2017.104	'Heterogeneity in the phenotype of Crohn's disease, including development of perianal disease, has fostered the study of genetic predispositions. The gene PUS10 (coding for pseudouridylate synthase 10) has a substantial protective effect against the development of perianal disease, and a C allele at the CDKAL1 rs6908425 variant and the absence of NOD2 variants have also been independently associated with perianal fistulas.'	A direct link has been established between a child of the annotation and IBD complications (cat 2)	Human (genetic)	0
GO:0070395	lipoteichoic acid biosynthetic process	1	2	https://doi.org/10.1073/pnas.0504084102	'Teichoic acids (TAs), and especially lipoteichoic acids (LTAs), are one of the main immunostimulatory components of pathogenic Gram-positive bacteria.' + 'Taken together, these results emphasize the importance of LTA composition in the proinflammatory or antiinflammatory properties of Lactobacillus cells, but also point to the potential for use of specific Lactobacillus cell wall mutants for treatment of IBD.'	The expression of this specific molecule in Lactobacillus plantarum has been proven to affect the host immunoresponse, making it capable of enhancing or downregulating inflammatory responses depending on their form (most notably, it has been shown to reduce inflammation when it is in a shape that inclused less d-Ala). (category 2)	Microbiota	N.A
2.4.1.109	dolichyl-phosphate-mannose–protein mannosyltransferase.	1	3	https://doi.org/10.2147/JIR.S327609	This annotation is involved in O-glycan biosynthesis (https://www.brenda-enzymes.org/enzyme.php?ecno=2.4.1.109): 'Several studies have shown that O-glycan is involved in the pathogenesis and development of IBD.'	This annotation is included in a biological pathway relevant to IBD (cat 3)	Human (mucosa)	1

GO:0046026	precorrin-4 C11-methyltransferase activity	1	4	https://doi.org/10.3389/fimmu.2023.1139799	This enzyme is involved in the metabolism of vitamin B12 (https://www.brenda-enzymes.org/enzyme.php?ecno=2.7.7.62) 'Further studies are warranted to elucidate the possible association between vitamin B12 and risk of IBD.'	The most relevant aspect of this annotation here is its implication in the metabolism of vitamin B12, which is heavily researched in the context of IBD. However, recent studies on the subject are still not conclusive in linking vitamin B12 levels to the disease. Therefore, though this promising lead has to be noted, there is no established correlation between this annotation and IBD (cat 4)	N.A	N.A
GO:0050560	aspartate-tRNA(Asn) ligase activity	1	3	https://doi.org/10.3390/nu9090920	This enzyme is involved in aspartate and asparagine metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=6.1.1.23): 'Aspartate, asparagine and proline also participate in immune responses [129] that may maintain intestinal health and protect against animal and human diseases.'	This annotation can be correlated to a biological pathway linkable to the gut's general health, but no direct link to IBD was found (cat 3)	Human (gut)	0
GO:0004139	deoxyribose-phosphate aldolase activity	0	3	https://doi.org/10.1186/s12866-023-02932-8 (1), https://doi.org/10.1016/j.isci.2020.101226 (2)	This enzyme is involved in purine metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=4.1.2.4): 'Hypoxanthine phosphoribosyltransferase (Hpt), Adenine phosphoribosyltransferase (apt), and xanthine phosphoribosyltransferase (xpt) are the enzymes crucial to the purine salvage pathway.' (1st citation) + 'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis. MSPs support energy balance, proliferation, and mucous barrier function' (2nd citation)	The annotation itself doesn't have a direct link to IBD in the bibliography. It is however involved in the purine pathway, which is a known factor in intestinal mucosa repair. The annotation is therefore involved in a larger relevant mechanism (category 3)	Human (epithelial cells)	0

