Title
Computational Methods for Optimization of Biological Organisms

Author
Eetemadi, Ameen

Publication Date
2021

Peer reviewed|Thesis/dissertation

Computational Methods for Optimization of Biological Organisms

By

AMEEN EETEMADI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Ilias Tagkopoulos, Chair

_____

Jonathan Eisen

_____

Justin Siegel

Committee in Charge

2021

Contents

Computational Methods for Optimization of Biological Organisms

Abstract

Computational methods play an irreplaceable role for optimization of biological organisms in the era of high-resolution omics, genetic engineering, and high-performance computing. A general overview of computational methods for optimization of biological organisms is presented in **Chapter 1** with a focus on three main challenges relating to data scarcity and heterogeneity, model interpretability, and the large number of factors that can affect an organisms' phenotype. Recent advances are discussed in **Chapter 2** with a forward-looking view on the application of computational methods for microbiome-based diet and health optimization. In **Chapter 3**, existing computational methods are applied for microbiome-based diet optimization in irritable bowel syndrome (IBS). The integrated data analysis results argue that there are two types of patients distinguishable by their fecal samples, those with high colonic methane and SCFA production, who will respond well on a low-FODMAP diet, and all others, who would benefit a dietary supplementation containing butyrate and propionate, as well as probiotics with SCFA-producing bacteria, such as lactobacillus. In **Chapter 4**, a novel artificial neural network (ANN) architecture called genetic neural network (GNN) is presented that captures the dependencies and non-linear dynamics that exist in gene networks into the GNN architecture. The results argue for 40% more accuracy of GNNs compared to several common ANNs in predicting genome-wide gene expression given gene knockouts and master regulator perturbations in bacterium *E. coli*. In **Chapter 5**, a novel group testing method called algorithmic lifestyle optimization (ALO) is presented for rapid identification of effective lifestyle interventions in individuals. ALO is robust to noise, data size and data heterogeneity, is between 58.9% and 68.4% more efficient compared to standard elimination diet for identification of food items that exacerbate IBS symptoms and

allergic reactions, and better than alternative state of the art group testing method for this application. The conclusions and future directions are discussed at the end of each chapter and summarized in the final chapter. Chapters 2, 3 and 4 are published (1–3).

Acknowledgements

I would like to thank my advisor, Professor Ilias Tagkopoulos, for his continued support and guidance throughout my Ph.D. that led to substantial improvements in my research and communication skills as well as the ability to realize my shortcomings and appreciate my strengths. Frankly, I initially felt that some of Ilias's expectations were too high for me to achieve, and now I can set myself goals that are above and beyond the horizon, and still find ways to achieve them. I'm grateful to be trained by such a smart, supportive, and visionary scientist, and proud for graduating from his lab.

Thanks to Professor Jonathan Eisen and Professor Justin Siegel for reviewing this manuscript as my dissertation committee members. I'm glad to have met such great leaders in science and receive their input on my research. Also, I'd like to thank my peer reviewed publication coauthors Dr. Minseung Kim, Dr. Beatriz Pereira, Dr. Navneet Rai, Dr. Harold Schmitz, Dr. Xiaokang Wang, and of course my advisor Dr. Ilias Tagkopoulos, as well as the editors and reviewers in the Bioinformatics, Clinical Nutrition, Frontiers in Microbiology, and PLOS Computational Biology journals for their constructive comments. I have benefited from discussions and help of my colleagues and friends in the Tagkopoulos lab especially Navneet, Xiaokang, Beatriz, Minseung, Jason, Linh, Cheng-En, Trevor, Gabe, Erol, Fang, Adil, and Tarini. I would like to thank them all.

I dedicate this dissertation to my dear wife, Zeinab, for her love, patience, support, and sacrifices throughout my academic journey. Ph.D. is one of the hard journeys in life that is left unfinished by many, it feels so good to be here at the finishing line, and be with the ones I love, Thank God. I clearly remember Zeinab's encouragements and positive energy on the first day that we came to

Davis, which has continued to this day, I cannot thank her enough. My two daughters Fatimah and Fereshteh were so good throughout and always made me happy, thanks to both of them.

I am especially grateful to my mom and dad, for their support, encouragement, love, and sacrifices from the time I was in the womb, until this day. They nourished my physical and spiritual development, and instilled great values that carry me through life. I'm really thankful for the help and advice of my older brother Sauleh at key points during my career in industry and academia.

Last but not least, I'd also like to thank Professor Farshad Fotouhi, Professor Hamid Soltanian-Zadeh, and Professor Mohammad-Reza Siadat for their advice and mentorship during and after my masters' studies in Wayne State University, and for their recommendations in my Ph.D. admission.

Chapter 1: Introduction

Biological organisms are naturally optimized through evolution by acquiring the characteristics that maximize their fitness when subjected to environmental constrains. Although evolution is a major method used by nature for optimization of biological organisms, it has several limitations. *First*, it does not optimize an individual organism during its lifetime. *Second*, it is relatively slow as it requires real experimentations that often take the lifetime of several generations until an environmentally fit generation of the organism is achieved. *Third*, it is mainly used for selection of organism characteristics given the environment, and not vice versa. Humans have attempted to address these limitations throughout history in medicine and agriculture using various methodologies based on intuition, logic, discovery of common principals, animal/plant breeding and more (4–6). Our civilization across the globe, is now intertwined with such methodologies that form the basis of our daily lives, from agricultural products to nutrition and medical practices.

With the advent of computers, computational methods have emerged as new powerful tools for optimizing biological organisms and play a key role in tackling outstanding health and environmental challenges, some of which are self-inflicted (7–10). These computational methods are reviewed with a future-looking perspective in **Chapter 2**, in the context of human microbiome, diet and health. Computational optimization of biological organisms often involves two major steps (**Figure 1.1 A**). The first step is to model the biological organism as a system with measurable inputs and outputs, as well as the target phenotype[s]. Such a model may correspond to a single organism, or a population of organisms. The second step is to optimize the modelled system for identifying a set of perturbations that will move it towards a state that corresponds to desired phenotypes.

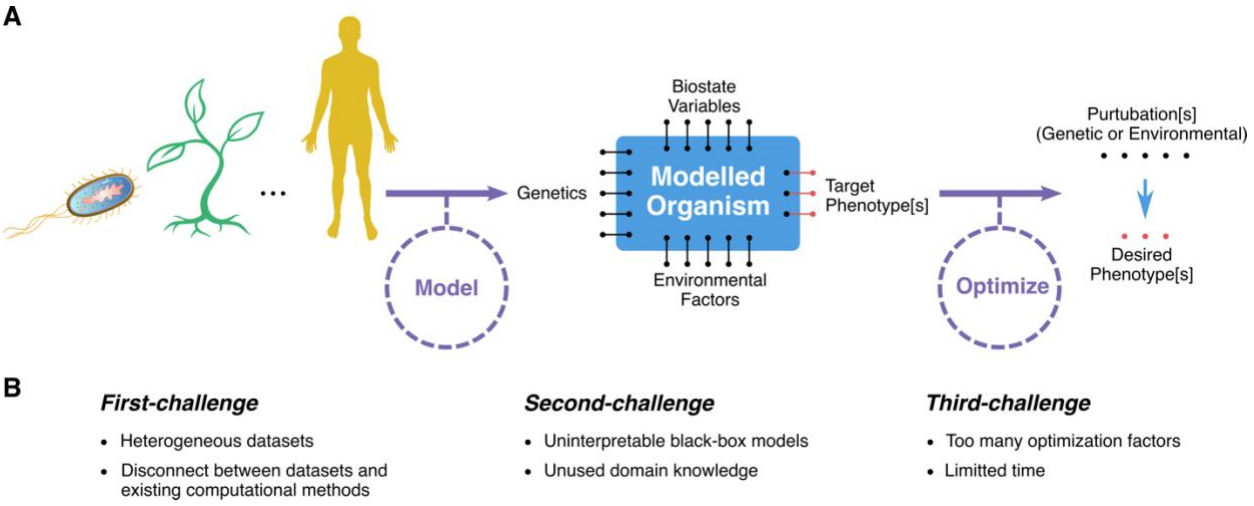

Figure 1.1 Computational modeling and optimization of biological organisms, with relevant challenges to address.[1]

Notable progress is made in development of computational methods for optimizing biological organisms with applications in nutrition, medicine, agriculture and biotechnology (1), while several challenges remain to be addressed sufficiently (**Figure 1.1 B**).

First-challenge. The disconnect between heterogenous datasets that are generated for optimizing the same condition in a given organism, and the appropriate computational methods, is an impediment towards realizing their full potential especially when the number of individual datasets and the size of each is small. This challenge is addressed in **Chapter 3**, for identification of irritable bowel syndrome patients that benefit from a low-FODMAP diet.

Second-challenge. Machine learning models can be used to predict the behavior of an organism given genetic and environmental perturbations; however, they are often incapable of incorporating the domain knowledge to enhance their accuracy and interpretability. Such models, are important for optimizing a given characteristic of an organism (e.g., minimizing adverse gastrointestinal

[1] The *Escherichia coli* icon is from Database Center for Life Science (DBCLS), CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, sourced via Wikimedia Commons. The human icon is made by Mikael Häggström, sourced via Wikimedia Commons.

symptoms in humans or maximizing the synthesis of target proteins in bacterium *Escherichia coli*), but their utility is substantially diminished if their accuracy is low, cannot incorporate domain knowledge, and are not interpretable. This challenge is addressed in **Chapter 4**, for prediction of gene expression in bacterium *Escherichia coli* using a novel machine learning method that incorporates gene regulatory relationships into its architecture leading into a more accurate and interpretable model compared to other machine learning methods.

Third-challenge. Optimal behavior of an organism depends on a large number of factors that can be hard or impossible to examine one-by-one due to time limitations. In cases where different factors have independent effects on a given binary target behavior of the organism (e.g., healthy/unhealthy state), they can be examined simultaneously in groups to minimize the time spent for identifying the effect of each factor. This challenge is addressed in **Chapter 5,** for rapid identification of effective lifestyle interventions amongst a large number of candidate lifestyle interventions, using a novel group testing method called algorithmic lifestyle optimization, and showcased for identifying food intolerances in irritable bowel syndrome and food allergies.

The following dissertation chapters that address the above challenges, have been published independently (1–3), or under consideration for peer review.

Chapter 2: Microbiome and diet-aware computational methods for health optimization

2.1 Abstract

Food and human health are inextricably linked. As such, revolutionary impacts on health have been derived from advances in production and distribution of food relating to food safety and fortification with micronutrients. During the past two decades, it has become apparent that the human microbiome has the potential to modulate health, including in ways that may be related to diet and the composition of specific foods. Despite the excitement and potential surrounding this area, the complexity of the gut microbiome, the chemical composition of food, and their interplay *in situ* remains a daunting task to be fully understood. However, recent advances in high-throughput sequencing, metabolomics profiling, compositional analysis of food, and the emergence of electronic health records, provide new sources of data that can contribute to addressing this challenge. Computational science will play an essential role in this effort as it will provide the foundation to integrate these data layers and derive insights capable of revealing and understanding the complex interactions between diet, gut microbiome, and health. Here, we review the current knowledge on diet-health-gut microbiota, relevant data sources, bioinformatics tools, machine learning capabilities as well as the intellectual property and legislative regulatory landscape. We provide guidance on employing machine learning and data analytics, identify gaps

in current methods and describe new scenarios to be unlocked in the next few years in the context of the current knowledge.

2.2 Introduction

During the past two decades, the human microbiome has emerged as a biological system with the potential to significantly influence health and disease (11). Despite our limited understanding regarding its intricate relationship with the host and its environment (12), recent discoveries related to the human microbiome have opened new horizons in food science (13), precision medicine (14), and biotechnology (15) among others. In parallel, advances in genomics and bioinformatics have provided inexpensive tools to acquire biological and clinical data, as well as the tools to translate the data into knowledge (16–24). Given these advances, the integration of diet, gut microbiome, and human health (DGMH) data has the potential to drive a paradigm shift in the way wellness states are measured, diseases are treated, products are designed, and health interventions are administered. To realize this potential, advances in knowledge are required in order to optimize the composition and metabolic dynamics of microbial communities in relation to desired health and performance outcomes; from dietary interventions and bioengineered products to lifestyle changes and the environment (**Figure 2.1**).

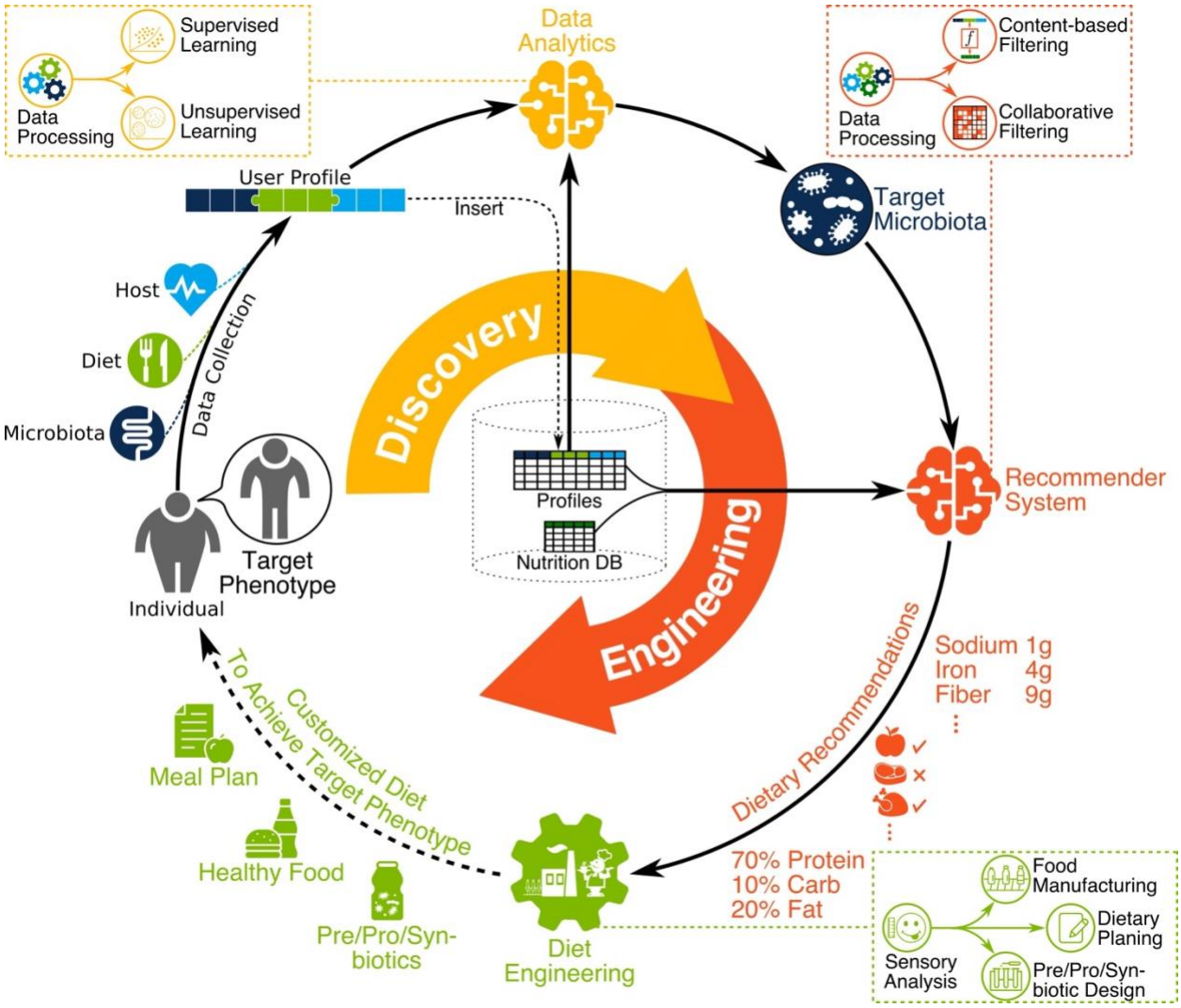

Figure 2.1 The vision for the next nutrition revolution involves microbiome-aware dietary planning and manufacturing. First, DGMH data is collected, homogenized and stored, with any new user data integrated as part of a cohesive compendium. Then, DGMH data are analyzed (*Data Analytics)* to identify the functional characteristics and target microbiota, personalized to the individual and the desired phenotype. This includes data processing followed by supervised and unsupervised learning using user profiles compendium. Bioinformatic tools are used during data processing to extract meaningful information from raw high-throughput data such as metagenomic sequence reads. Then, the *recommendation System* provides dietary recommendations to help achieve target microbiota. This includes the integration of user profiles in a compendium along with nutrition DB proceeded by data processing then content-based and collaborative filtering. Finally, *Diet Engineering* is performed to create dietary products for the user. This includes the design of prebiotics, probiotics, synbiotics, manufactured food, and detailed dietary planning. In practice, taste and flavor of dietary products is very important to help users commit to any given diet, therefore sensory analysis should inform all dietary engineering efforts.

In this article, we summarize the research that has been done related to DGMH, with a focus on

DGMH data and computational methods. We begin with a brief overview of key areas of current

knowledge regarding the interaction between diet, health and the gut microbiome. We then proceed to more extensively review the available data sources and the computational methods currently used, an investigation into the role that machine learning and artificial intelligence (AI) can play in this area, and a summary of the intellectual property (IP) and legislative regulatory landscape. We conclude with recommendations to accelerate research and development efforts through better integration of research resources and tools, especially in the context of computational science and data analytics. A glossary of terms is provided in **Table 2.6**.

In general, the most recent articles reviewing the computational tools for microbiome data focusing on metagenomic data processing methods provide limited guidance on employing machine learning, data analytics and recommendation in the context of DGMH data. The purpose of this manuscript is to help fill this gap by considering relevant literature, describing key challenges and potential solutions, and proposing a framework to improve the potential for research initiatives to accelerate progress in this exciting, and potentially revolutionary, field.

2.2.1 Current Knowledge: Gut microbiota and human health

Emerging evidence suggests that the intestinal microbiota plays a significant role in modulating human health and behavior (see comprehensive reviews (25–27)). Several studies have demonstrated that the human intestinal microbiota is seeded before birth (28), and the mode of delivery influences the composition of the gut microbiota (29,30). The gut of a vaginally born newborn is enriched primarily with the vaginal microbiota from the mother, while a cesarean procedure results in the newborn's gut microbiota being dominated by the microbiota of the mother's skin as well as points of contact at the hospital (31). Microbial diversity is very dynamic during the infancy and increases and converges to an adult-type microbiota by 3 to 5 years of age (32). Evidence is also building suggesting that diet plays a key role in shaping the composition of

microbial communities in the infant's gut. For example, species of beneficial bacteria such as *Lactobacillus* and *Bifidobacterium* have been found to be dominant in breastfed infants while species of harmful bacteria such as *Clostridium*, *Granulicatella*, *Citrobacter*, *Enterobacter*, and *Bilophila* have been found to be dominant in formula-fed infants (33). In addition, breastfed babies have higher gut microbial diversity compared to formula-fed babies, and several studies indicate that the diversity of bacteria is directly connected to health (33,34). An unbalanced composition of the infant's gut microbiota has been linked to several childhood diseases including atopic dermatitis (AD) (35,36) obesity (37) and asthma (38).