GO:0046555	acetylxylan esterase activity	1	3	https://doi.org/10.3390/ph15091151	'Degradation of xylan into xylo oligosaccharides (Figure 1e) and into free xylose requires the combined action of degradative enzymes such as α-L-arabinofuranosidase (EC 3.2.1.55), α-D-glucuronidase (EC 3.2.1.139), acetylxylan esterase (EC 3.1.1.72) and ferulic acid esterases (EC 3.1.1.73), which release the side chains from the xylan backbone. Endo-β-1,4-xylanase (EC 3.2.1.8) acts synergistically with β-xylosidase (EC 3.2.1.37) to degrade the xylan backbone with the former hydrolysing the internal β-(1,4) linkages of the xylan backbone to produce short xylo-oligosaccharides, and β-xylosidase then removes xylose units from the non-reducing termini of these xylo-oligosaccharides (Figure 1e).', 'The use of engineered B. ovatus for focal delivery of KGF-2 and TGF-β has considerable potential in the treatment of inflammatory bowel disease. Bacteroides spp. are prominent commensal anaerobes found in the mucin layer coating the colonic mucosa and thus ideally placed for therapeutic protein delivery to the injured epithelium. The ability of B. ovatus to utilise xylan as its sole carbon source contributes to its predominance as a representative of the colon microbiota.'	The annotation itself doesn't have a direct link to IBD in the bibliography. It is however expressed by xylan-consuming bacteria, several of which are researched as probiotics for IBD (category 3)	Microbiota	0
4.1.2.4	deoxyribose-phosphate aldolase.	0	3	https://doi.org/10.1186/s12866-023-02932-8 (1), https://doi.org/10.1016/j.isci.2020.101226 (2)	This enzyme is involved in purine metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=4.1.2.4): 'Hypoxanthine phosphoribosyltransferase (Hpt), Adenine phosphoribosyltransferase (apt), and xanthine phosphoribosyltransferase (xpt) are the enzymes crucial to the purine salvage pathway.' (1st citation) + 'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis. MSPs support energy balance, proliferation, and mucous barrier function' (2nd citation)	The annotation itself doesn't have a direct link to IBD in the bibliography. It is however involved in the purine pathway, which is a known factor in intestinal mucosa repair. The annotation is therefore involved in a larger relevant mechanism (category 3)	Human (epithelial cells)	0

GO:0008124	4-alpha-hydroxytetrahydrobiopterin dehydratase activity	1	3	https://doi.org/10.3390/microorganisms4020020	This enzyme is involved in phenylalanine metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=4.2.1.96): 'In a human study published in 2014 [137], the metabolites that allowed to distinguish between the group of patients with active IBD and the group with IBD in remission were: N-acetylated compounds and phenylalanine which were up-regulated in serum'	This annotation is included in a biological pathway which was measured with higher prevalence in IBD (cat 3)	Human	1
GO:0015846	polyamine transport	1	1	https://doi.org/10.1080/10408360701250016	'In inflamed mucosal specimens of patients with inflammatory bowel disease, ODC activity and polyamine concentrations are increased.'	A direct link is established between the annotation and the disease (cat 1)	Human (mucosa)	1
GO:0046113	nucleobase catabolic process	1	2	https://doi.org/10.1111/imcb.12167 (1), https://doi.org/10.1016/j.biopha.2023.114620 (2)	A child term of this GO term is purine nucleobase catabolism (https://www.ebi.ac.uk/QuickGO/term/GO:0046113): "The purine metabolic pathway is involved in various inflammatory processes including IBD." (2)	A direct bibliographic link between a child of the annotation and IBD was established (category 2)	Human (signalling pathways)	1
GO:0004365	glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity	1	3	https://doi.org/10.1016/j.mucimm.2022.11.001 (1) https://doi.org/10.1016/j.biopha.2023.114620 (2)	This enzyme is involved in the metabolism of short chain fatty acids (SFCAs) (1), 'SCFAs have various functions in order to improve the inflammatory conditions of the digestive system in IBD and celiac diseases' (2)	This annotation is included in a biological pathway relevant to IBD (cat 3)	Human (signalling pathways)	0
GO:0015606	spermidine transmembrane transporter activity	1	1	https://doi.org/10.1080/10408360701250016	'Putrescine, spermidine, and spermine are representatives of a group of aliphatic biogenic amines that is known by the designation "polyamines.", "In inflamed mucosal specimens of patients with inflammatory bowel disease, ODC activity and polyamine concentrations are increased.'	A direct link is established between the annotation and the disease (cat 1)	Human (mucosa)	1

Table S3 – **Bibliographic research on a subset of the IBD dataset's non-candidate functional annotations.**

ID	Name	Family	Bibliography category	Bibliography	Quotes	Bibliographic grade justification	Impacted organism	Family according to bibliography

GO:0051266	sirohydrochlorin ferrochelatase activity	1	3	https://doi.org/10.3389/fmicb.2017.01809	This enzyme is involved in heme metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=4.99.1.4) 'In conclusion, our data suggest that luminal heme, originating from dietary components or gastrointestinal bleeding in IBD and, to lesser extent in CRC, directly contributes to microbiota dysbiosis.'	There is no direct link between this annotation and IBD, however it is part of the heme metabolism pathway, which is a factor of disbyosis in the context of IBD (cat 3)	Microbiota	1
2.3.1.179	beta-ketoacyl-[acyl-carrier-protein] synthase II.	1	3	https://doi.org/10.1097/MIB.0000000000000394	This annotation is involved in lipid metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=2.3.1.179) 'We demonstrate that a number of ether lipids (alkylphospholipid and plasmalogens) are significantly and negatively associated with CD. These alterations of lipid profiles particularly plasmalogens may contribute to the pathogenesis of IBD.'	A direct bibliographic link between the annotation's product and IBD was established (category 1)	Human (plasma)	0
1.4.4.2	glycine dehydrogenase (aminomethyl-transferring).	0	3	https://doi.org/10.1016/j.compbiomed.2022.105865 (1), https://doi.org/10.2174/1872131079116307 1 (2)	'S. typhi amino methyl-transferring glycine dehydrogenase protein is similar to the human Glycine dehydrogenase protein, implicated in the Toll-Like Receptor Pathway and TLR signaling.' (ref 1), 'Studies have revealed that intestinal microorganisms play a key role in the initiation and maintenance of disease, and some signaling pathways including TLR, NF-__x0002__B can act on the intestinal barrier, and may be associated with the intes- tinal environment disorder, and affect the disease [i.e: IBD]' (ref 2)	This protein is involved in the microbiota's Toll-Like Receptor (TLR) signalling pathway, which is a known factor of IBD (cat 3)	Microbiota (signalling pathways)	1

GO:0009061	anaerobic respiration	1	1	https://doi.org/10.1111/j.1365-2036.2008.03612.x (1), https://doi.org/10.3389/fimmu.2019.00277 (2),	' Anaerobic nitrate respiration yields nitrite, nitric oxide (NO) and nitrous oxide. Colonic bacteria produce NO and UC (Ulcerative Colitis) in remission has a higher lumenal NO level than control cases. [...] The prolonged production of bacterial NO with sulphide can explain the initiation and barrier breakdown, which is central to the pathogenesis of UC.' (ref 1) + 'SCFAs are carboxylic acids with aliphatic tails of 1–6 carbons of which acetate (C2), propionate (C3), and butyrate (C4) are the most abundant produced by anaerobic fermentation of dietary fibers (DF) in the intestine.' (ref 2) + ' The metabolic welfare in health primarily depends on n-butyrate, a SCFA produced by fermentation of complex carbohydrates. In early UC, there is failure to utilize n-butyrate for oxidative and synthetic processes (mucus, lipids and proteins) with parallel enhancement of glucose oxidation.' (ref 1)	Anaerobiotic mechanisms are directly relevant to IBD (category 1). However, it is difficult to link anaerobiosis itself to the disease, as it appears to be the balance of different anaerobiotic pathways that determine the risk, as some of these pathways are beneficial (i.e: SFCA via fermentation) and others detrimental (I.e: nitrate respiration).	Microbiota	N.A
GO:0018759	methenyltetrahydromethanopterin cyclohydrolase activity	3	3	https://doi.org/10.1007/s00253-018-8871-2	This annotation is involved in methanogenesis (https://www.brenda-enzymes.org/enzyme.php?ecno=3.5.4.27) 'Initial studies found that there is a correlation between methane excretion and IBD. In UC and CD patients, a mere 10 to 30% were methane producers, compared to 50% in control groups.'	Though no direct link was found between this enzyme or its products and IBD, it participates in methanogenesis, which is less prevalent in IBD microbiotas (cat 3)	Microbiota	0
GO:0004042	acetyl-CoA:L-glutamate N-acetyltransferase activity	1	1	https://doi.org/10.1038/s41467-023-42788-0	Goterm synonym: N-acetylglutamate synthase (https://www.ebi.ac.uk/QuickGO/term/GO:0004042) 'The K22477 (argO, N-acetylglutamate synthase) is responsible for producing N-acetylglutamate (NAG) from glutamate and acetyl-CoA. Our study revealed that IBD patients have reduced levels of K22477, leading to an excess of L-glutamate.'	A direct link was established between the annotation and IBD (cat 1)	Human	0
GO:1904823	purine nucleobase transmembrane transport	1	1	https://doi.org/10.1016/j.isci.2020.101226	'Microbiota-sourced purines (MSPs) are salvaged by the gut mucosa for nucleotide genesis. MSPs support energy balance, proliferation, and mucous barrier function'	A direct link was established between the annotation and IBD (cat 1)	Human (epithelial cells)	0