The composition of the gut microbiota of an adult human is relatively stable (39), but several factors can influence it, including antibiotic treatment, long-term change in diet, microbial infections, and lifestyle (26,40–42). Several health conditions were linked to the changes in a stable and established gut microbiota such as Crohn's disease (43), psoriatic arthritis (44), type 1 diabetes (45), atopic eczema (34), coeliac disease (46), obesity (47), type 2 diabetes (48) and arterial stiffness (49). However, it is important to note that further research is required to establish direct links between these conditions and the composition of microbial communities in the gut. Several methods, such as oral administration of probiotics and prebiotics and fecal transplant, are being tried out to achieve a healthy gut microbiota and remediate several health-related issues. Occasionally, these methods have reduced the severity of some diseases such as diarrhea, acute upper respiratory tract infections, eczema, Crohn's disease, and ulcerative colitis (50–55). See **Figure 2.2** for illustration of factors affecting the gut microbiota.

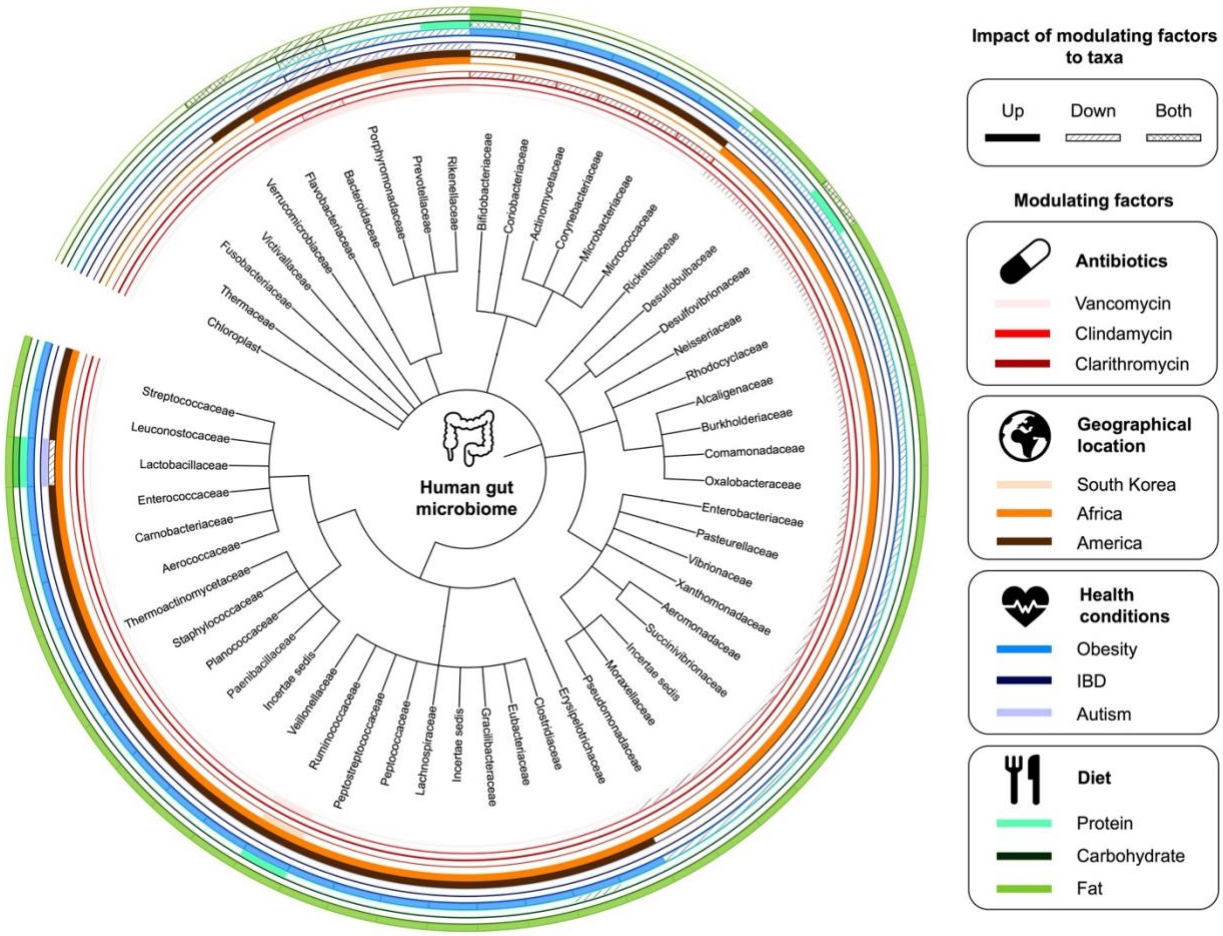

Figure 2.2 Factors affecting the gut microbiota. A summary of human gut microbiome taxonomy at the family level and the corresponding modulating factors.

2.2.2 Data

The increase in size and heterogeneity of information gathered by microbiome studies present great opportunities and serious data analysis challenges (56), with many tools developed to address them (57,58). These bioinformatics tools, quantify low dimensional biological variables such as the relative abundance of microbial species and metabolites, using high dimensional data such as DNA sequence reads and mass spectrometry (MS) signatures as illustrated in **Figure 2.3**. Depending on data quality, sample size and research hypothesis, different information dimensionalities are used. For example, in type1 diabetes research, (59) included gene-level

information while (60) mainly focused on gene functional groups. Different types of data can be grouped based on their sources as described next.

Gut Microbiota Data. Functional characteristics of microbial communities can be revealed using high-throughput meta-metabolomics (61) and meta-proteomics (62,63) using MS technologies. Metagenomic and meta-transcriptomic content of gut microbiota (which give rise to the functional characteristics), can be quantified using DNA sequencing. The most widely used approach for gut microbiota profiling is *marker gene sequencing* which relies on sequencing counts of the hypervariable 16S genes to calculate Operational Taxonomic Units (OTUs) (64). Searching OTUs against reference databases such as Greengenes (65) and SILVA (66), allows inferring relative taxa abundances in a microbiome sample (67). *Whole-genome metagenomics* (i.e. untargeted shotgun metagenomics) (58) is another technique that is both more expensive and of higher resolution, as it not only reveals the microbial community structure, but it can also quantify relative abundances of genes, taxa, conserved functional groups or over-represented pathways. Within-sample (alpha) and cross-sample (beta) diversity of microbiome can be calculated with respect to genetic, taxonomic, functional or metabolic pathway profile of samples (68–73). Shannon index, Chao1 and Abundance-based Coverage Estimator (ACE) are used to measure alpha diversity while UniFrac, weighted UniFrac, and Bray-Curtis calculate beta diversity. In longitudinal studies, the same measures of diversity, or more sophisticated eigenvalue based analyses, can quantify the microbiota stability across timepoints (70,74–76). Jackknifing and bootstrapping are used to estimate the bias in diversity estimates, particularly when estimating the number of species (i.e. species richness) in samples (77). Some of the most significant publicly available microbiome datasets are listed in **Table 2.1.**

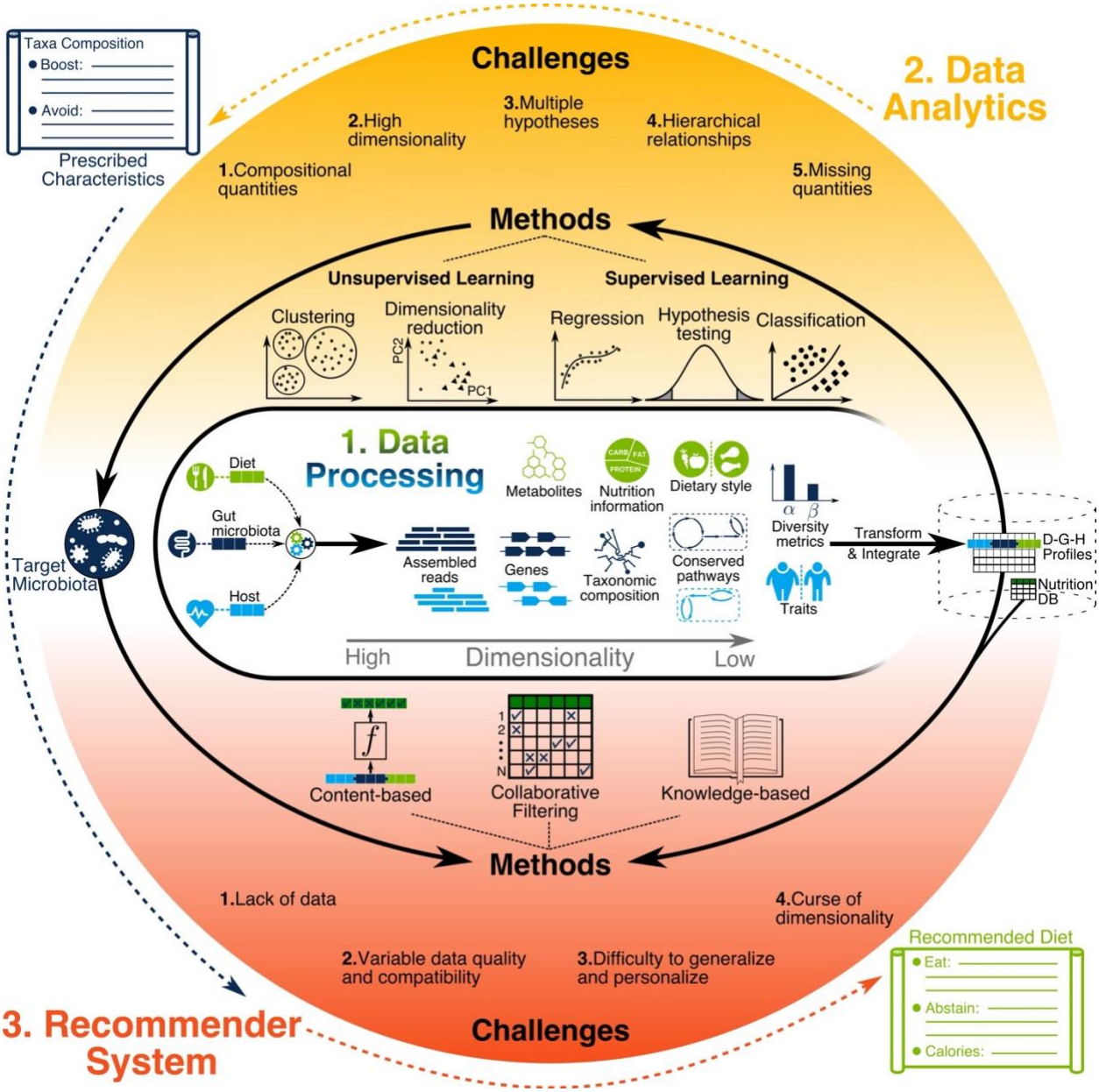

Figure 2.3 Illustration of data processing, data analytics, and recommendation system. Data processing generates diverse types of information with different levels of resolution and dimensionality. Such information needs to be transformed and integrated across all users for building a compendium. Next, data analytics methods are used to discover the characteristics of target microbiota prescribed for individuals to achieve their health objectives. Finally, recommendation system methods use the compendium to find the dietary recommendations for helping individuals achieve the target microbiota.

Project, Database or Repository Name	Number of Cases	Sample Types	Disease Related (Y/N/B)	Data Availability (Y/N/ Conditional)	Website
Human Microbiome Project (HMP1)	300	nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract	N	Y	(78)
Integrative Human Microbiome Project (iHMP): pregnancy and preterm birth (MOMS-PI)	~2000	mouth, skin, vagina, and rectum	Y	Y	(79)
Integrative Human Microbiome Project (iHMP):onset of IBD (IBDMDB)	~90	stool, blood	Y	Y	(79)
Integrative Human Microbiome Project (iHMP): onset of type 2 diabetes (T2D)	~100	fecal, nasal, blood, serum, and urine	Y	Y	(79)
American Gut Project (AGP)	>3000	stool, swabs from skin/ mouth	B	Y	(80)
Personal Genome Project microbiota component (PGP)	>5000	skin/oral/fecal	?	Y	(81)
TwinsUK	>11000	multiple	?	C	(82)
Global Gut project (GG)	531	fecal	N	Y	(83)
Project CARDIOBIOME	>4000	?	?	N	(84)
Pediatric Metabolism and Microbiome Repository (PMMR)	~350	human microbial cell lines, stool, and/or DNA and RNA	Y	N	(85)
Lung HIV Microbiome Project (LHMP)	162	lung, nasal, and/or oropharyngeal cavities	Y	Y	(86)
The Study of the Impact of Long-Term Space Travel on the Astronauts' Microbiome (Microbiome)	9	saliva and gastrointestinal	N	N	(87)
Michigan Microbiome Project (MMP)	?	?	?	N	(88)
uBiome	?	gut, mouth, nose, genitals, and skin	B	C	(89)
Human Oral Microbiome Database (eHOMD)	?	upper digestive and upper respiratory tracts, oral cavity, pharynx, nasal passages, sinuses, and esophagus	?	Y	(90)
Human Pan-Microbe Communities (HPMC)	>1800	gastrointestinal	B	Y	(91)
Curated Metagenomic Data	>5000	multiple	B	Y	(92)
European Nucleotide Archive	?	?	?	Y	(93)
EBI-metagenomics portal samples	>20000	multiple	B	Y	(94)
MG-RAST	>10000	multiple	B	Y	(95)

Table 2.1 Publicly available data from gut microbiota studies

Diet Data. Various types of dietary information are collected in gut microbiome studies. This includes fine-grain information such as mass spectrometry (MS) signatures and metagenomic reads (96) or coarse grain information such as dietary style (e.g. Western vs. Mediterranean diet (97)) from study participants. Diet data collection is often questionnaire-based either through self-reporting or done by a trained interviewer. For self-reporting, food frequency questionnaire (FFQ) and 24-hour dietary recall (24HR) can be used where participants report their dietary intake either every 24-hours or over a longer period through a checklist of food items (98). Dietary record (DR) can also be used where data collection is done when food is consumed (e.g. using smartphones) hence can minimize reliance on participant's memory. After data collection, the intake amount of macronutrients (fat, carbohydrates, protein), micronutrients (vitamins and minerals) and food metabolites can be estimated by querying the food items against food composition databases such as USDA food composition database (99) and the Canadian nutrient file (100). It is also important to consider the usage of drugs, antibiotics, and nutritional supplements in the diet which can be collected through patient health records or as part of the questioners. Note that microbiota of dietary intake can be characterized using metagenomic sequencing as reviewed previously, although it becomes unnecessary when such information is readily available (e.g. probiotics with predefined strains (101)). Some studies perform metabolic characterization of dietary intake directly (96) while others rely on a pre-characterized metabolic profiles (102). Researchers should note that food composition databases characterize only 0.5 % of known nutritional compounds (103). Therefore studies that rely on these databases will not be able to identify health impacts associated with more than 99% of biochemical compounds in food (103).

Host Data. Profiled host information types can be very high dimensional (e.g. high throughput genome sequences (104)) or low dimensional (e.g. obese vs. non-obese (105,106)). Host genotype

data can come from whole-exome sequencing (WES) (107) or genome-wide association study (GWAS) (108,109). It can also be extended by predicting the whole-genome sequence for each individual through genotype imputation software (110) as done in several studies (109,111,112). Host transcriptomic profiles can be assessed directly using microarrays (113,114) and RNA-Seq (115,116), or imputed using tools such as PrediXcan (117) with GWAS data. The genetic and transcriptomic profiles can be summarized into informative lower-dimensional features through gene ontology categories and metabolic pathways using databases such as MetaCyc (118), KEGG (119), Reactome (120) or GO (121). Today, limited microbiome studies perform such analysis (122–124). Other important information such as age, gender, ethnicity, body weight, blood pressure, dietary restrictions and diseases of a host organism can be extracted from medical records, surveys, and interviews.

2.3 Computational Analysis

There have been various reviews concerning microbiome data processing and analysis (16,57,58,125,126). Here we focus on data analytics, machine learning, and AI-based recommendation system methods that enable microbiome-aware systems involving diet and wellness. We provide readers insight into important methods, challenges that arise, suggested solutions as well as blueprints of example scenarios to be used in their research. See (127–129) for further explanation and examples of machine learning methods discussed here.

2.3.1 Microbiome data processing tools

There are a substantial number of microbiome data processing methods and pipelines publicly available that can generate the various types of data discussed. **Table 2.2** provides a representative summary of such methods and pipelines. QIIME (130) and MOTHUR (131) provide a wider range

of options for the user compared to UPARSE (132) but all are popular pipelines. QIIME 2 (133), is now emerging as a powerful replacement to its predecessors partly due to its extensibility and support. Depending on project aims, some of the steps mentioned in **Table 2.2** can be avoided. The focus of most popular data processing methods is on marker gene sequencing data. For whole metagenomic sequencing, however, methods such as Kraken (134), MEGAN (135), MetaPhlAn2 (136) and HUMAnN (137) are used. Such methods are expected to gain more popularity as whole metagenomic sequencing makes its way to become the standard practice with access to more powerful computational hardware and software tools.

Challenges in microbiome data processing. Growth in the variety and complexity of data processing tools presents opportunities but also significant challenges for new investigators. First, although best practices have been suggested (16), tools are still far from a fully automated user experience that would lead to reliable results. Second, microbial genomes with different abundances are sequenced together making metagenomic assembly much more challenging compared to single genome assembly where the sequence coverage is approximately uniform. Third, the amount of uncharacterized microbes (known as microbial dark matter) exacerbates problems associated with unaligned and misaligned sequence reads. Fourth, evaluation of methodology and findings from different studies is difficult since each study may use a different method or a different implementation of the same method in their data processing pipeline. Fifth, data collection and integration of microbiome data from different studies is difficult due to many factors including differences in wet-lab library preparation (e.g. primers used), differences in sequencing devices and their settings (e.g. coverage) and non-uniform methods of formatting and storage for microbiome data and metadata. See (58) for further discussion concerning microbiome data processing challenges.