GO:0004018	N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity	1	1	https://doi.org/10.1080/00365520510023198	Fumarate is a product of the reaction catalyzed by this enzyme (https://www.ebi.ac.uk/QuickGO/term/GO:0004018) 'Oral ferrous fumarate [...] increased clinical disease activity in IBD patients. '	A direct link was established between the annotation's product and IBD (cat 1)	Microbiota	1
GO:0004309	exopolyphosphatase activity	0	3	https://doi.org/10.1111/imcb.12167	This enzyme is involved in purine metabolism (https://www.brenda-enzymes.org/enzyme.php?ecno=3.6.1.11) "The purine metabolic pathway is involved in various inflammatory processes including IBD."	Though no direct link was found between this enzyme or its products and IBD, it is involved in purine metabolism, which is involved in IBD (cat 3)	Human (signalling pathways)	1
GO:0098800	inner mitochondrial membrane protein complex	0	3	https://doi.org/10.3390/microorganisms10101910	This annotation is characteristic of the presence of eukaryotes in the microbiota, as bacteria and archaea do not possess mitochondiae. 'The individuals with IBD had a higher prevalence of fungi, particularly Saccharomyces cerevisiae, and a lower prevalence of protozoa, particularly Blastocystis species (subtypes 1, 2, 3, and 4).' + Fig 1: higher proportion of eukaryotes in control samples compared to IBD	Though no direct link was found between this component and IBD, it is expressed exclusively by eukaryotes, which tend to be more prevalent in IBD (cat 3)	Microbiota	1
GO:0008655	pyrimidine-containing compound salvage	0	4	NA	The purine salvage pathway is connected to IBD, but no solid references were found for the role of the pyrimidine salvage pathway	No link was found (cat 4)	NA	NA
GO:0061595	6-deoxy-6-sulfofructose-1-phosphate aldolase activity	1	3	https://doi.org/10.1039/C8NP00074C	Enzyme products: 3-sulfolactaldehyde + dihydroxyacetone phosphate (https://www.ebi.ac.uk/QuickGO/term/GO:0061595). 'Interestingly, the genomes of ∼20% of the human population encode for a non-functional fucosyltransferase (Fut2) that normally adds terminal L-fucose molecules to glycans,84 and this genotype has been associated with decreased microbiota diversity and a higher risk for Crohn's disease.' + 'The L-fucose sugars released from polysaccharides can have multiple fates depending on the degrading organism and the environmental conditions. One of the pathways, termed the "fucose utilization" (fuc) pathway, starts with steps similar to those of glycolysis, involving aldol cleavage to yield lactaldehyde and dihydroxyacetone phosphate.'	Though no direct link was found between this enzyme and IBD, its products appear to be involved in the metabolism of L-Fucose, which can be correlated to Crohn's (cat 3)	Human (metabolism)	0

GO:0097157	pre-mRNA intronic binding	0	3	https://doi.org/10.1016/j.ejcb.2010.11.010 (1), https://doi.org/10.1002/ctm2.1479 (2)	'Alternative pre-mRNA splicing is regarded as a pivotal mechanism for generating proteome diversity and complexity from a limited inventory of mammalian genes.'(1), 'Available evidence suggests that some abnormal splicing RNAs can lead to multiple intestinal disorders during the onset of IBD' (2)	This enzyme is involved in m-RNA splicing, which can be involved IBD (cat 3)	Human (genetic)	N.A
GO:0071597	cellular birth scar	0	4	NA	NA	No link was found (cat 4)	NA	NA
GO:0071001	U4/U6 snRNP	0	3	https://doi.org/10.1002/ctm2.1479	'Compared with the constitutive splice, AS (Alternative Splicing) is far more complex. The process is performed by the spliceosome, which is a big complex consisting of 5 ribonucleoproteins (RNPs) involving the small nuclear RNA U1, U2, U4, U5, U6 and multiple auxiliary proteins cooperating to precisely recognise the splicing sites and catalyse the two splicing reaction steps.18, 19 First of all, the splicing process starts with the identification of the 5' splicing site by the snRNP U1 and the combination of the splicing factor 1 (SF1) with the branch point 3 and of the U2 auxiliary factor (U2AF) heterodimer with the 3' terminal AG and polypyrimidine tract. This assembly contributes to the E complex formation, which can be transformed to an ATP-reliant, pre-spliceosome A complex after replacing SF1 with the U2 snRNP at the branch site. Subsequently, the recruitment of U4/U6–U5 tri-snRNP complex causes the B complex formation.' + "Available evidence suggests that some abnormal splicing RNAs can lead to multiple intestinal disorders during the onset of IBD'	This enzyme is involved in m-RNA splicing, which can be involved IBD (cat 3)	Human (genetic)	N.A
GO:0008252	nucleotidase activity	0	2	https://doi.org/10.3389/fphar.2020.619458	'Ectonucleotidases are extracellular enzymes with a pivotal role in inflammation that hydrolyse extracellular purine and pyrimidine nucleotides, e.g., ATP, UTP, ADP, UDP, AMP and NAD+.'	A type of nucleotidase was found to be relevant to IBD (cat 2)	Human (signalling pathways)	1