Steps	Sub-step Descriptions	Highlighted Methods & Their Availability in Popular Pipelines (**Q**IIME, **M**OTHUR, **U**PARSE)
1. Quality Control	Chimera removal & Noise mitigation	Trimmomatic[Q] (138), AmpliconNoise[Q, M] (139), UNOISE[M, U] (140), UCHIME[Q, M, U] (141), Deblur[Q, M] (142) and DADA2[Q] (143)
	Remove host DNA contaminant reads	Bowtie2[Q] (144), BMTagger (145) and DeconSeq (146)
2. Sequence Assembly	*De novo* read assembly	MEGAHIT(147), MAFFT[Q, M] (148), UCLUST[Q, U] (149) and metaSPAdes[Q, M] (150)
	Read alignment to annotated database	DIAMOND(151), NAST[Q, M] (152), USEARCH[Q, U] (149) and VSEARCH[Q, M] (153)
3. OTU Analysis	Assignment of reads to OTUs	UPARSE-OTU[U] (132), Kraken (134), MetaPhlAn2[Q] (136) and DOTUR[M] (154)
4. Functional Profiling	Functional profiling and prediction.	MEGAN (135), HUMAnN (137), MetaCLADE, MOCAT (155) and PICRUSt (67)
5. Diversity Analysis	Diversity, evenness and richness metrics	Alpha (e.g. Chao1[Q, M, U]) and Beta (e.g. Jaccard[Q, M, U])

Table 2.2 A summary of highlighted methods and pipelines for microbiome data processing.

2.3.2 Data Analytics and Machine Learning

Data analytics and data processing are closely related. In this review, data processing is considered to be the steps necessary for converting the raw data such as metagenomics sequence reads, into biologically meaningful representations such as OTU counts using bioinformatics tools, some of which are done in the sequencing device itself. Data analytics, however, often starts after the integration of processed sample data from various information sources (i.e., microbiota, diet, and host) as illustrated in **Figure 2.3**. In most cases, all samples are from a single study which helps ensure consistency with respect to the experimental settings and data processing protocols used. Furthermore, limited resources force the researchers to narrow their data collection to particular information types in order to have sufficient statistical power for hypothesis testing. A recent

increase in the number of microbiome studies with publicly available data has enabled cross-study data integration (156–161). In such cases, extra precautions are necessary to minimize biases introduced by inconsistencies among datasets during data collection, sample preparation, sequencing and data processing.

Challenges in microbiome data analysis. A number of challenges arise when analyzing microbiome data as summarized in **Table 2.3**. The first challenge is due to *compositional quantities* in microbiome data. Quantities such as the number of reads assigned to a given species, which can only be interpreted as proportions, are called compositional. These quantities cannot be compared directly across multiple samples. Conclusions should not be made based on the number of reads assigned to individual sample features (e.g. OTUs, genes and functional groups) since they do not represent absolute abundances due to instrumental limitations (162). Instead, the assigned number of reads should be converted to relative abundances and analyzed with that in mind. Some studies perform rarefaction to adjust for differences in library size due to unexhaustive metagenomic sampling. Although several pipelines provide this functionality, it has been found inadmissible for metagenomics microbiome studies as it discards many reads leading to decreased sensitivity in differential abundance testing (163), and biased estimates for alpha diversity (164). The second challenge is due to the *high dimensionality* associated with OMICS data. Datasets in which items are characterized by a high number of features while the number of items is limited are called high dimensional. In microbiome studies, a limited number of individuals are characterized using many host, diet, and microbiome features leading to high dimensional datasets (165). Dimensionality can be reduced by grouping OTUs into phylogenetic variables, regularization or unsupervised dimensionality reduction (explained below). The third challenge is about testing *multiple hypotheses* in exploratory analysis. It relates to the fact that, as the number

of hypotheses increases, the chance of false discoveries also increases. This can be addressed by increasing sample size and p-value adjustment (explained below). The fourth challenge relates to *hierarchical relationships* amongst bacterial species due to their shared ancestors. Assumptions such as independence among samples may not hold which leads to wrong estimations of correlation (166) and phylogeny-aware methods to address the issue. The fifth challenge is about *missing quantities* in sampled data. For example, when marker gene sequencing is used, quantities relating to the amounts of functional genes in the microbiome are not directly available (i.e. missing). Identifying functions of microbial organisms is important for understanding the gut microbiota. Such information can be estimated using meta-transcriptomics data which is often not available. Data imputation tools such as PICRUSt (67), help to mitigate this through gene imputation based on annotated databases.

Next, we review methods for identifying microbiota characteristics associated with host phenotypes of interest. They can be categorized into two main groups: supervised learning and unsupervised learning. Supervised learning methods require labeled data while unsupervised learning methods can be used when records are not labeled. More advanced methods such as semi-supervised learning (167); which can take advantage of both labeled and unlabeled data and transfer learning (168); which can transfer knowledge learned from one task to another are not discussed here.

Challenges in Microbiome Data Analysis	Examples & Solutions
1. Compositional quantities: Metagenomic data processing provides read counts for discovered entities such as genes, species, and OTUs from a given sample. These read counts are only meaningful within a sample.	**Example:** Metagenomic analysis of feces samples tells us that Person A has five reads mapped to bacterium *E. coli* while person B has ten. Can we conclude that this bacterium is more populated in the gut of person B compared to person A? *Answer*: No, read counts cannot be compared across samples. **Solutions:** (I) Convert read counts to relative abundances before comparison. (II) If an optimization problem is defined using read counts, add constraint for total counts per sample.
2. High dimensionality: Metagenomic data processing results in many entities such as genes and species discovered for each sample which may not be shared amongst multiple samples. During data aggregation, one dimension is associated to each entity resulting in a high number of dimensions compared to the number of samples.	**Example**: Metagenomic data processing of feces samples from twenty individuals results in relative abundances for ten microbial families per sample. Can we use classical linear regression to predict an individual's age using relative abundances from aggregated data? *Answer:* No, aggregating twenty samples results in more than twenty microbial families. **Solutions:** (I) Use dimensionality reduction such as PCA prior to regression. (II) Use regularized linear regression such as Lasso. (III) Use microbial abundances of higher-order taxonomic ranks such as phylum instead of family.
3. Multiple hypotheses: The high-dimensional nature of metagenomic data allows the researcher to generate a large number of hypotheses which leads to seeing patterns that simply occur due to random chance. This is sometimes called "the high probability of low probability events".	**Example**: Metagenomic data processing provides relative microbial abundances at species level using feces samples of two hundred individuals, half of which are diagnosed with Crohn's disease and the rest are healthy. Performing a t-test identifies that the relative abundance of 40 species (amongst 1000) are significantly different between microbiota of sick and healthy individuals (*p-value* < 0.05). Is this result correct? *Answer:* No, the standard threshold of 0.05 for *p-value* is only acceptable when a single hypothesis is involved while the t-test is performed 1000 times leading to many false discoveries. **Solution:** Calculate FDR adjusted *p-value* (i.e. *q-value*) of 0.05 to control the false discovery rate.
4. Hierarchical relationships: Assumptions of independence do not hold in microbiome data since taxonomic variables (e.g. species and OTUs) have known hierarchical relationships due to genetic and phenotypic similarities. Therefore common statistical techniques that assume independence between variables are problematic.	**Example:** Beta-diversity can be used to calculate the similarity between groups of microbiome samples. Can we simply calculate the Beta-diversity using standard Euclidean distance between relative abundances at a given taxonomic order? *Answer:* No, Euclidean distance doesn't take into account the similarity between species. **Solution:** Use phylogeny-aware metrics such as UniFrac distance instead which takes into account the phylogenetic tree when calculating distances.
5. Missing quantities: Metagenomic data often lacks information about the functions of the microbial communities which can only be estimated using meta-transcriptomics or meta-proteomics. However, deciphering microbiota's function is a major goal in microbiome studies.	**Example:** In one case, metagenomic data processing from marker-gene data has provided us use with relative abundances at the genus level but we do not know the possible functions of the microbiota in terms of proteins that it can produce. Should we abandon further analysis? *Answer:* No, although we don't have direct information about proteins, we can infer. **Solution:** Databases such as Greengenes contain the whole-genome sequence of identified species at various taxonomic orders which can be used for gene and protein inference.

Table 2.3 Key challenges that arise in microbiome data analysis with examples and suggested solutions.

2.3.3 Supervised Learning

2.3.3.1 Hypothesis Testing and Variation Analysis.

Analysis of variation may involve single or multiple variables. For a single variable hypothesis, the Student's t-test or non-parametric tests such as Wilcoxon rank-sum or Kruskal-Wallis can be used. For example, the t-test is used to show that patients with ADHD have a lower alpha-diversity index of gut microbiota compared to healthy controls (169). Investigators should ensure that underlying assumptions of t-test (i.e. normal distribution) are supported by data particularly when the sample size is small. Non-parametric tests are good alternatives when such assumptions do not hold. For example, the Wilcoxon rank-sum test is used on predicted pathway data suggesting that enzymes in the "Glycan Biosynthesis and Degradation" pathway increase in summer compared to winter (170). In cases where a statistical test is repeated with different hypotheses (i.e. multiple hypothesis testing), the statistical significance should be adjusted by methods such as FDR adjustment (i.e. q-value) (171) or Holm's procedure (172).

When our hypothesis corresponds to multiple variables, MANOVA (173) or non-parametric alternatives such as PERMANOVA (174) or ANOSIM (175) can be used. The samples are first assigned to multiple groups (e.g. based on some feature values). The goal is to quantify how much this grouping can explain the distribution of values in a given sample feature (response variable). The simplest case is the popular method called analysis of variance (ANOVA) considering a single response variable with a normal distribution. In one study, two bacterial phyla (Bacteroidetes and Firmicutes) are identified using ANOVA with different relative abundance in microbiota of children living in a rural African village compared to European children (176). ANOVA can be generalized to multivariate analysis of variance (MANOVA) when we can have multiple response variables. For example, it is used to investigate the overall difference in composition between the

microbiota of children with Prader–Willi syndrome and children with simple obesity, before and after treatment (177). In many cases, normal distribution assumptions do not hold hence non-parametric methods are used. In one study, PERMANOVA is used to detect taxonomic differences in the microbiota of patients with Crohn's disease when compared to healthy controls (178).

2.3.3.2 Regression and Correlation Analysis.

A general understanding of the extent of association among pairs of variables can be achieved using correlation analysis. Correlation metrics measure different types of relationships. For example, Bray–Curtis measures abundance similarities (179), Pearson correlation coefficient quantifies linear relationships and Spearman correlation coefficient quantifies rank relationships (180). In (181), the authors perform a simulation-based comparison on a range of correlation metrics for microbiome data. Metrics such as SparCC (182) and LSA (183) perform particularly better as they are designed to capture complex relationships in compositional microbiome data. For example, SparCC is used for analyzing the TwinUK dataset to identify bacterial taxa whose abundances are influenced by host genetics (184). This was done by creating a correlation network between microbial families based on their intraclass correlation. More recently, the phylogenetic isometric log-ratio (PhILR) transform has been introduced (185) to transform compositional data into non-compositional space where standard data analytic techniques are applicable. Usage of such transformations should be limited to features that are compositional and phylogenetic in nature.

Regression methods aim to predict the change in one continuous variable using other variables. Correlation analysis can be considered a special case of regression with a single input variable. Standard linear regression can be used for various DGMH predictive tasks. However, when variables relate to OTU abundances, the typical assumptions of a linear relationship, normal

distribution and dependence do not hold. For example, when the goal is to predict the composition of OTUs (normalized for summing up to one (126))), zero-inflated continuous distributions are used. Often a two-part regression model is used where part I is a logistical model to calculate the probability that the given OTU is present. Part II is a generalized linear regression using beta distribution to predict relative abundance assuming the presence of OTU in the sample (186–188). Phylogenetic comparative methods (PCMs) such as phylogenetic generalized least squares (PGLS) are used to control for dependence among observations given the phylogenetic hierarchies (189). Ignoring the phylogenetic ancestry of microbial species can increase the chance of false discovery in regression models (166). PCMs are not widely used in microbiome studies today which may be one reason for a high number of false positives that can be alleviated using them (190).

To investigate the correlation between two groups of variables (e.g. abundances of microbiome OTUs and metabolites), canonical correlation analysis (CCA), can be used (191). CCA finds linear transformation pairs that are maximally correlated when applied to data while ensuring orthogonality for different transformation pairs. The original CCA, however, fails for high dimensional microbiome data when the number of variables exceeds the number of samples. This can be addressed using regularization giving rise to sparse CCA methods (192). For example, a sparse CCA is applied to investigate correlations between the gut microbiome and metabolome features in type 1 diabetes (193).

2.3.3.3 Classification.

In supervised classification, the goal is to build a predictive model (classifier) using labeled training data. The labels can have binary or categorical values (in contrast to regression where labels are continuous and numerical). In one study a classifier was built to predict the geographical origin of sample donors using relative OTU abundances estimated from 16s rRNA gut samples

(83). This was done using the method called Random Forests (RF) which constructs an ensemble of decision trees (194). In a different study, the classification task was to identify healthy vs. unhealthy donors given relative OTU abundance data (including species level) coming from shotgun metagenomics sequencing of gut (as well as other body sites) (158). In addition to RF, they used the support vector machine (SVM) classifier which is a powerful method for building generalizable and interpretable models and is mathematically well understood (195). In their study, RF classifiers performed better than SVM except in a few datasets. Both RF and SVM have built-in capability to deal with overfitting issues that arise in high-dimensional datasets. RF achieves this using an ensemble-based technique where the prediction is made based on predictions from many trained classifiers. In SVM, parameters of the predictive model are constrained based on a priori defined criteria. Note that constraining the model parameters is often mathematically equivalent to regularization (196). In both cases, the objective is to minimize the value of a loss function that calculates the overall error in model predictions. When regularization is used, the loss function not only depends on prediction errors but also on the magnitude of model parameters. For example in L1 regularization, the absolute values of model parameters are scaled and added to the loss function. Therefore, when two models have a similar error, the model with smaller parameter values will be selected during training. L1 regularization is commonly used for feature selection by picking only the non-zero features of the trained model because such a model achieves a low prediction error while using a subset of features.

Artificial neural networks (ANN) can also be used for classification and shown to outperform other techniques in many areas of biology (3,197–199) as well as computer vision and natural language processing to name a few (200). Recently, a new ANN-based method called Regularization of Learning Networks (RLN) is designed and evaluated on microbiome data. RLN provides an

efficient way for tuning regularization parameters of a neural network when a different regularization coefficient is assigned for each parameter (201). They use RLN to predict human traits (e.g. BMI, Cholesterol) from estimated relative OTU abundances and gene abundances. We expect the development of new classification methods that can deal with the aforementioned challenges arising in DGMH data by considering the biological phenomenon, properties of measurement instruments and upstream data processing pipelines.

2.3.4 Unsupervised Learning

2.3.4.1 Dimensionality Reduction

High-dimensional datasets can provide a high resolution and multifaceted view of a phenomenon such as gut microbiota. Predictive performance in data analytics can increase significantly given such data. Many data analytics methods, however, fall short when presented with high-dimensional data which necessitates using dimensionality reduction (DR). Once dimensionality is reduced, data visualization and analytics become more accessible. Principal component analysis (PCA), is one of the most widely used DR methods. It replaces the original features with a few uncorrelated features called principal components (PCs) which are linear combinations of the original features and preserve most of the variance within the data. In one study, PCA was applied to predicted abundances of about 10 million genes from the gut microbiota of donors (202). Reducing dimensionality from 10 million to two dimensions only, enabled visualization of data on standard two-dimensional scatter-plot (i.e. PCA plot) showing a clear distinction between the microbiota of Danish and Chinese donors. In another study, the top five PCs of individual bacteria's genome (sequenced from infant fecal samples) were used to create a classifier for predicting antibiotic resistance (203).

The relationships amongst features in a microbiome study can be used in DR giving rise to various factor analysis (FA) methods as we review here briefly. Multiple factor analysis (MFA) is an extension of PCA that considers predefined grouping of features during DR to ensure equal representation for all groups of features in derived PCs (204). In one study (205), MFA is used for simultaneous 2D visualization of host and microbiome features (see (206,207) for other examples). Exploratory factor analysis (EFA) is used to identify unobserved latent features called factors to explain the correlations amongst observed features (208). Factors that are identified by EFA are uncorrelated to each other similar to PCs in PCA; however, PCs are used to explain overall variance instead of correlations. EFA has been used in a recent study to extract four factors explaining the correlations amongst 25 top taxa for studying the association of microbiome with early childhood Neurodevelopmental outcome in 309 infants (209). Confirmatory factor analysis (CFA) and structural equation modeling (SEM) can be used to examine the extent to which a hypothesized model of latent features and their relationships with observed variables, are supported by the data (210). In a recent study, a theoretical framework is proposed and examined using CFA to model the influence of maternal and infant factors on the milk microbiota (211). The R packages lavaan (212) and FactoMineR (213), as well as the IBM SPSS software (214), are widely used for factor analysis.

Another related method is principal coordinate analysis (PCoA) also called multidimensional scaling (MDS) (215) which is commonly employed for 2-3 dimensional visualization of beta diversity. It can deal with situations where distances between individual feature vectors from samples cannot be used directly (e.g. due to significant sparsity and phylogenetic relationships). PCoA takes a matrix of distances between samples (e.g. UniFrac distance between OTU abundances of a pair of sample donors) as input. It then assigns new coordinates such as PC1 and

PC2 to each sample such that the Euclidean distances in the new coordinate are similar to the ones in the matrix. For example, PCoA was applied given UniFrac distances between OTU abundances (from 16S rRNA samples) from the gut microbiota of donors (83). Two-dimensional visualization using PC1 and PC2 showed that the gut microbiota of donors who lived in the US is distinct from the ones from Amerindian and Malawian villages.

Linear discriminant analysis (LDA), is also a DR technique although supervised and closely related to regression and ANOVA. Unlike PCA and PCoA, it requires class labels. It generates new features that are linear combinations of the original ones while separating samples with respect to their class labels. In one study, LDA was used to distinguish gut microbiota samples based on diet but not for dimensionality reduction (216). Successful usage of LDA for high dimensional microbiome data may require regularization to account for overfitting as similarly used for high-dimensional microarray (217).

The optimal amount of reduction in dimensionality (e.g. the number of principal components) varies given the data and the task downstream. For data visualization tasks, it is largely constrained by the limitations of human visual perception (three dimensional). For downstream supervised learning tasks, we are often interested in the maximum amount of dimensionality reduction without a significant decrease in predictive power. This is showcased in (218) where the impact of the amount of dimensionality reduction on classification performance is evaluated for gene expression data.

2.3.4.2 Cluster Analysis.

Similar microbial communities are expected to exhibit analogous effects on the host organism (219). Once a similarity measure is defined, various cluster analysis methods can be used to find groups of samples with similar microbiota. In one study, three robust microbiota clusters (called

enterotypes) were identified using cluster analysis from 16s rRNA data of fecal samples (220). It was later shown that such clustering results are not only sensitive to data, but also to choices made during analysis (221). We enumerate four important choices impacting cluster analysis results (other than upstream data processing). First, is the distance measure. Standard distance metrics such as the Euclidean and Manhattan distance are simple, well understood and supported in many clustering libraries. Applicability of such metrics depends on prior compositionality aware transformations such as ILR. Beta-diversity metrics such as weighted and unweighted UniFrac distances are designed for microbiome analysis considering compositionality and phylogenetic dependencies of microbiome data. Researchers should pay attention to the properties of the distance metric used in order to better understand the clustering results. Second, is the clustering algorithm. Algorithms such as Partition Around Medoids (222) and Hierarchical Clustering (223) can employ various distance metrics. Others such as k-means are tied to a single distance measure but computationally less demanding. Third, is the number of clusters. Clustering algorithms often require the number of clusters to be provided as input. When unknown, the number that provides higher cluster scoring is picked. Prediction strength (224), silhouette index (225) and Calinski-Harabasz (226) are popular cluster scoring metrics. Fourth, is the method used to identify the robustness of clustering results. Often a cluster scoring metric that is not used to identify the number of clusters, is used as a robustness measure. Recent studies consider the effect of the above choices during cluster analysis to better understand how results can be generalized (227,228).

The integration of data from disparate omics data types (also called integrative omics) and other heterogeneous metadata enables a more comprehensive look into the underlying biology (229). Integrative omics data analysis methods have been categorized into three types (230). First is *data-to-data,* where disparate data types are analyzed together. For example, CCA can be used to

investigate the correlations between metagenomics and metabolomics data as discussed before. Second is *data-to-knowledge,* where the knowledge gained from analyzing some data types are used to inform analysis of other data types. For example, a metagenomics analysis of Colon cancer patients can lead to candidate genes to be investigated further using targeted proteomics analysis. Third is *knowledge-to-knowledge,* where the data types are initially analyzed separately but the acquired knowledge is integrated together afterward to either identify hypotheses that are supported by multiple data types or create a more complete view of a given phenomenon. For example, differentially expressed genes, and differentially abundant metabolites in the digestive tract of patients with Crohn's disease can be used together for narrowing down pathways involved in disease etiology. See (229–232) for comprehensive reviews.

2.3.5 Recommendation systems and artificial intelligence

The human microbiome is referred to as “our second genome” and has a major influence on our health (233). Although it is known for its resilience (70,75), unlike the human genome it has considerable plasticity hence providing ample opportunities in the design of new types of food, medical interventions and dietary recommendations (234). Despite recent progress in microbiome research, switching from population-wide dietary recommendations to microbiome-aware recommendations is not yet realized. Once a personalized healthy target microbiome is identified using data analytics methods, a recommendation system (RS) can utilize this information to suggest the path towards establishing it in the host and ensuring the health benefits. One approach is to use a knowledge-based RS where recommendations are made using a limited number of approved drugs and dietary prescriptions. Although this would be a good starting point, such a system would be limiting in its ability to provide precise and personalized recommendations, that usually need a platform that can create new products or processes on a case-by-case basis. Recent

studies simulate a virtual gut microbiome by integrating known metabolic pathways of microbial species with the individual's microbiome and diet (24,21–23). Such mechanistic modeling is very promising however it is currently hindered by numerous challenges such as incomplete characterization of individual's gut and metabolic pathways of their microbiome. There is considerable research on AI-based RS related to food, drug design and health (235,236) but its application with microbiome data is in its early stages (18,20,19). Commercial investments in this area have already started, with companies such as UBiome and DayTwo, using 16S rRNA technology to provide insights into our personal microbiota and suggest dietary recommendations.

Recommendation system is defined as "any system that guides a user in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output" (237). Microbiome-aware diet recommendations can be generated from knowledge-based, content-based or collaborative filtering as described next.

2.3.5.1 Knowledge-based Recommendation System

An ideal knowledge-based RS would be based on *in silico* model that can correctly simulate an individual's gut. It requires proper characterization of the gut microbiome, human intestinal cells, intestinal and dietary metabolite concentrations, their interactions through metabolic pathways and realistic objective functions for modeling such complex dynamics. Such knowledge-based RS is devised in a recent study involving 28 patients with Crohn's disease and 26 healthy individuals (21). They integrate genome-scale metabolomic reconstructions (GENREs) of 818 microbes from http://vmh.life (238) with the individual's microbiome abundances after metagenomic data processing in the R package BacArena (239). Their *in silico* simulations provide personalized metabolic supplements for improving patient's SCFA levels. Earlier studies have created a metabolic model of gut microbiome on a smaller scale (24). See (240) for a comprehensive review.

Despite the promise, there are several challenges for the application of such knowledge-based RS. The first challenge, is the limited availability and accuracy of GENREs for gut microbes. A recent study has identified 1,520 unique microbes in the human gut (241) while the number of microbes that have GENREs is only 818 (238). In one study (242) 75% of the GENREs required updates (from previously constructed GENREs (243)) so that *in silico* simulations can recapitulate growth on new media. This suggests that *in silico* GENREs of the gut microbiome are far from complete however progress is being made towards closing this gap. The second challenge is the metabolic characterization of the media inside the intestine on which gut microbes grow. This includes identifying the dietary metabolites available to microbes at different sites in the gut which necessitates meticulous dietary data processing. The third challenge relates to the computational complexity of *in silico* simulations which increases as host and microbial GENREs become more comprehensive. Although more challenges can be enumerated is would beyond the scope of this article.

2.3.5.2 Content-based Recommendation System

In content-based RS, the recommendations are made based on the item's content (often characterized using item features). This is in contrast to collaborative filtering RS where recommendations are based on preferences of other users for each item. In one landmark study (18), authors use content-based RS for meal recommendations with the goal of improving post-meal glucose levels. Each meal is first characterized based on its nutritional profile (macronutrients and micronutrients). Then a regression model is trained to predict post-meal glucose level based on the meal's nutritional profile, individual's microbiome features, and other personal information. For each new user and meal, post-meal glucose levels are predicted by the model and the meal with minimum post-meal glucose level is recommended to the user. The same methodology is

used in a later study using only microbiome features of individuals to predict post-meal glucose levels in a bread-type recommendation system (19). Several challenges arise when building content-based RS. The first challenge is variable data quality and compatibility. When a group of users (or items) are overrepresented in the data, the predictive model tends to be biased towards their favorite items. As a result, the quality of recommendations will be highly variable. Stratified sampling can be used to alleviate this issue. The second challenge is the difficulty to generalize and personalize recommendations particularly when feature vectors are not informative for predictions (also relevant to the "missing quantities" challenge mentioned in Table 3). This is in contrast to collaborative filtering RS, where latent features are learned instead of being defined a priori. Hybrid RS methods are designed to take advantage of collaborative filtering RS to address such inherent challenges in context-based RS (and vice versa) (237). For an extensive review of context-based RS, methods see (244).

2.3.5.3 Collaborative Filtering Recommendation System

In collaborative filtering RS, each user is characterized by the items (foods or ingredients here) they have previously rated, bought or generally acted upon. Recommendations are given based on the idea that users who assign the same rating to existing items are expected to have a similar rating profile for all items. Matrix completion is one of the most popular collaborative filtering methods (245,246). User assigned scores are first organized in a sparse matrix where columns correspond to all different items and rows to various users. In cases where most users only have evaluated a few items, most of the matrix remains empty. Matrix completion fills the rest of the matrix through the similarities discovered amongst users and items. See (245,246) for comprehensive review. Collaborative filtering RS has not been used for microbiome-aware food recommendations. We describe an example here to showcase how it can be used. Consider a matrix where each column

corresponds to a dietary plan and each row to a person, a specific value can represent gut microbiome alpha diversity during the time which the user followed a particular dietary plan. Assuming that each person has only tried a few dietary plans, most of the matrix will be empty. Here we can use matrix completion to fill the matrix with predicted alpha diversities to have a complete matrix. This can be used to recommend dietary plans for a person with the goal of maximizing gut microbiota diversity. Several challenges arise in collaborative filtering RS. The first challenge is the lack of data for new users ("cold-start"). Note that the recommendations rely on similarities amongst users while new users have not tried any of the items available in the database. The second challenge is the curse of dimensionality. As the number of items increases, the chance of having user scores for the same item combinations decreases hence items and users become equally dissimilar (also relevant to the "high dimensionality" challenge in **Table 2.3**). In such cases, hybrid RS can be used. Next, we bring up a few example scenarios.

2.3.6 Example Scenarios

We discussed various data analytics and recommendation system methods for microbiome discovery and diet engineering as illustrated in **Figure 2.1** and **Figure 2.3**. Applicability of each method depends on research objectives and data availability. Here we explain particular scenarios illustrated in **Figure 2.4** as blueprints for integrating relevant techniques in a single pipeline. In **scenario A,** the goal is to identify metabolic pathways that are enriched in the gut microbiome of healthy adults using 16S rRNA data (see (159–161) for similar works). In **scenario B**, the goal is to provide recommended probiotic intake for supporting a healthy gut microbiome. First, the study participants would be profiled based on the probiotic products they consume (each containing specific OTUs) as well as their gut microbiome. Next, microbiome scores will be calculated for each participant based on the distance between enriched pathways of their microbiome and the

target healthy microbiome. Then a regression model is trained to predict microbiome scores based on OTU intakes. Finally, the OTU intake concentration that is predicted to have an optimal microbiome score, would be used as the recommended probiotic intake. In **scenario C**, the goal is to identify optimal diets for health, performance, and disease. A compendium needs to be built following a consistent data collection and processing pipeline for study participants**.** The compendium serves the training data necessary for building machine learning models to predict health metrics such as post-meal glucose level (18,19) or post-dieting weight regain (20). The predictive models can then be used as the key part of a recommendation system by identifying the expected impact of a given diet on health for new individuals. In **scenario D**, the goal is to recommend metabolic supplements needed by an individual's microbiota to secrete vital compounds. First, OTU abundances of each individual are identified using a metagenomic data processing pipeline. Then individual gut metabolic pathways are reconstructed using online resources such as Virtual Metabolic Human database (238). Finally, constraint-based reconstruction and analysis (COBRA) tools (22,239) are used to perform *in silico* simulation of GENREs to identify metabolic intake requirements to secrete vital compounds of interest. This mechanistically sound approach is used in a few recent studies (24,21).

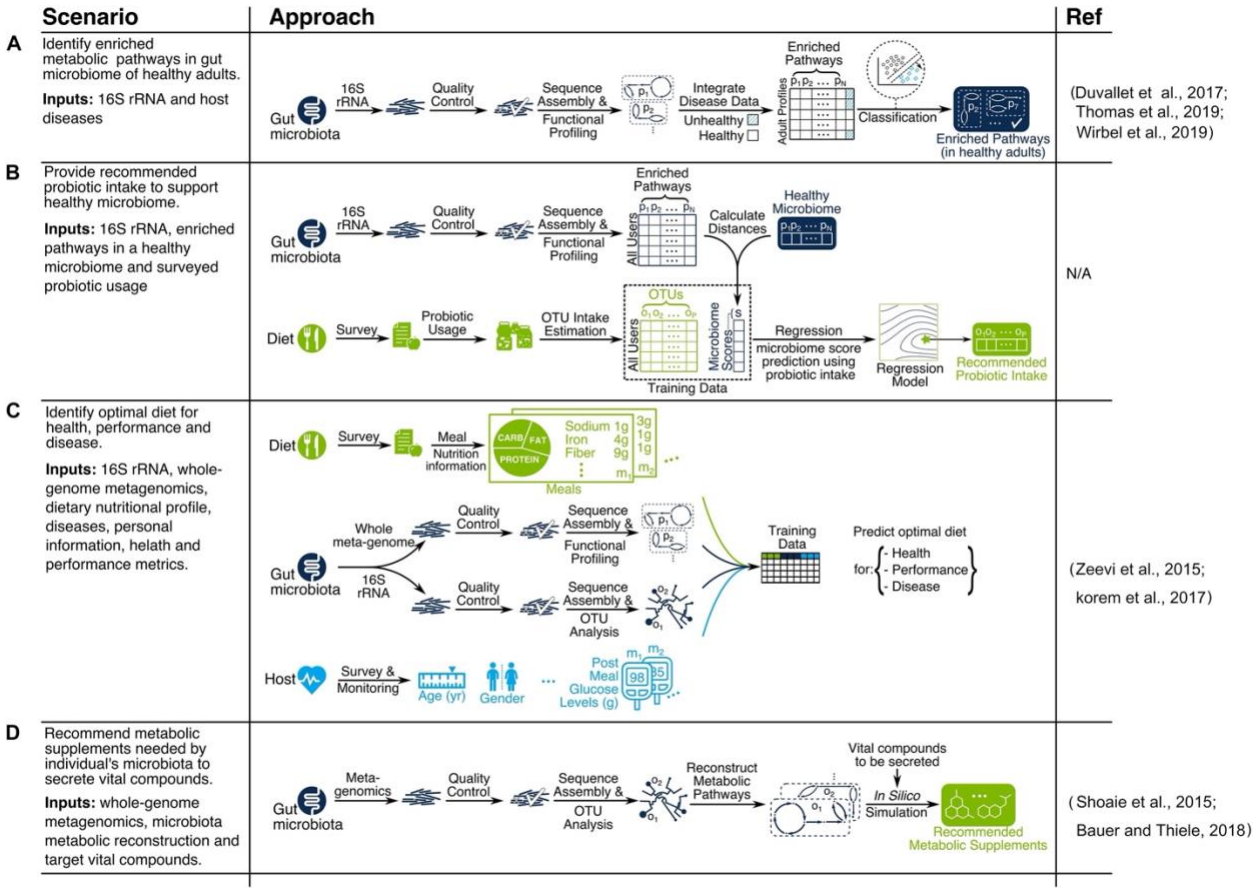

Figure 2.4 Examples of microbiome-aware diet recommendation pipelines.

Study Description	Dietary Variables	Metagenomic Technology	Ref
Personalized meal recommendation system uses personal, microbiome and dietary features to select an optimal meal for lowering post-meal glucose levels in patients with type II diabetes.	Micro and macronutrients	16S rRNA & whole metagenomics	(18)
Microbiome features enable accurate prediction of an individual's glycemic response to different bread types.	Bread type	16S rRNA & whole metagenomics	(19)
Accurate prediction of weight regain given normal vs. high-fat diet in mice is enabled using a microbiome-based predictor.	Dietary fat	16S rRNA	(20)
Personalized metabolite supplement recommendations for Crohn's disease are made using *in silico* simulation of reconstructed metabolic pathways from gut microbiome (773 microbes).	Metabolic supplements	whole metagenomics	(21)
Fecal amino acid levels are predicted given dietary macronutrients through *in silico* simulation of metabolic pathways from gut microbiome (four microbes) and host cells.	Macronutrients	16S rRNA	(24)

Table 2.4 Highlighted microbiome-aware diet recommendation studies.

2.4 Intellectual property development

The potential application impact generated by research on the relationship between the gut microbiome and diet can be visualized by the abundant number of patent applications on the topic, as well as more generally in the field of microbiome and health research. A search for “gut microbiome” and “diet” returns over 2,500 patents on Google, deposited by universities, institutes and companies such as MicroBiome, Microbiome Therapeutics, Gutguide, Whole Biome Inc., UBiome and others, from as early as 2004. However, it is important to note that most of these hits are less than a decade old, demonstrating the relatively early stages in which this area still resides. The exponential growth in patent applications related to the microbiome since 2007 correlates to a similar curve for the academic publications in the same period (247).

One of the earliest patent applications (US20050239706A1) available related to the topic of the microbiome and nutrition describes methods to regulate weight by manipulating the gut microbiome. Additional patents also aim to use the gut microbiome as a therapeutic target, monitoring and altering the composition with the goal of manipulating the host phenotype such as weight gain/loss and obesity. In general, weight management with the manipulation of the gut microbiome (US20110123501A1, US20100172874A1) appears as a favored theme for early patent applications in the area of microbiome and diet. Several patents describe novel probiotics and their uses (WO2007136553A2), often relating them to specific target phenotypes such as weight loss (EP2178543B1, US9371510B2, US9113641B2, EP2216036A1, EP2296489A1, WO2010091991A1). Multiple applications for probiotics focused on weight loss were deposited by Nestec SA, which offers research and consulting services to the food company Nestlé S.A.

Patent number	Name	Owner	Year
US20100172874A1	Gut microbiome as a biomarker and therapeutic target for treating obesity or an obesity related disorder	Washington University in St Louis	06
WO2007136553A2	Bacterial strains, compositions includig same and probiotic use thereof	Benson et al.	06
US20110123501A1	Gut flora and weight management	Nestec SA	07
EP2178543B1	Lactobacillus rhamnosus and weight control	Nestec SA	07
US9371510B2	Probiotic compositions and methods for inducing and supporting weight loss	Brenda E. Moore	07
US9113641B2	Probiotic bacteria and regulation of fat storage	Arla Foods amba	07
EP2296489A1	Lactobacillus paracasei and weight control	Nestec SA	08
EP2216036A1	Lactobacillus rhamnosus NCC4007, a probiotic mixture and weight control	Nestec SA	09
WO2010091991A1	Lactobacillus helveticus cncm i-4095 and weight control	Arigoni et al.	09
US20100331641A1	Devices for continual monitoring and introduction of gastrointestinal microbes	Gearbox LLC	09
US20160074505A1	Method and System for Targeting the Microbiome to Promote Health and Treat Allergic and Inflammatory Diseases	Kovarik et al.	09
US20120058094A1	Compositions and methods for treating obesity and related disorders by characterizing and restoring mammalian bacterial microbiota	New York University Dow Global Technologies LLC	10
US9040101B2	Method to treat diabetes utilizing a gastrointestinal microbiome modulating composition	MicroBiome Therapeutics LLC	11
US20170348359A1	Method and System for Treating Cancer and Other Age-Related Diseases by Extending the Health span of a Human	Kovarik et al.	11
US20170281091A1	Capsule device and methodology for discovery of gut microbe roles in diseases with origin in gut	Lowell Zane Shuck	12
US20170372027A1	Method and system for microbiome-derived diagnostics and therapeutics for locomotor system conditions	uBiome Inc	14
US20170286620A1	Method and system for microbiome-derived diagnostics and therapeutics	uBiome Inc	14
US20190030095A1	Methods and compositions relating to microbial treatment and diagnosis of disorders	Whole Biome Inc	14
WO2017216820A1	Metagenomic method for in vitro diagnosis of gut dysbiosis	Putignani et al.	16
WO2017171563A1	Beta-caseins and cognitive function	Clarke et al.	16
WO2017160711A1	Modulation of the gut microbiome to treat mental disorders or diseases of the central nervous system	Strandwitz et al.	17
US20180318323A1	Compositions and methods for improving gut health	Plexus Worldwide LLC	17

Table 2.5 Highlighted patents relating to diet, gut microbiome, and human health.