4.1.1.65	phosphatidylserine decarboxylase	0	4	https://doi.org/10.3390/ijms222111682	Enzyme product: Phosphatidylethanolamine (https://www.brenda-enzymes.org/enzyme.php?ecno=4.1.1.65): 'However, thus far there is no in vivo evidence on the positive role of PE (Phosphatidylethanolamine) in IBD, either in mouse models or in clinical data.'	There is currently no evidence to link this enzyme to IBD (cat 4)	NA	NA
GO:0032447	protein urmylation	0	2	https://doi.org/10.1016/j.imbio.2020.152026	GO term definition: 'Covalent attachment of the ubiquitin-like protein URM1 to another protein. '; 'In the past few years, accumulative evidence illustrated that six E3 ubiquitin ligases, namely, ring finger protein (RNF) 183, RNF 20, A20, Pellino 3, TRIM62 and Itch, exhibited clear mechanisms in the development of IBD. They regulate the intestinal inflammation by facilitating the ubiquitination of targeted proteins which participate in different inflammatory signaling pathways.'	A direct bibliographic link between a comparable process and IBD was established (category 2)	Human (signalling pathways)	0
GO:0060702	negative regulation of endoribonuclease activity	1	1	https://doi.org/10.1073/pnas.1809575115	'We show that specific deletion of the endoribonuclease Regnase-1 in intestinal epithelial cells relieves the symptoms of experimental colitis during acute inflammation.'	A direct link was established between the annotation and IBD (cat 1)	Human (epithelium)	0
GO:0031126	sno(s)RNA 3'-end processing	0	2	https://doi.org/10.1093/ibd/izaa009	'Noncoding RNAs can be divided according to their function into 2 groups: housekeeping ncRNAs (e.g., tRNAs, rRNAs, snRNAs, snoRNAs)...', 'Specific deregulation patterns of ncRNAs have been linked to pathogenesis of various diseases, including pediatric IBD.'	This annotation is attached to a larger family of molecules that has an impact on IBD (category 2)	Microbiota	N.A

Appendix C

This appendix contains detailed zooms on the different parts of Figure 5.1, in Section 5.1. They are made available to improve the readability of the figure.