With the development of computational techniques to analyze larger datasets, and more research on the relationship of the microbiome and the host homeostasis and disease, patent applications related to gut microbiome and diet have subsequently extended to other health conditions beyond obesity and weight control. Among the newest patent applications related to the gut microbiome and diet is a patent describing the characterization, diagnostics, and treatment of a locomotor system condition based on microbiome data (US20170372027A1). Other applications include metagenomic methods specific for the comparison of healthy individuals and those with gut dysbiosis (WO2017216820A1), diagnostic tools for Crohn's disease, inflammatory bowel disease, irritable bowel syndrome, ulcerative colitis, and celiac disease using microbiome and other types of data (US20170286620A1), and devices such as capsules to acquire and monitor microbiome and metabolites in the gut (US20170281091A1). Research on the Gut-Brain axis relationship also resulted in several applications aiming at monitoring and manipulating the gut microbiome to enhance cognition or treat mental-health conditions (WO2017171563A1, WO2017160711A1). A recent and thorough review of patents related to the microbiome identified cancer diagnosis and treatment and CRISPR technology as recent trends in the field (247). **Table 2.5** shows a summary of highlighted patents relating to DGMH.

Even though there is already a considerable number of patent applications for technologies aiming to manipulate the gut microbiome for multiple health conditions, regulatory legislation has not yet become specific to deal with the new scientific advances in the field. In Europe, the European Food Safety Authority (EFSA) is responsible for regulating and approving food products with health claims, including probiotics, while in the U.S., the Food and Drug Administration (FDA) assumes a similar role. Legislation and regulatory aspects are changing in an attempt to keep up with the ever-evolving field. Recently, the FDA has released a statement (FDA 2018) clarifying

existing regulations and announcing the intention to work closely with the US National Institutes of Health to ensure public safety. Currently, there is no probiotic approved to be marketed in the US as a live biotherapeutic product, defined by the agency as a "biological product other than a vaccine that contains live organisms used to prevent or treat a disease or condition in humans" (FDA 2016, FDA 2018). This means that, even though probiotics are legally available as dietary supplements or food ingredients, they cannot yet have claims to cure, treat, or prevent any diseases per current regulation (FDA 2018), since those claims are reserved for drugs. Classification of food ingredients targeting the microbiome, but not composed of living organisms, microbiota-directed foods or MDFs, prebiotics and dietary fiber, is also challenging based on the available legislation. Depending on the health claims, such products can fall under the categories of drugs or dietary supplements, which have different requirements for approval (248)

2.5 Conclusion

Significant advances in microbiology, genomics, analytical chemistry, computational science, bioinformatics, and other critical disciplines have begun to converge such that it is possible to foresee a new era of health and nutrition research enabling the design of food products capable of optimizing health via predictable interactions with the gut microbiome. Despite the exciting potential in this context demonstrated by pioneering research efforts of many investigators, including those cited in this brief review, the complexity of the microbiome, the chemical composition of food, and their interplay *in situ* remains as a daunting challenge in the context of achieving needed breakthroughs. However, recent advances in high-throughput sequencing and metabolomics profiling, compositional analysis of food, and the emergence of electronic health records as an opportunity to integrate health information provide new sources of data that can

contribute to addressing this challenge. Indeed, it is now clear that computational science will play an essential role in this effort as it will provide the foundation to integrate these data layers and derive insights capable of revealing and understanding the complex interactions between diet, microbiome, and health.

The human microbiome is exceptionally plastic, which presents both challenges and opportunities (234). Due to its temporal and inter-individual variability, it is difficult to discover statistically significant signatures that unambiguously constitute a healthy versus non-healthy microbiota. At the same time, its potential for adaptation to diet and other environmental factors makes the gut microbiome an excellent target for diet-related interventions to improve health. In this article, we presented a brief overview of the current state of knowledge and potential avenues for research at the interface of diet, gut microbiome, and human health, with particular emphasis on the role that computational science and data analytics can play in accelerating this research. Using these tools, we envision a future in which diets, as well as food and dietary supplement products, can be better designed for specific populations; and in some cases for individuals; in order to optimize gut microbiota and health via a platform integrating two distinct systems. The first system will be responsible for identifying the optimal target microbiota (*discovery*) given the desired target, individual and environment, while the second will provide recommendations for achieving that target microbiota (*engineering*). Recognizing this distinction and the requirement for seamless interaction between the two can reinforce collaborative research in this evolving field where some teams focus on microbiota discovery and others on diet engineering.

Microbiome research has attracted much interest in the past few years and given rise to various software tools and pipelines for metagenomic data processing and analysis. Many of these tools address similar problems and researchers may choose a variety of tools depending on the context.

Interestingly, recent research has shown that synthetic datasets can be used to assess the performance of competing tools given a project's assumptions and hence provide useful benchmarks (249,250). We believe further progress in simulation-based studies, can give rise to new data processing and analytics pipelines customized for each project based on factors such as sequencing technology, data availability, dimensionality and variability. This can help to build standard protocols for addressing challenges like the ones mentioned in **Table 2.3** and **Table 2.4**.

Our current knowledge about the relationship between diet, gut microbiome, and human health is evolving fast. Many data analysis methods exist for discovering characteristics that can define a healthy microbiota and the factors influencing it. We believe that proper integration of recommendation systems with existing research developments will have an unprecedented impact on our way of life. Given the accelerated pace of advances in sequencing and computational tools, we expect the next decade to be the era of computational nutrition that will revolutionize our relationship with food and diet.

Alpha Diversity. A measure that represents the diversity of species at a particular site (e.g. human intestine).

Beta Diversity. A measure that represents the difference in community composition between sites (e.g. healthy vs. malignant intestine).

Classification. Supervised learning tasks in which the dependent variables are categorical.

Cluster analysis. Unsupervised learning methodology to identify groups of similar data-points automatically.

Collaborative filtering. Recommendation system methodology in which relies on similarities amongst user preferences for new recommendations.

Compositional quantities. Quantities that are described as proportions or probabilities with constant or irrelevant sum.

Content-based filtering. Recommendation system methodology in which recommendations are made based on item's content (often characterized using item features) and user features.

Curse of dimensionality. Attributed to the phenomena that increase in data dimensionality exacerbates data analysis challenges such as overfitting, the required number of samples, memory, and runtime.

Data imputation. The process of replacing missing data with substituted values.

Diversity metric. Quantitative measure that represents the number of unique entity types (e.g. species) in a community and evenness in their relative population.

Dimensionality. Number of attributes available for each sample in a given dataset. A dataset with relatively few attributes is considered *low-dimensiona*l while a dataset with many attributes is referred to as *high-dimensional*.

Labeled/unlabeled samples. Samples that have been tagged using particular labels describing the value of a dependent variable are called *labeled*. This is in contrast to *unlabeled* samples for which such labels are unavailable. Note that labels many categorical or numerical.

Marker gene sequencing. Primer-based strategy (such as 16S rRNA) that targets a specific region of a gene of interest to characterize microbial phylogenies of a sample.

Multiple-hypothesis testing. A problem that arises in tests of statistical significance when applied multiple times using different hypotheses.

Overfitting. A problem that arises in machine learning where parameter values of a model are too closely fit for training data and therefore not useful in practice.

Rarefaction. A bias correction technique used to enable comparison of diversity measures between communities with unequal sample sizes.

Recommendation system. "Any system that guides a user in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output" (267)

Regression. Supervised learning tasks in which the dependent variables are numerical.

Regularization. Machine learning technique that dampens variability of model parameters leading to a more smooth model. It is usually used to mitigate overfitting.

Stability metric. Quantitative measure to assess whether properties of a community (e.g. gut microbes) are preserved over time.

Supervised learning. Learning tasks that require labeled data. They involve learning a function to predict the correct label for a new sample given input attributes.

Unsupervised learning. Learning tasks that do not rely on labeled data. They involve learning hidden structures, features or patterns within the data.

Variation analysis. Statistical methods such as analysis of variance (ANOVA) used to identify the amount of variance in a dependent variable which can be explained using independent variables.

Whole metagenomic sequencing. Sequencing the whole genome of all microbial species within a sample. This is also called shotgun metagenomics.

Table 2.6 Glossary of terms.

Chapter 3: Microbiome-based diet optimization for irritable bowel syndrome

3.1 Abstract

Objective. Identification of microbiota-based biomarkers as predictors of low-FODMAP diet response and design of a diet recommendation strategy for IBS patients.

Design. We created a compendium of gut microbiome and disease severity data before and after a low-FODMAP diet treatment from published studies followed by unified data processing, statistical analysis and predictive modeling. We employed data-driven methods that solely rely on the compendium data, as well as hypothesis-driven methods that focus on methane and short chain fatty acid (SCFA) metabolism pathways that were implicated in the disease etiology.

Results. The patient's response to a low-FODMAP diet was predictable using their pre-diet fecal samples with F1 accuracy scores of 0.750 and 0.875 achieved through data-driven and hypothesis-driven predictors, respectively. The fecal microbiome of patients with high response had higher abundance of methane and SCFA metabolism pathways compared to patients with no response (p-values $< 6\times10^{-3}$). The genera *Ruminococcus 1*, *Ruminococcaceae UCG-002* and *Anaerostipes* can be used as predictive biomarkers of diet response. Furthermore, the low-FODMAP diet followers were identifiable given their microbiome data (F1-score of 0.656).

Conclusion. Our integrated data analysis results argue that there are two types of patients, those with high colonic methane and SCFA production, who will respond well on a low-FODMAP diet, and all others, who would benefit a dietary supplementation containing butyrate and propionate,

as well as probiotics with SCFA-producing bacteria, such as *lactobacillus*. This work demonstrates how data integration can lead to novel discoveries and paves the way towards personalized diet recommendations for IBS.

3.2 Introduction

Irritable bowel syndrome (IBS) is a chronic gastrointestinal disorder that is prevalent in approximately 11% of adult population (251). It is associated with abdominal pain and changes in stool form and frequency of bowel movements (251,252). One of the emerging treatments for IBS is to reduce the amount of fermentable oligosaccharides, disaccharides, monosaccharides and polyols (FODMAPs) in the diet, also called the low-FODMAP diet, as recommended by the American College of Gastroenterology (253) and the Canadian Association of Gastroenterology (254). The low-FODMAP diet has been effective for 50%-80% of IBS patients (255), however the patients who will benefit from this diet cannot be accurately identified beforehand. Several studies have attempted to create predictors for the efficacy of this diet in IBS using pre-treatment samples (256–258), however there is no evidence to show the utility of such a predictor across multiple studies. Furthermore, there is no common theory to explain the reason why the low-FODMAP diet is only effective for some patients in terms of disease etiology that is supported by data form multiple studies. It is believed that a low-FODMAP diet works by reducing the amount of carbohydrates that are not digested by the small intestine hence reach the colon to be used in gas producing microbial fermentation (259).

Here, we investigate whether the efficacy of low-FODMAP diets on IBS patients can be predicted by analysis of easy to obtain biomarkers. Towards this goal, we created a compendium of microbiota metagenomics, by integrating data from 6 sources and fecal metagenomics samples

from 152 unique IBS patients and 37 healthy adults. In addition, we investigated whether the amount of FODMAPs in an individual's diet, can be predicted using their gut microbiome data, showcasing the potential utility of microbiome data for assessing dietary adherence.

3.3 Materials and methods

3.3.1 Data curation

We searched PubMed for studies that have collected gut microbiome data before and after a period of low-FODMAP dietary treatment in humans. We found nine such studies and only six of them provided us with both the gut microbiome data as well as the corresponding metadata that is needed for this meta-analysis (**Table 3.1**). In all studies, the microbiome data came from fecal samples, characterized by 16S rRNA, or by the GA-map™ microbiome profiling (260). In GA-map™ microbiome profiling, each fecal sample is characterized by 54 numbers each representing the signal intensity of a DNA probe. The probes were designed for detection of bacterial taxonomies for distinguishing between IBS patients and healthy controls given fecal samples. The 16S rRNA and GA-map™ were analyzed independently.

Id	Reference	Microbiome Technology	Access
1	(261)	16s rRNA	N/A
2	(262)	16s rRNA	N/A
3	(263)	16s rRNA	Granted
4	(264)	16s rRNA	Granted
5	(265)	16s rRNA	N/A
6	(256)	GA-map™	Granted
7	(266)	16s rRNA	Granted
8	(267)	16s rRNA	Granted
9	(268)	GA-map™	Granted

Table 3.1. Studies with gut microbiome data involving low-FODMAP dietary treatment. N/A: Authors did not grant access to metadata and/or raw microbiome data.

3.3.2 Metadata processing

In all studies, the severity of IBS was quantified using IBS-SSS (IBS symptom severity scale) which is a number between zero and 500 representing the overall severity of IBS symptoms in a patient. We evaluated the patient's response to the diet based on the improvement in their IBS-SSS score ($\Delta_{IBS\text{-}SSS} = IBS\text{-}SSS_{before} - IBS\text{-}SSS_{after}$) after a period of following the low-FODMAP and labeled the patient's response as "High" (i.e. $\Delta_{IBS\text{-}SSS} \geq 150$), "Low" (i.e. $22 < \Delta_{IBS\text{-}SSS} < 150$), or "No" (i.e. $\Delta_{IBS\text{-}SSS} \leq 22$). The high threshold of 150 is reasonable since the reported mean plus standard deviation of $\Delta_{IBS\text{-}SSS}$ for a placebo treatment can range from 124 to 162 (269,270), and therefore a "High" response is unlikely to be associated with a placebo effect. The low threshold of 22 was chosen to create a balance between the "No" and "High" response groups.

3.3.3 Preprocessing of 16S rRNA microbiome data

We analyzed 16s rRNA data separately for each study before integration. We used DADA2 (271) version 1.10.1 implemented in R version 3.5.2 following the package's online tutorial (benjjneb.github.io/dada2/bigdata.html). First, primer and adapter sequences were removed from each read and quality control was performed by removing 16S rRNA reads that were chimeric,

shorter than 260 bp, or had at least two expected errors. In addition, longer reads were truncated at 260 bp since read qualities decreased sharply afterward. For one dataset (264), the reverse reads were truncated at 160 bp instead due to the decrease of read qualities at lower base pairs compared to the forward reads. Next, we performed de novo sequence assembly to identify operational taxonomic units (OTUs). Then SILVA database (272) version 32 was used to identify bacterial taxonomies associated with 16S rRNA assembled sequences. Taxa that were only observed in a single sample were filtered out.

3.3.4 Functional profiling from 16S rRNA microbiome data

We imported OTU read counts of the DADA2 analysis into qiime2, searched against Greengenes (273) and filtered out OTUs that could not be matched at the 97% identity threshold as needed for PICRUSt (67). Samples with no remaining OTUs were removed if any, and predictive metagenome profiling and KEGG pathway enrichment analysis (for level L3) were performed using PICRUSt. Finally, we converted the read counts to relative abundances and transformed using centered log-ratio transform (CLR) to account for the compositionality of microbiome data (274). In the case of zero relative abundances of a given pathway, we used the minimum amongst CLR transformed values of non-zero read counts, subtracted by 10% of their standard deviations. Given that reported KEGG pathways from PICRUSt did not include specific pathway for SCFAs, we relied on fatty-acid pathway abundances instead.

3.3.5 GA-map™ microbiome data processing

We normalized the signal intensities of 54 probes from each study separately to have zero-mean and unit-variance for a given probe before integration. To estimate the relative enrichment of methane metabolism in gut microbiome, we used the AG0581 probe (designed for detection of genus Dorea). The genus Dorea has been shown previously to be negatively associated with breath

methane levels (see (275), Table 3). To estimate the enrichment of SCFA metabolism pathways in gut microbiome, we used two pairs of probes AG0686, AG1099 (designed for genus Parabacteroides) and AG1225, AG1226 (designed for genus Alistipes) as their corresponding genus have been shown to be negatively associated with fecal SCFA levels (see (266), Table S5).

3.3.6 Differential analysis and statistical validation

We used unpaired non-parametric Wilcoxon rank-sum test for identifying pathways and taxa that are differentially abundant between IBS patients with high (n=8), low (n=29), or no (n=9) response to low-FODMAP diet where degrees of freedom is equal to the number of samples used minus two (e.g. degrees of freedom for high versus no response was 8+9-2=15). The calculated p-values were one sided for hypothesis-driven statistical validations and two sided for data-driven differential analysis. We also calculate FDR-corrected p-values (i.e. q-values) in data-driven differential analysis to account for multiple hypothesis testing given the number of KEGG pathways (n=237) and genus taxa used (n=217), with thresholds of 0.15 or lower.

3.3.7 Diet response prediction

We first integrated data from multiple studies and performed dimensionality reduction using sparse principal component analysis (276,277) reducing the number of microbiome features (microbial taxa, enriched pathways or GA-map probes) to 30% of the number of profiles in the dataset. Then for a given pair of classification labels, we created random forest (RF) classifier and evaluated using leave-one-out cross-validation. We also evaluated the classification performance by iterative removal of the feature that is identified as least important by the RF classifier until only one feature remained. In all cases the areas under the precision-recall (PR) and receiver operating characteristic (ROC) curves, as well as the F1 score (the harmonic mean of precision and recall) were calculated.

3.4 Results

Figure 3.1 illustrates our data analysis methodology. A consistent data processing pipeline was applied to the curated metagenomics data enabling downstream analysis (hypothesis-driven and data-driven). The hypothesis-driven analysis was informed from the illustrated literature-based hypotheses: (a) the methane gas can inhibit intestinal motility hence contributing to stool abnormality in the form of constipation or bloating (278), (b) methanogenesis requires hydrogen and carbon dioxide that can be generated by anaerobic fermentation of undigested carbohydrates in colon (279), and (c) short-chain fatty acids (SCFAs) such as formate can also induce methanogenesis independently or in tandem with hydrogen (259,280). Therefore, in hypothesis-driven analysis we only used methane and fatty acid metabolism pathway abundances as input while in data-driven analysis all pathways and taxa (at genus level) were used for differential abundance analysis and predictive modelling.

3.4.1 Comparison of high/low response to Low-FODMAP diet reveals structural differences in the microbiota

Pre-diet fecal metagenomes of IBS patients were integrated and processed from five studies along with disease severity scores (IBS-SSS) ranging from zero to 500 before and after following a low-FODMAP for a total of 152 patients (**Figure 3.2 A**). For differential analysis, we focused on the patients with most extreme responses (high versus no response) that had 16S rRNA metagenomic profiles (n=17). Top 5 KEGG pathways were differentially abundant with q-values < 0.11 with fatty acid metabolism being the most differentially expressed. However, there was no differentially abundant genus taxa when a q-value significance threshold of 0.15 is used (**Figure 3.2 B-C**). Three genera (*Ruminococcaceae* UCG-002**,** *Ruminococcus* 1 and *Anaerostipes*) were identified amongst the top 5 to be positively associated with stool SCFA levels based on other studies.(266,281)

Therefore a 3-genus microbiome biomarker was designed by adding their CLR-transformed abundances providing higher values for patients with a high response versus low response (p-value = 1.0×10^{-10}) or no response (p-value = 2.5×10^{-4}) following the diet. Note that the microbiome profiles of patients with low response were never used in the discovery of top five genera reported in **Figure 3.2 C**. A data-driven predictor of high/no response was built given all KEGG pathway abundances providing an F1 score of 0.750, AUROC of 0.708 (baseline: 0.5) and AUPR of 0.629 (baseline: 0.471).

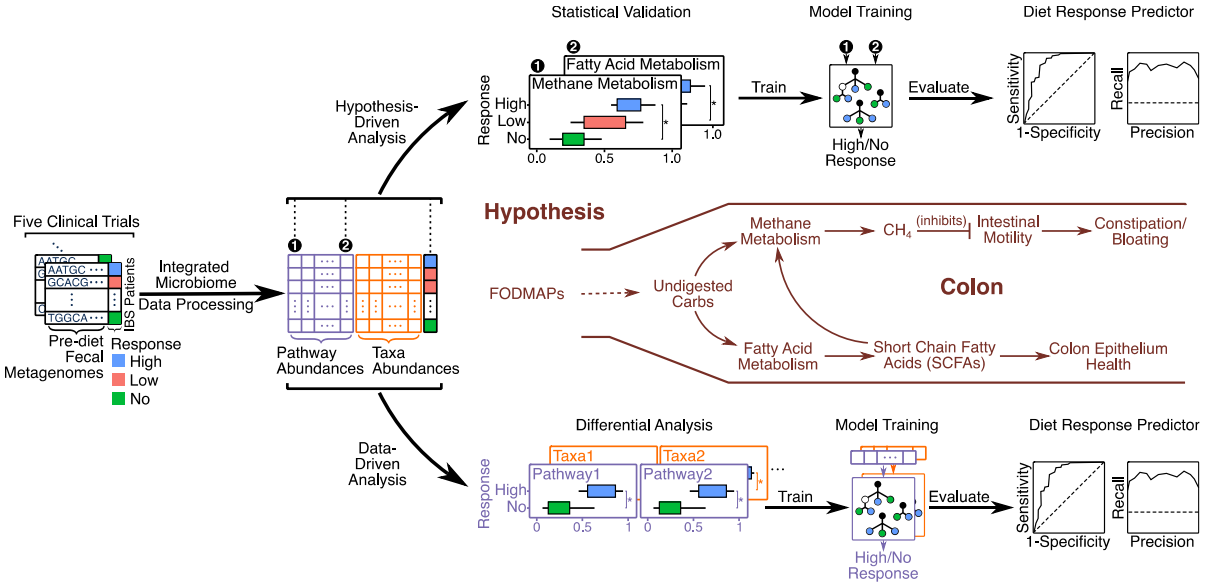

Figure 3.1. Overview of low-FODMAP diet response prediction for irritable bowel syndrome (IBS): The response of IBS patients to a low-FODMAP diet and their pre-diet fecal metagenomes were integrated and analyzed from five independent studies. Consistent data processing pipeline was applied on raw metagenome data to infer the relative pathway and taxa (at genus level) abundances for individual gut microbiomes. In a data-driven analysis, differentially abundant taxa and pathways were identified for patients with high versus no response to the low-FODMAP die. Diet response predictors were built to identify whether an IBS patient will benefit from a low-FODMAP diet given their pre-diet fecal metagenome. Furthermore, a hypothesis-driven analysis was performed given the hypothesized relationships between FODMAPs, methane metabolism, fatty acid metabolism and illustrated colon functions base on literature. Although similar to the data-driven analysis, only the pathway abundances relating to methane and fatty acid metabolism were used for statistical validation, model training and the final diet response predictor.

We also created predictor for high versus low or no response for patients with 16S rRNA metagenome profiles (**Figure 3.3 A, B, D** and **D**). Using pathway abundances as input provides an

F1 score of 0.625, AUROC of 0.850 (baseline: 0.5) and AUPR of 0.693 (baseline: 0.174) while with genus taxa abundances as input an F1 score of 0.533, AUROC of 0.873 (baseline: 0.5) and AUPR of 0.425 (baseline: 0.174) was achieved. For patients with GA-map data (**Figure 3.3 C** and **F**) an F1 score of 0.581, AUROC of 0.625 (baseline: 0.5) and AUPR of 0.530 (baseline: 0.462) was achieved.

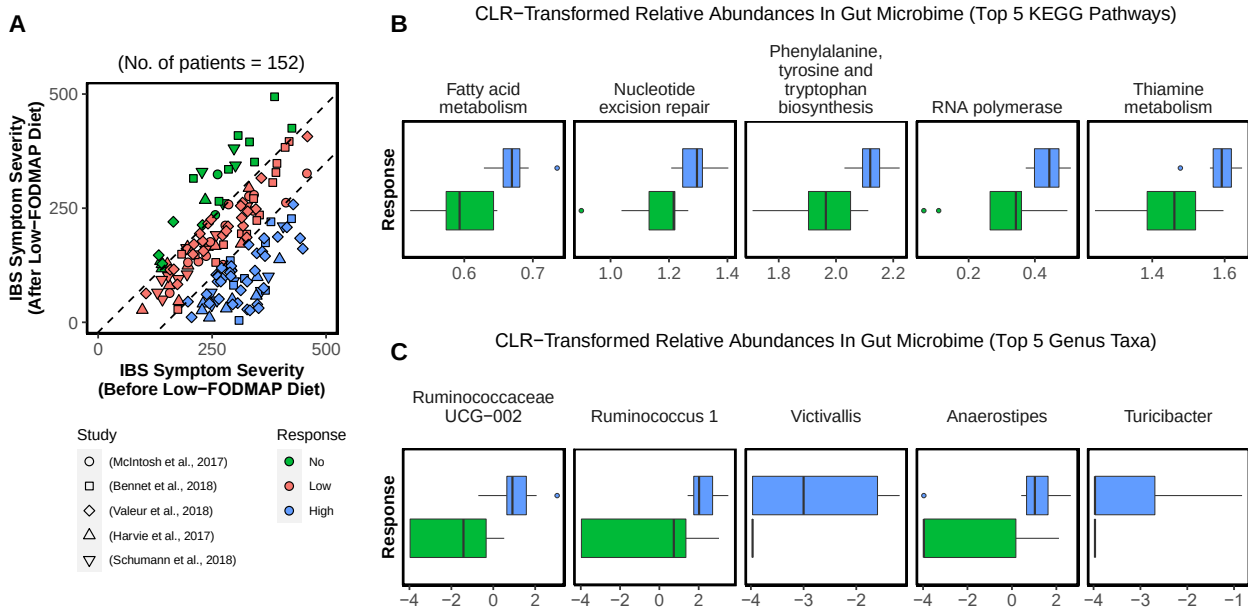

Figure 3.2. Pre-diet microbial differential abundances for IBS patients with high versus no response to the low-FODMAP diet: (A) IBS patient records from five studies are sorted into three groups based on their response to the low-FODMAP diet (High/Low/No) given the amount of improvement in IBS symptom severity after following the diet. (B) Top 5 pre-diet gut microbiome KEGG pathways that are differentially abundant (following a clr-transformation of their relative abundances) amongst High versus No response patient groups (q-values < 0.11; Fatty acid metabolism p-value = 1.5×10^{-3}; Nucleotide excision repair p-value = 3.7×10^{-3}; Phenylalanine, tyrosine and tryptophan biosynthesis p-value = 3.7×10^{-3}; RNA polymerase p-value = 3.7×10^{-3}; Thiamine metabolism p-value = 3.7×10^{-3}). (C) Similar to (B) for differentially abundant genus taxa (genus related q-values are not significant using a threshold of 0.15; Ruminococcaceae UCG-002 p-value = 3.1×10^{-3}; Ruminococcus 1 p-value = 1.3×10^{-2}; Victivallis p-value = 2.3×10^{-2}; Anaerostipes p-value = 3.0×10^{-2}; Turicibacter p-value = 6.0×10^{-2}).

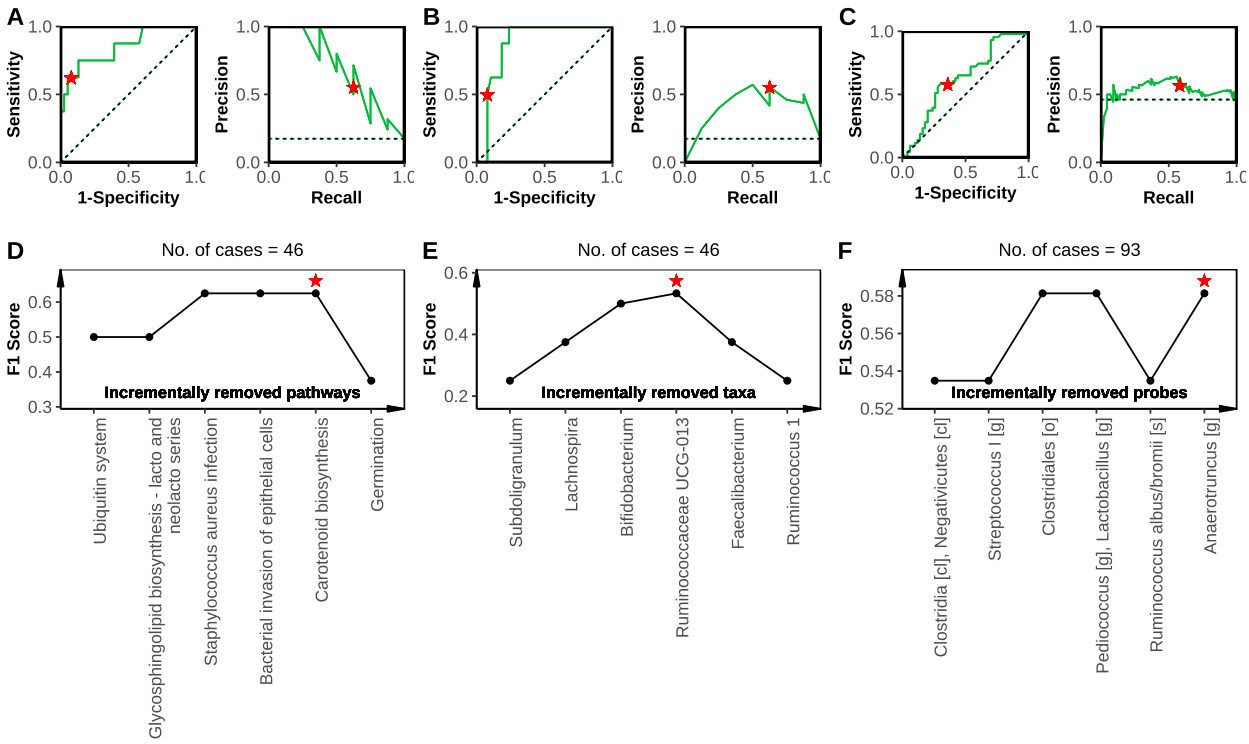

Figure 3.3 Prediction of response to low-FODMAP diet given pre-diet microbiome data: (A-C) ROC and PR curves for prediction of response to low-FODMAP diet using pathway abundances, genus taxa abundances and GA-map probe signals of pre-diet gut microbiome. The star relates to the threshold used for calculating the F1 scores. (D-F) The F1 scores relating to predictive models when the least important feature (pathway, taxa or GA-map probe) is incrementally removed until only a single feature remains in the predictive model. The stars highlight the best F1 score achieved and each corresponds to a pair of ROC and PR pair curves on the top (i.e. A&D, B&E and C&F correspond respectively).

3.4.2 IBS patients with methanogenic fecal microbiome respond better to Low-FODMAP diets

Low intestinal motility of IBS patients has been associated with intestinal production of methane (278) due to methane producing microbes (methanogens) in the gut (266,282), which use undigested carbohydrates for their metabolism (259). Therefore, we hypothesized that response to low-FODMAP diet is associated with gut microbiome methane metabolism capability. To validate this hypothesis, we performed meta-analysis on 46 patients; integrated from three studies (263,264,267) that rely on 16S rRNA data. In agreement with our hypothesis, the high response group of patients had a significantly higher enrichment in methane metabolism pathway of their

pre-treatment microbiome samples compared to low response (p-value =1.3×10^{-2}) and no response (p-value = 5.6×10^{-3}) groups (**Figure 3.4 A)**. We then used GA-map microbiome data from a separate study (256) with 31 IBS patients, using only the probe associated with methane production. The analysis of GA-map data also supports our hypothesis with high response patients having higher abundance in methane production associated taxa when compared to the no response patients (p-value = 7.4×10^{-3}).

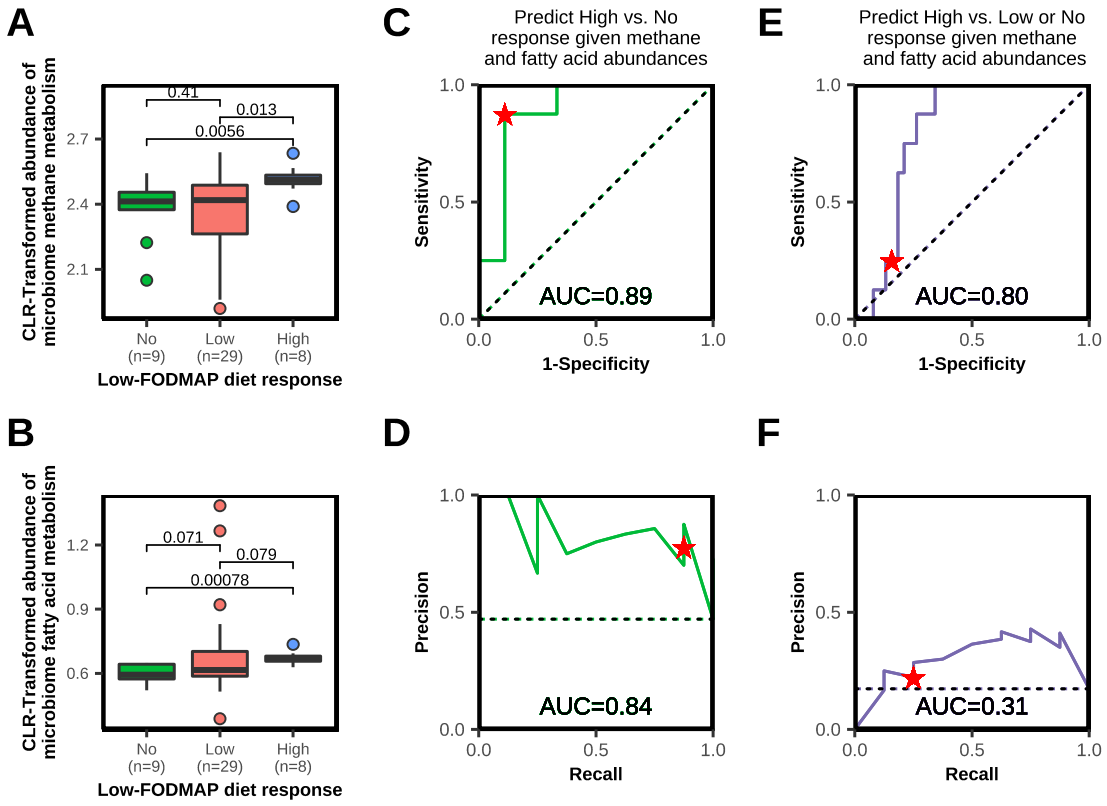

Figure 3.4. Prediction of response to low-FODMAP diet given pre-diet microbial abundances for methane and fatty acid metabolism pathways. (A&B) Methane and fatty acid metabolism pathway enrichment of pre-treatment gut microbiome for patients with High, Low or No response to low-FODMAP diet. (C&D) ROC and PR curves for predicting High vs. No response to low-FODMAP diet using methane and fatty acid metabolism pathway abundances (CLR-transformed) in gut microbiome. (E&F) ROC and PR curves for predicting High vs. Low or No response to low-FODMAP diet using methane and fatty acid metabolism pathway abundances (CLR-transformed) in gut microbiome.

3.4.3 The efficacy of Low-FODMAP diet can be accurately predicted by methane and short-chain fatty acid metabolic pathways

Short-chain fatty acids (SCFAs) are key products of microbial fermentation in human intestine and important for health of epithelial cells (283). Therefore, we also analyzed the enrichment of fatty-acid metabolism pathway in 16S rRNA fecal microbiome data of IBS patients. Our analysis shows higher enrichment in fatty-acid metabolism for high versus no response patients (p-value = 7.8×10^{-4}) (**Figure 3.4 B**). Next, we created a classifier to predict the patient's response (high versus no response) based on methane and fatty acid metabolism in 16s rRNA data. Our random forest (RF) classifier achieved 0.89 and 0.84 for area under the curve (AUC) of ROC and PR curves, respectively (**Figure 3.4 C-D**). We also performed analysis for GA-map probe data using taxa probes that have been associated with SCFA levels in fecal samples, but did not find a significant difference between the "High-response" and "Low-response" IBS patients.

3.4.4 Predicting diets from their effect on the microbiome

Diet is considered to be an important factor for modulating intestinal microbiota (284), however it is not clear whether a low-FODMAP diet leads into common changes in gut microbiome across different individuals. To investigate this, we used 188 16S rRNA fecal microbiome profiles from IBS patients and healthy individuals before (n=95) and after (n=93) low-FODMAP dietary intervention. Microbiome samples were characterized by their KEGG pathway and genus taxa abundances. We used random forest classifier to predict whether the microbiome sample is taken before, or after the low-FODMAP dietary intervention (**Figure 3.5 A**). When pathway abundances were used as input the classifier achieved F1 score of 0.656, AUROC of 0.687 (baseline: 0.5) and AUPR of 0.663 (baseline: 0.495) (**Figure 3.5 B**). Only three pathways were needed to achieve an F1 score of 0.66 (**Figure 3.5 D**). Using taxa abundances at genus level for classification provided

F1 score of 0.602, AUROC of 0.608 (baseline: 0.5) and AUPR of 0.597 (baseline: 0.495) (**Figure 3.5 C**).