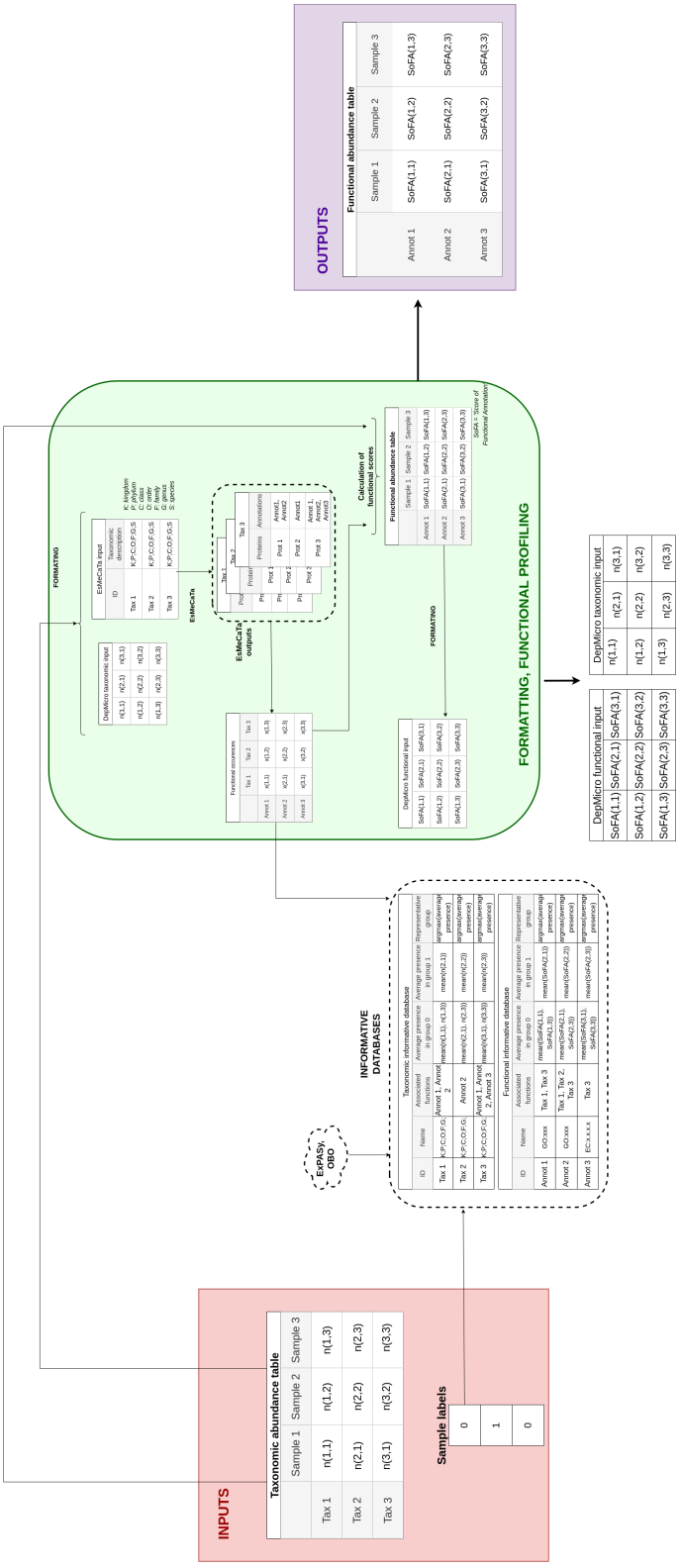


Figure S5 – **Visual representation of the implementation of SPARTA's first step (formatting and functional profiling).** This is a zoom on a part of Figure 5.1.

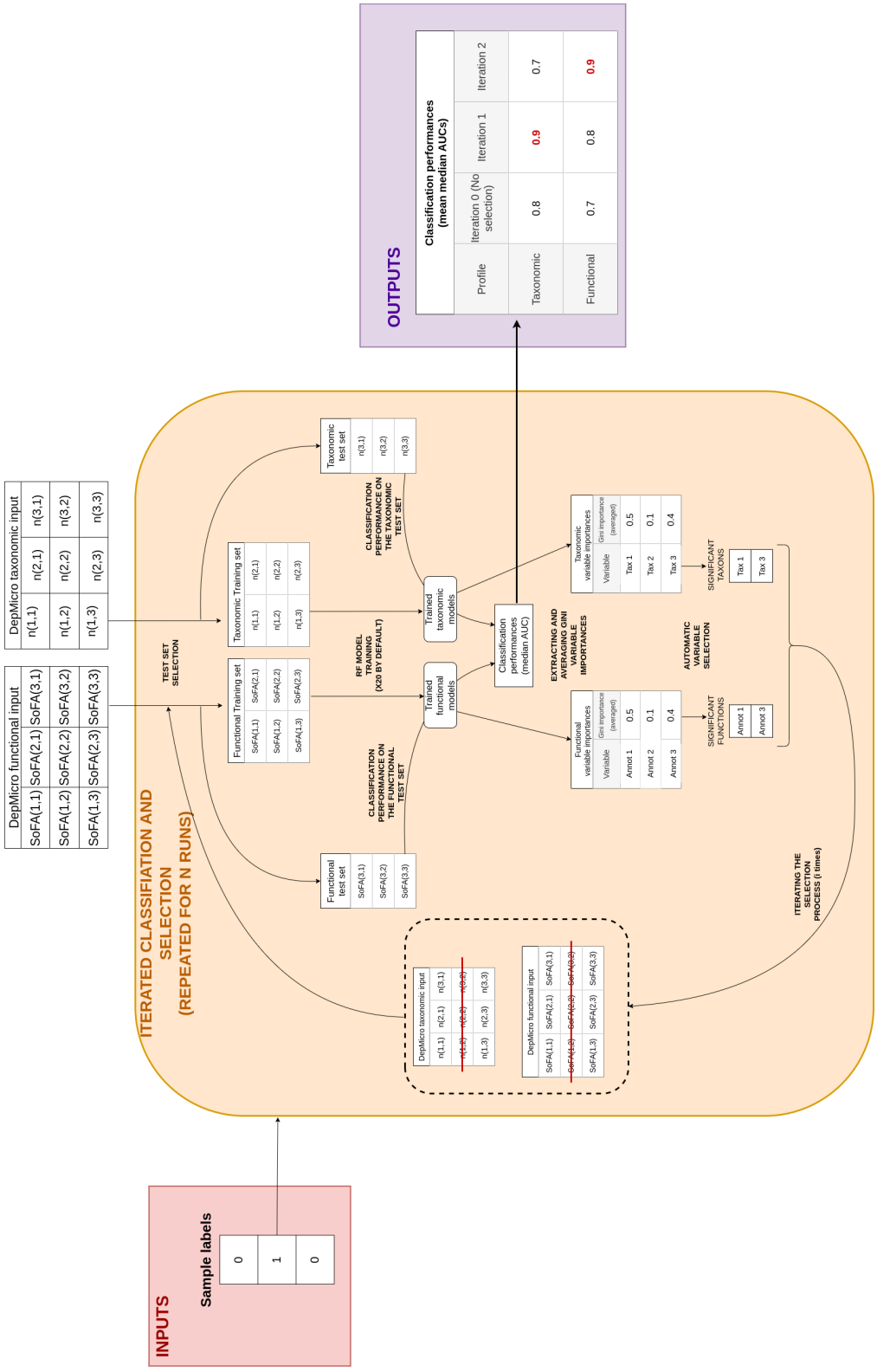


Figure S6 – **Visual representation of the implementation of SPARTA's second step (iterative classification and selection).** This is a zoom on a part of Figure 5.1.