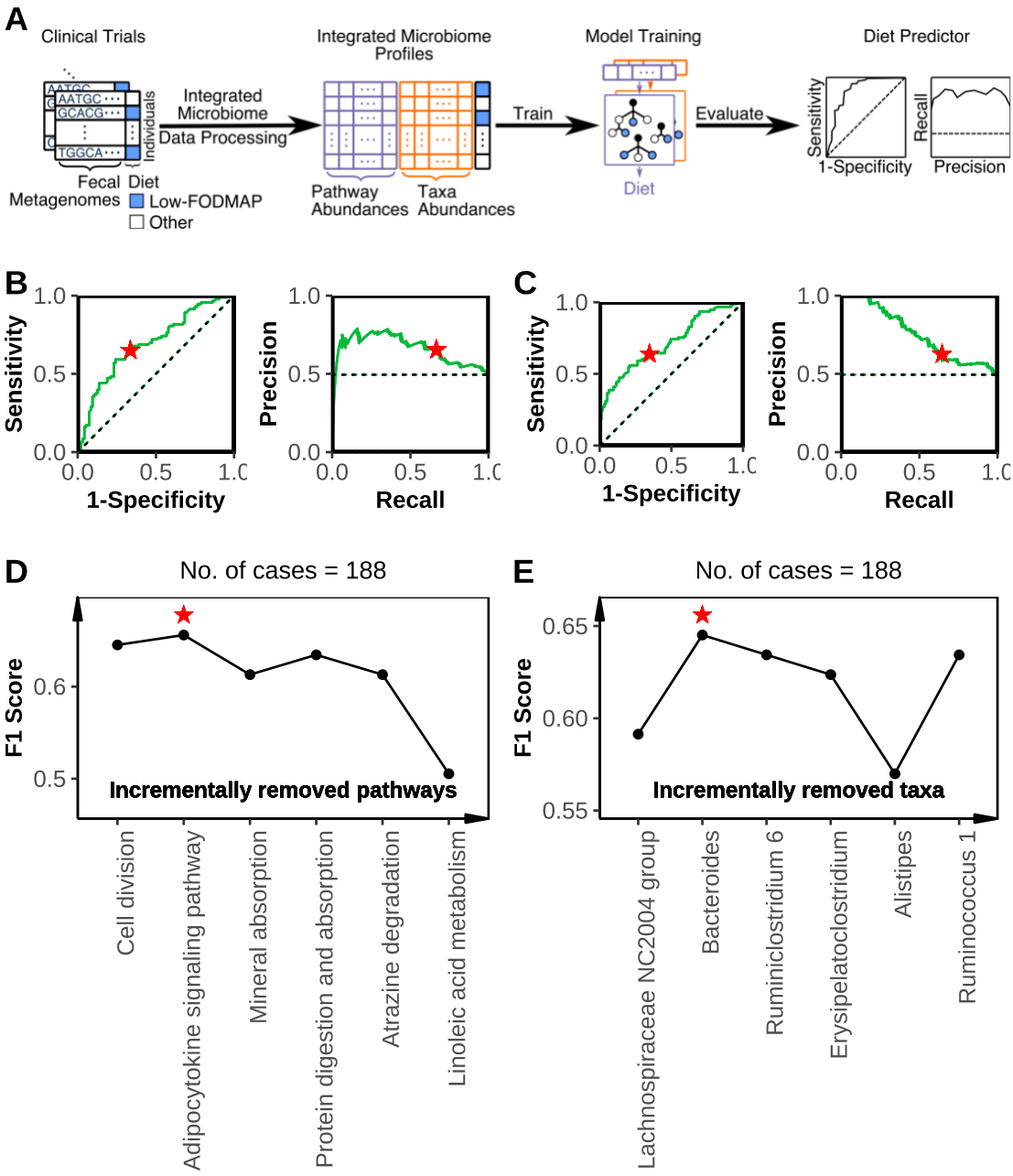

Figure 3.5. Prediction of diet (low-FODMAP vs. other) given microbiome data: (**A**) Fecal metagenomes were integrated from four studies along with the dietary regimen that was followed prior to sampling. Consistent data processing pipeline was applied on raw metagenome data to infer the relative pathway and taxa (at genus level) abundances for each sample. Diet predictors were built to identify the individual's diet given their fecal metagenome. (**B** &**C**) ROC and PR curves for diet prediction using pathway and genus taxa abundances in gut microbiome. The star relates to the threshold used for calculating the F1 scores. (**D**-**E**) The F1 scores relating to predictive models when the least important feature (pathway or taxa) is incrementally removed until only a single feature remains in the predictive model. The stars highlight the best F1 score achieved and each corresponds to a pair of ROC and PR pair curves on the top (i.e. **B**&**D** and **C**&**E** correspond respectively).

3.5 Discussion

While several studies have confirmed the efficacy of low-FODMAP diet for symptom management in IBS, between 55%-66% of IBS patients have a response that is similar to a placebo treatment. We hypothesized that the patient's response level (high/low/no) to a low-FODMAP diet can be predicted using their fecal microbiome samples. Although this hypothesis had been validated to an extent by individual studies, there is no predictor that (a) works across multiple studies and (b) comes with a mechanistic explanation of the patient's response based on their microbiomes. To this end we integrated data from five distinct studies and performed a meta-analysis showing that the patient's response to low-FODMAP diet is predictable given their fecal microbiome. We also formed a literature-based hypothesis supported by the integrated data that a high response to low-FODMAP diet is associated with higher abundance of methane and SCFA metabolism pathways in gut microbiome. Our mechanistic explanation is that a low-FODMAP diet works by lowering the amount of colonic methane that is shown to slow down intestinal motility (278), a precursor to constipation and/or bloating. Therefore, patients with highest response have a colonic microbiome with substantial methane production capability due to (a) methane metabolism pathways, and (b) SCFA metabolism pathways that promote methanogenesis, both of which rely on microbial digestion of carbohydrates. Gut microbes can also use formate or hydrogen to produce acetate (285), an SCFA with anti-inflammatory properties (286), which may inhibit their availability for methanogenesis and decrease bloating. The microbiome SCFA pathways can have positive or negative impact on microbial secretion and absorption of gases, which necessitates more in-depth investigation of their role in IBS dietary treatments (e.g. low-FODMAP diet and probiotics). Additionally, we showed that gut microbiome data can be used to predict whether a patient is following a low-FODMAP diet, suggesting that this diet modulates gut

microbiome and leaves identifiable traces which can be used for assessing dietary compliance. This work showcases the utility of integrated meta-analysis using raw data from individual studies with a consistent methodology to arrive at new insights. Although there were several differences amongst the low-FODMAP studies that can create risks for data analysis, we found no significant change in the amount of improvement of IBS-SSS score after following a low-FODMAP diet amongst the studies despite their differences. In addition, when it comes to microbiome data processing and analysis, we minimized the impact of such differences by applying the same standard pipeline starting from the raw microbiome data of each study. We acknowledge that the other differences (e.g. stool sample handling and metagenomic sequencing) can also be problematic in revealing any signal, however once such pattern is discovered, these differences increase the robustness and reproducibility of the analysis, as it becomes less sensitive to the specific details of the techniques used.

Prior studies show that lower abundance of microorganisms that produce butyrate (an important SCFA) is associated with irritable bowel syndrome (287), *Lactobacillus* based probiotics promote production of SCFAs in the gut (288) and improve disease symptoms in IBS (289). Consistent with our meta-analysis results, we suggest a biomarker-based diet recommendation system where a low-FODMAP diet is recommended to patients with high colonic methane and SCFA production, and a probiotic supplementation with SCFA producing microbes is recommended to patients with low colonic methane and SCFA production. Such a personalized recommendation system will be inline with dietary recommendations from the American College of Gastroenterology and the Canadian Association of Gastroenterology for IBS which consider both dietary treatments as beneficial (253,254), while expected to decrease the array of treatments that patients need to try before finding the treatment that works for them. Clinical trials will be

necessary to identify best biomarkers, probiotic species and dosages and evaluate the patient's response compared to alternative treatments. A comprehensive array of tests including gas analysis of breath samples, shotgun metagenomics, qPCR with primers that can detect SCFA producing microbiomes and methanogenic microorganisms that are archaeal, and gas chromatography–mass spectrometry (GC/MS) for detecting SCFA levels from microbiome samples (fecal or through colonic biopsy), will be necessary to provide more accurate insight into the microbiome pathways discussed. Given the advent of low-cost breath testing and accessibility of primer-based qPCR testing of fecal samples, gut microbiome methane and SCFA metabolism levels can be readily assessed in the clinic in order to provide more effective dietary recommendations for IBS patients. Intestinal bacterial infections are commonly diagnosed through low-cost qPCR testing of stool samples for detection of known pathogens given target-specific primers (290). Intestinal malabsorption of carbohydrates is also diagnosed in the clinic using hydrogen and methane breath testing although with variable repeatability (291). Upon development of a qPCR kit for gut microbiome SCFA metabolism estimation (e.g. by detection of *Ruminococcus 1*, *Ruminococcaceae UCG-002* and *Anaerostipes* genera levels), a personalized IBS diet can be employed in the clinic where SCFA supplementation (prebiotic or postbiotic) is recommended when SCFA microbiome metabolism is low, and a low-FODMAP diet is recommended when SCFA and methane metabolism of the gut microbiome are above a calibrated threshold. We believe that the recent advances in high resolution omics and computational methods across diet, microbiome, and health (1), as well as novel ways of food representation that rely on artificial intelligence (292,293), will give rise to more personalized dietary treatments potentially revolutionizing clinical nutrition.

It is important to note that, the analyzed data here included microbiome profiles from IBS patients with diarrhea, constipation, or both symptoms, however, we did not perform a separate analysis based on the IBS type since multiple studies did not provide the IBS type information per patient. Further studies will be necessary to validate the hypothesized mode of action for this diet in reducing constipation and bloating symptoms of IBS, and to understand the possibly different modes of action in reducing diarrhea.

Chapter 4: Genetic neural networks for modeling biological systems

4.1 Abstract

Motivation: Gene expression prediction is one of the grand challenges in computational biology. The availability of transcriptomics data combined with recent advances in artificial neural networks provide an unprecedented opportunity to create predictive models of gene expression with far reaching applications.

Results: We present the Genetic Neural Network (GNN), an artificial neural network for predicting genome-wide gene expression given gene knockouts and master regulator perturbations. In its core, the GNN maps existing gene regulatory information in its architecture, and it uses cell nodes that have been specifically designed to capture the dependencies and non-linear dynamics that exist in gene networks. These two key features make the GNN architecture capable to capture complex relationships without the need of large training datasets. As a result, GNNs were 40% more accurate on average than competing architectures (MLP, RNN, BiRNN) when compared on hundreds of curated and inferred transcription modules. Our results argue that GNNs can become the architecture of choice when building predictors of gene expression from exponentially growing corpus of genome-wide transcriptomics data.

Availability and implementation: https://github.com/IBPA/GNN

4.2 Introduction

Prediction of cellular state in novel environments presents a need and an opportunity in systems biology (294–296). Surge in the availability of data, advances in computational techniques and exponential increase of computing power have led to the adoption of omics analysis and predictive modeling in a variety of fields, including food safety, drug discovery, biofuel development and precision medicine (297–300). The key role of gene expression (GE) in cellular machinery (294) and the cost-effective nature of high-throughput transcriptomics have renewed interest in predictors of gene expression as a proxy of the cellular state (301,302). If successful, an accurate predictive GE model can be useful in basic research on understanding how gene expression changes based on environmental stimuli, and in industrial biotechnology by guiding wetlab experimentation to those settings that are more likely to produce the desired results, from recombinant protein expression (303,304) to strain engineering (305) and drug production (306).

When it comes to prediction, artificial neural networks (ANN) outperform other methods in areas such as computer vision and machine translation (307). Despite their success in complex prediction tasks (308,309), their application for steady state GE prediction has been quite limited (310). Instead most researchers rely on methods based on linear models, molecular thermodynamics, differential equations, logical circuits and Bayesian networks (302,311). The idea of using ANNs for GE prediction is not novel (312) but its adoption has remained stale due to lack of data and limited predictive power of the algorithms so far. Recently, ground-breaking ideas around deep neural networks (DNN) accompanied with the availability of vast transcriptomics repositories have created an unprecedented opportunity to create accurate predictors for genome-wide expression (313). For example, a recurrent neural network (RNN) was employed as part of a genome-scale

model trained on twenty million data points for steady state GE prediction in novel conditions for bacterium *Escherichia coli* (295). In another study, A 3-layer feedforward neural network (FNN) was used for GE prediction when the expression of landmark genes were given (314). More recently, a convolutional neural network (CNN) called DeepChrome was used to predict GE from histone modifications (315) and a similar tool, DeepPep was developed for predicting protein occurrence in proteomics samples (316).

There are several technical challenges to overcome when building an ANN GE model. First, one has to optimize the ANN hyper-parameters that determine the underlying ANN architecture. Although general architectures can be trained, architectures that are tailored to incorporate the key properties of a given problem tend to be substantially more accurate. For example CNNs are designed to excel in image tasks (317) based on the idea that high level features in an image can be identified by hierarchical combination of local features (e.g. for a face to be identified, two eyes, nose and mouth would be identified nearby each other). Recently, a new type of ANN called visual interaction network is developed to capture dynamics of physical objects (e.g. billiard balls) from video frames in order to predict object's physical trajectory (318). Schema networks take this idea further by capturing causalities between object dynamics as well to enhance prediction in transfer learning (319). Currently, there is no ANN architecture that maps well to the gene regulatory dynamics and the complex expression signatures they produce within cells. To address this challenge, we developed a novel feed-forward architecture, coined *Genetic Neural Network*(GNN), that is founded on the observation that gene expression in prokaryotic systems is influenced, at least partially, by the expression level of its transcription factors (TF) (320,321). The fundamental building block of the GNN is the *GNN layer or cell*, a type of node that has been designed to capture the dynamics that govern gene regulation.

A second challenge in training ANNs is to produce sufficient data to avoid over-fitting. DNNs are notorious for their need of large datasets: for instance, ImageNet that is used to train computer vision DNNs contains 14 million images (322). In contrast, for the most widely studied bacterium, *Escherichia coli*, there are only 4,389 GE profiles, each with the expression of its approximately 4,500 genes, across 649 conditions (295). One of the common approaches for mitigating the data gap in ML is constrained optimization. In the GNN architecture that we introduce here, we achieve that by constraining the connectivity of the GNN layers based on the transcriptional regulatory network of the organism, which is (partially) known from public databases. These two features, namely the introduction of a new node type and an architecture that have been designed to mimic gene regulatory and expression dynamics, are the key innovations behind the superior performance of GNNs and the main contributions of this paper.

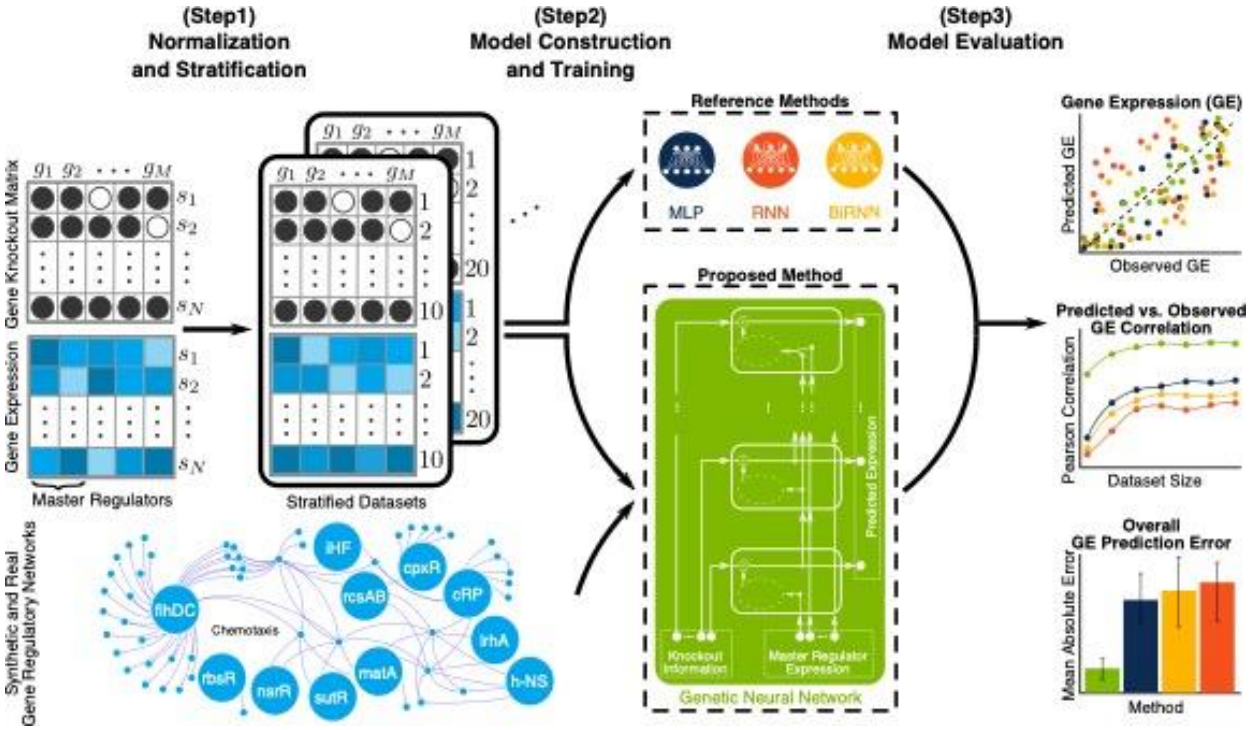

Figure 4.1 The proposed Genetic Neural Network (GNN) architecture, is evaluated against various ANN architectures and ANN types (MLPs, RNNs, BiRNNs) in its ability to predict gene expression levels given master regulator expression and knockout information. In Step 1, a compendium of normalized expression levels over a wide spectrum of conditions is created, together with the contextual gene regulatory network information (Chemotaxis pathway here, retrieved from (323)). Stratified datasets of various sizes are generated after normalization to drive Step 2, where ANN models are constructed and trained. When applicable, the model architecture is informed by known regulatory relationships. In Step 3, the methods are evaluated through 5-fold cross validation on their predictive performance on gene expression.

In this article, we focus on steady state GE prediction for small to medium size transcriptional network modules (between 2 to 1000 genes) and with the assumption that the expression of *master regulator (MR)* genes are known. Since MR genes sit on top of the regulatory hierarchy, they play a key role in transcriptional regulation. Given the causal role of MR genes on the GE profile, models that accurately predict the impact of their perturbation are important. In section 2, we describe how we define, construct and train the GNN model. In section 3 we introduce competing methods that we compare against and in section 4 we describe the results of these performance review. The overall methodology is summarized in **Figure 4.1**.