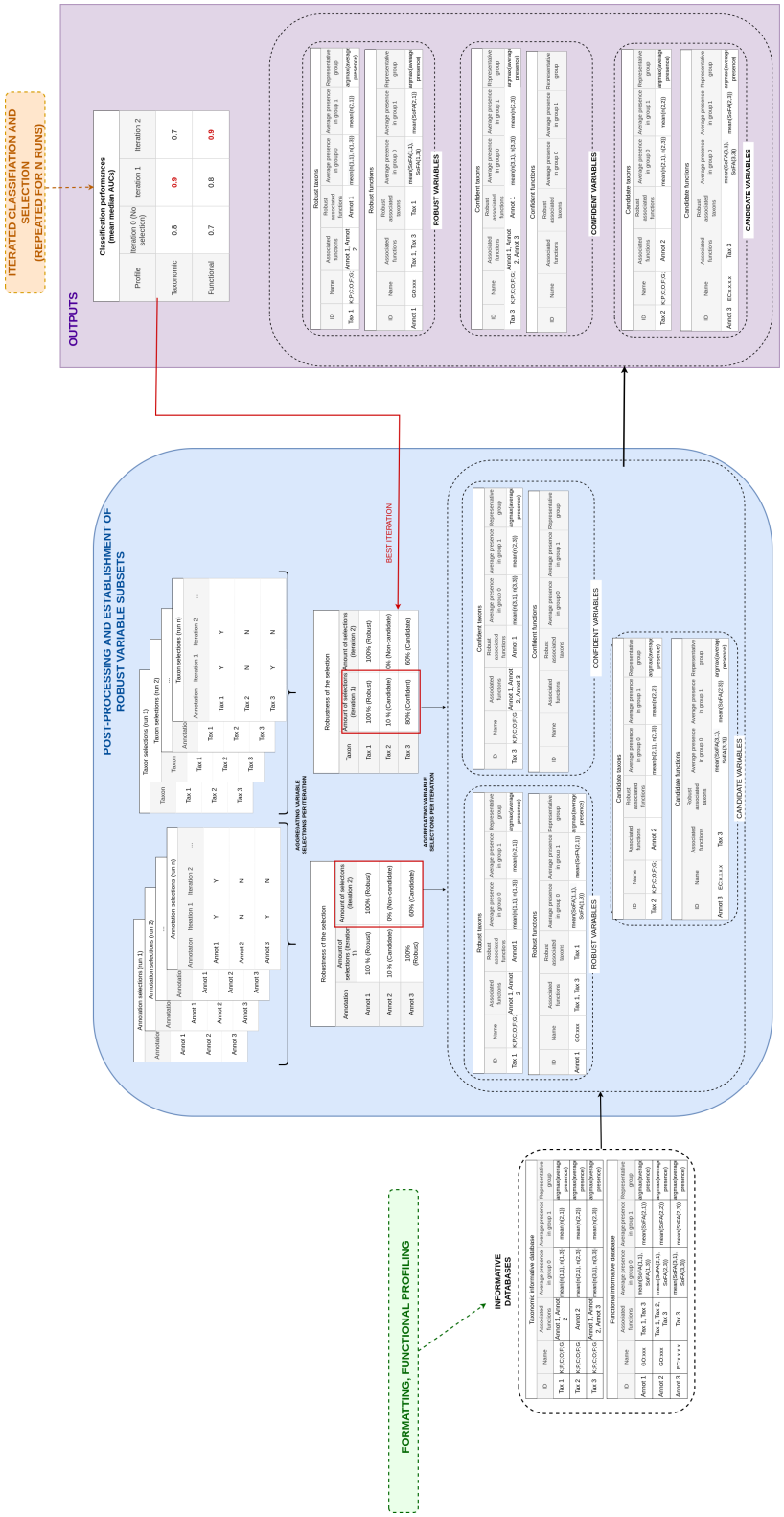


Figure S7 – **Visual representation of the implementation of SPARTA's third step (post-processing and establishment of robust variable subsets).** This is a zoom on a part of Figure 5.1.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Algorithmes d'apprentissage automatique dans le secteur de la santé : intégration des connaissances fonctionnelles pour améliorer l'analyse des données sur le microbiote intestinal

Mot clés : Microbiote intestinal, métagénomique, intégration de connaissances, Apprentissage Automatique

Résumé : La composition du microbiote intestinal influence diverses maladies et peut être utilisée pour la classification automatique de l'état de santé. Cette thèse propose une méthode intégrant l'annotation fonctionnelle du microbiote intestinal dans un processus de classification automatique pour améliorer l'interprétation des résultats. En utilisant les données taxonomiques et les annotations fonctionnelles via le pipeline EsMeCaTa, un profil fonctionnel du microbiote est établi. Ces profils, microbien et fonctionnel, servent à entraîner des Forêts Aléatoires pour différencier les échantillons malades des témoins. Une sélection automatique des variables basée sur leur importance est itérée jusqu'à la diminution des performances de classification. Les résultats montrent que les profils fonctionnels offrent des performances comparables aux profils microbiens et permettent d'identifier un sous-ensemble robuste de variables discriminantes. Ces variables se sont révélées plus fiables que celles obtenues par des méthodes de référence et ont été validées par une recherche bibliographique. L'analyse des interconnexions entre taxons et annotations fonctionnelles a révélé que cartaines annotations importantes sont issues de l'influence cumulative de taxons non sélectionnés.

Title: Machine Learning Algorithms in the health sector : integration of functional knowledge to enhance the analysis of gut microbiota data

Keywords: Gut microbiota, metagenomics, knowledge integration, Machine Learning

Abstract: The gut microbiota composition is a recognized factor in various diseases and serves as a robust basis for automatic disease state classification. A deeper functional understanding of this community is needed to enhance the biological interpretability of these approaches. This thesis presents a method for integrating functional annotation of the gut microbiota into an automatic classification process, facilitating downstream result interpretation. The process utilizes taxonomic composition data and links each component to its functional annotations via the EsMeCaTa pipeline, creating a functional profile of the gut microbiota. Both microbial and functional profiles are used to train Random Forest classifiers to distinguish between unhealthy and control samples. An automatic variable selection, based on variable importance, is iterated until classification performance declines. The results demonstrate that functional profiles provide comparable performance to microbial profiles and yield a robust subset of discriminant variables through repetition. These selections proved more reliable than those from state-of-the-art methods and were validated through manual literature review. Analysis of the interconnections between selected taxa and functional annotations revealed that significant annotations arise from the cumulative influence of non-selected taxa.