4.3 Method

4.3.1 GNN Architecture

The **input layer** of a GNN consists of the expression level of MR genes and gene knockout information (referred to as "a" and "KO" in **Figure 4.2**). The **output layer** of GNN consists of the predicted gene expression levels. Each **intermediate layer** in GNN predicts expression of a single gene, for instance, L_1, L_2 and L_3 predict the expression of gene a, b and c, respectively (**Figure 4.2**). This architecture is built under the assumption that expression of a gene with d regulators, can be estimated using the activation function $f_{\boldsymbol{\theta}}(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}_{\geq 0}^{d}$ represents the expression level of regulator genes. Therefore inputs of the activation function $f_{\boldsymbol{\theta}}$ for each gene, are made available by ensuring a topological order. For example, in **Figure 4.2 A**, the expression of gene c is regulated by gene b. Therefore designated layer of b (i.e. the L_1 cell) comes before designated layer of c (i.e. the L_2 cell) as in **Figure 4.2 B**. Note that topological order is not unique for cyclic graphs. Therefore, when a cycle is detected (here by using depth-first-search), we remove the feedback edges before generating the topological gene ordering.

The **activation function** $f_{\boldsymbol{\theta}}$ is based on the generalized logistic function that recapitulates the non-linear dynamics that govern gene expression, usually modeled through the Hill function (324). More specifically, $f_{\boldsymbol{\theta}}$ is given by:

$$f_{\theta}(\mathbf{x}) = \frac{t_0 + \sum_{k=1}^{d} t_k \, e^{p_k x_k}}{1 + \sum_{k=1}^{d} b_k \, e^{p_k x_k}} \tag{4.1}$$

where θ is the set of function parameters including the input weight vector $\mathbf{p} \in \mathbb{R}^{d}$, numerator weight vector $\mathbf{t} \in \mathbb{R}_{\geq 0}^{d}$, denominator weight vector $\mathbf{b} \in \mathbb{R}_{\geq 0}^{d}$ and bias $t_0 \in \mathbb{R}_{\geq 0}$. Assuming θ is

known for all layers, one forward pass results in predicted expression levels $\hat{y}$ for all genes. The final predictions are clamped into a valid range $[y_{min}, y_{max}]$. This will be [0,1] if data is normalized to this range. Otherwise y_{min} and y_{max} are the minimum and maximum values observed for each gene in the whole dataset. Layer-wise GNN training for θ estimation is explained in the next section.

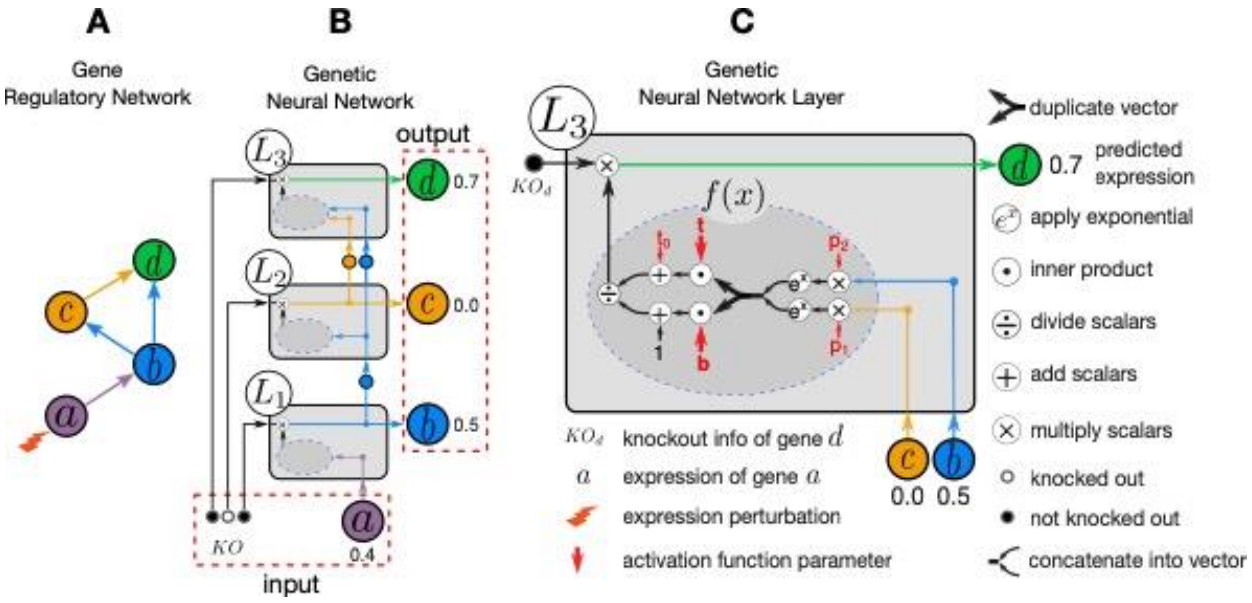

Figure 4.2 Genetic Neural Network architecture schematic for a regulatory network example. The aim is to predict gene expression levels given the expression of master regulator[s] (MR) and knockout information (KO) for other genes. (**A**) An example gene regulatory network, consisting of a single MR "a" and three other genes b, c and d in topological order. Each arrow indicates a direct regulatory relationship. (**B**) The Genetic Neural Network topology that would map the regulatory relationships of the example gene regulatory network. The input consists of the MR expression level a and the knockout vector KO. Each layer corresponds to a single gene (i.e. L_1, L_2, L_3 correspond to b, c, d, respectively). Prediction of non-MR gene expression is achieved by a forward-pass, from first layer (L_1) to last (L_3). This order ensures that for each layer the expression levels of the regulator[s] are available for the layer's forward pass. (**C**) A dissection of a layer (i.e. GNN node). It consists of the the MR gene expression levels x, the activation function f, knockout information (e.g. KO_d) and finally the output vector by appending the predicted expression (e.g. $\hat{d}$) to the initial inputs of current layer when needed by subsequent layer[s]. Although only L_3 is illustrated in detail, the general form of layers are the same while the inputs and weights vary depending on regulators and training data.

4.3.2 Layer-wise Training

A separate regression problem is defined in each layer (i.e., for each gene). For the layer corresponding to a particular gene, a corresponding dataset $C = \{X, y\}$ that consists of regulator gene expression levels X and expression levels of the current gene y is created:

$$X = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(m)} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \tag{4.2}$$

To predict $\hat{y}^{(i)} = f_\theta(\mathbf{x^{(i)}}),\ \forall i = 1 \dots m$, we devise a loss function:

$$loss(C, \boldsymbol{\theta}) = \sum_{i=1}^{m} \left[f_\theta(\mathbf{x}^{(i)}) - y^{(i)} \right]^2 \tag{4.3}$$

Hence optimal $\boldsymbol{\theta}$ can be determined by minimizing $loss(C, \boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = obj_1(C) = \underset{\boldsymbol{\theta}}{ArgMin}\ loss(C, \boldsymbol{\theta}) \tag{4.4}$$

In order to solve Equation (4.4) first we show that for a given parameter vector $\mathbf{p}$, the parameters $\mathbf{w}^* = [t_0^* | \mathbf{t}^* | \mathbf{b}^*]$ can be uniquely determined using a linear program. To see this, let us assume $\mathbf{w}^*$ is determined. Hence we have:

$$\boldsymbol{\theta}^+ = \{t_0^*, \mathbf{t}^*, \mathbf{b}^*, \mathbf{p}\} \tag{4.5}$$

where $\boldsymbol{\theta}^+$, is the set of parameters for f where $t_0^*, \mathbf{t}^*, \mathbf{b}^*$ minimize the $loss$ for a given $\mathbf{p}$. Therefore, predicted expression of each gene $\hat{y}^{(i)}$ can be calculated given the corresponding TF expressions $\mathbf{x}^{(i)}$ for $i = 1 \dots m$:

$$f_{\theta^+}(\mathbf{x}^{(i)}) = \frac{t_0^* + \sum_{k=1}^{d} t_k^* h_k^{(i)}}{1 + \sum_{k=1}^{d} b_k^* h_k^{(i)}} = \hat{y}^{(i)},\ h_k^{(i)} = e^{p_k x_k^{(i)}} \tag{4.6}$$

Assuming $\mathbf{b}^* \geqslant 0$, the denominator above will be non-zero hence we can rewrite as:

$$t_0^* + \sum_{k=1}^{d} t_k^* \, h_k^{(i)} - \hat{y}^{(i)} \sum_{k=1}^{d} b_k^* \, h_k^{(i)} = \hat{y}^{(i)} \tag{4.7}$$

Considering $\hat{y}^{(i)} \approx y^{(i)}$, we have:

$$t_0^* + \sum_{k=1}^{d} t_k^* \, h_k^{(i)} - y^{(i)} \sum_{k=1}^{d} b_k^* \, h_k^{(i)} \approx y^{(i)} \tag{4.8}$$

To convert this into matrix form, we define vector $\mathbf{w}^*$ and matrices H, Y_e and A.

$$\mathbf{w}^* = [t_0^*, t_1^*, t_2^*, \dots, t_d^*, b_1^*, b_2^*, \dots, b_d^*]^T \tag{4.9}$$

Matrix H consists of $h_k^{(i)}$ values (constant for given $\mathbf{p}$ and X):

$$H = \begin{bmatrix} h_1^{(1)} & h_2^{(1)} & \dots & h_d^{(1)} \\ h_1^{(2)} & h_2^{(2)} & \dots & h_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ h_1^{(m)} & h_2^{(m)} & \dots & h_d^{(m)} \end{bmatrix}_{m \times d} \tag{4.10}$$

Matrix Y_e consists of expression levels $y^{(i)}$ repeated d times in columns (constant for given $\mathbf{y}$):

$$Y_e = \begin{bmatrix} y^{(1)} & y^{(1)} & \dots & y^{(1)} \\ y^{(2)} & y^{(2)} & \dots & y^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y^{(m)} & y^{(m)} & \dots & y^{(m)} \end{bmatrix}_{m \times d} \tag{4.11}$$

Matrix A is calculated using H and Y_e where "$\odot$" represents entry-wise multiplication and "|" represents column-wise matrix concatenation:

$$A = [\mathbf{1}|H|(-Y_e \odot H)]_{m\times(2d+1)} \tag{4.12}$$

Therefore Equation (4.8) can be represented in matrix form:

$$A.\mathbf{w}^* \approx \mathbf{y} \tag{4.13}$$

To see this, note that for each i: (I) the inner product of $\mathbf{w}$ with i^{th} row of A corresponds to the terms on the left side of Equation (4.8) and (II) the i^{th} element of $\mathbf{y}$ corresponds to the term on right side of Equation (4.8).

To convert the approximation in Equation (4.13) to equality, we add $\boldsymbol{\epsilon}_l, \boldsymbol{\epsilon}_\mathbf{r} \in \mathbb{R}^m_{\geq 0}$ to both sides:

$$A.\mathbf{w}^* + \boldsymbol{\epsilon}_l = \mathbf{y} + \boldsymbol{\epsilon}_r \tag{4.14}$$

Therefore, the desirable $\mathbf{w}^*$ should minimize approximation error $\sum_{i=1}^{m} \boldsymbol{\epsilon}_l^{(i)} + \boldsymbol{\epsilon}_r^{(i)}$. To find $\mathbf{w}^*$, we devised obj_2:

$$\begin{aligned} \mathbf{w}^* = obj_2(C,p) \quad = \quad & \underset{w}{ArgMin} && \mathbf{1}^T\boldsymbol{\epsilon}_l + \mathbf{1}^T\boldsymbol{\epsilon}_r \\ & \text{subject to} && A.\mathbf{w}^* + \boldsymbol{\epsilon}_l - \boldsymbol{\epsilon}_r = \mathbf{y} \\ & && \boldsymbol{\epsilon}_l \geqslant 0, \boldsymbol{\epsilon}_r \geqslant 0, \mathbf{b} \geqslant 0, \mathbf{t} \geqslant 0 \end{aligned} \tag{4.15}$$

This can be transformed into standard linear programming (LP) form:

$$\begin{aligned} \mathbf{w}^* = \quad & \underset{\mathbf{z}}{ArgMin} && \mathbf{a}^T.\mathbf{z} \\ & \text{subject to} && G.\mathbf{z} = \mathbf{y} \\ & && \mathbf{z} \geq \mathbf{l} \end{aligned} \tag{4.16}$$

where

$$\begin{aligned} \mathbf{z} \quad &= \begin{bmatrix} t_0 \\ \mathbf{t} \\ \mathbf{b} \\ \boldsymbol{\epsilon}_l \\ \boldsymbol{\epsilon}_r \end{bmatrix}, \mathbf{a} = \begin{bmatrix} 0 \\ \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times 1} \\ \mathbf{1}_{m\times 1} \\ \mathbf{1}_{m\times 1} \end{bmatrix}, \mathbf{l} = \begin{bmatrix} 0 \\ \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times 1} \\ \mathbf{0}_{m\times 1} \\ \mathbf{0}_{m\times 1} \end{bmatrix}, \\ G \quad &= [A_{m\times(2d+1)} | I_{m\times m}| - I_{m\times m}] \end{aligned} \tag{4.17}$$

Therefore, for a given input coefficient $\mathbf{p}$ and gene expression values $C = \{X, \mathbf{y}\}$, the optimal $\mathbf{w}^*$ can be estimated by solving $obj_2(C, \mathbf{p})$ using the linear program in Equation (4.16). With this insight, we can solve $obj_1(C)$ (Equation (4.4)) using an iterative algorithm starting from an initial $\mathbf{p}$ vector. In each iteration, first $\mathbf{w}^*$ is estimated using $obj_2(C, \mathbf{p})$. Then $loss(C, [\mathbf{w}^*|\mathbf{p}])$ and its gradient w.r.t $\mathbf{p}$ are calculated. Finally, a new $\mathbf{p}$ is generated using the calculated $loss$ and its gradient. Although various gradient based optimization methods can be used for this iterative procedure, we used the conjugate gradient method (325). This is described in **Algorithm 1**. Line 6 refers to the first step. Lines 7 and 8 refer to second step. The last step is done by the $ConjugateGradient$ function in each iteration.

In practice we run the algorithm 10 times with different initial random values for $\mathbf{p}$ and use the one which gives the best fit (i.e. lowest value for $loss$ in Equation (4.3)). The complexity of layer-wise training algorithm is $O(m^{5.5})$.

Note that, there are two choices for matrix X values. First is to use the actual GE values from the training dataset. Second is to replace actual GE values with corresponding predicted ones when calculated from a previous layer. In our experiments the second choice provided slightly better predictive power (hence used for the presented results of section 4.5).

4.4 Competing Methods

We compare the GNN method against LASSO, a linear model with ℓ_1 regularization (326), a Multi-layered Perceptron (MLP) (327), a recurrent neural network (RNN) (328), a bi-directional neural network (BiRNN) (329) and a linear version of our GNN network (LinGNN). Recall that in our prediction task, the input vector consists of the expression level of the master regulator (MR) genes and the knockout information vector. Here, we use the vector $\mathbf{v}$ as the concatenation of all inputs, the vector $\mathbf{y}$ as the expression level of non-MR genes, and $\hat{\mathbf{y}}$ referring to their predicted values. Unlike GNN, the common ANN architectures have hyper-parameters that need to be first optimized. For our comparison, we use the hyper-parameters that correspond to the best-performing architecture (i.e. the architecture with minimum MAE) by using a traditional search method (330).

Algorithm 1. Layer-wise training algorithm to estimate activation function parameter vector $\boldsymbol{\theta} = [\boldsymbol{w}|\boldsymbol{p}]$ where p and w consist of input and exponential coefficients, respectively (see Equation (4.1)). The gene expression dataset C related to a gene, consists of X and $\boldsymbol{y}$. The matrix X contains observed TF expression levels (i.e., inputs to the activation function). The vector $\boldsymbol{y}$ contains corresponding observed GE values for this gene (i.e., activation function outputs).

Inputs: gene expression dataset $C = \{X, \boldsymbol{y}\}$
Outputs: activation function estimated parameter vector $\boldsymbol{\theta}^* = [\boldsymbol{w}|\boldsymbol{p}]$

$\boldsymbol{w} \leftarrow \mathbf{0}$ (Note: $\boldsymbol{w}$ is a global variable)
$\boldsymbol{p} \leftarrow$ random initial vector
$\boldsymbol{p} \leftarrow ConjugateGradient(GetLossP, \boldsymbol{p}, C)$
return $[\boldsymbol{w}|\boldsymbol{p}]$

Function $GetLossP(\boldsymbol{p}, C)$:
    $\boldsymbol{w} \leftarrow obj2(C, \boldsymbol{p})$ // solve LP from Equation (4.16)
    $\boldsymbol{loss\_p} \leftarrow loss(C, [\boldsymbol{w} \mid \boldsymbol{p}])$ // Equation (4.3)
    $\boldsymbol{loss\_p\_grad} \leftarrow \frac{\nabla \text{loss}(C, [\mathbf{w}|\mathbf{p}])}{\nabla \mathbf{p}}$
return $\boldsymbol{loss\_p}, \boldsymbol{loss\_p\_grad}$

Function $ConjugateGradient(f, \boldsymbol{x}, C)$:
    /* f: cost function takes variables $\boldsymbol{x}$ and C as input and returns
        the cost and the gradient with respect to $\boldsymbol{x}$. */
    /* *** Conjugate Gradient Implementation
    *** */
return $\boldsymbol{x}^*$

Multi-layer Perceptron (MLP): MLP is used by (314) for GE prediction when expression of landmark genes are known. An MLP instance here takes an *input* vector **v** and calculates the *output* vector $\hat{\mathbf{y}}$. To identify the *hyper-parameters*, we examine architectures with 0 to 3 hidden layers, 5 to 50 hidden nodes per layer with ℓ_2 regularization coefficient between 0.0 to 0.5.

Recurrent Neural Network (RNN): RNN is used by (316) for GE prediction when genetic and environmental perturbations are characterized as input. Similar to an RNN used by (316), a fully connected RNN instance here, takes a sequence of *input vectors* (331). The same vector **v** is

repeated multiple times (depending on the depth *hyper-parameter* t) as input. The *output* vector of RNN $\hat{\mathbf{y}}$ corresponds to the output of the last rollout of the RNN (only). For *hyper-parameters*, we examine architectures with depth t between 1 and 20 and ℓ_2 regularization coefficients between 0.0 to 0.5.

Bidirectional Recurrent Neural Network (BiRNN): Our BiRNN instances are set-up exactly same as our simple RNN ones except that they are bidirectional.

LASSO: Linear regression with ℓ_1 regularization (i.e. LASSO) is a widely-used regression method that improves the generalization power of the linear model by reducing the number of features through an ℓ_1 penalty in the objective function (326). In our setting, it is equivalent to an MLP with no hidden layers, identity activation function and ℓ_1 regularization. For *hyper-parameter* optimization, we examined regularization coefficient from 0.0 to 5.0.

For training competing ANN architectures and Lasso, we used RMSProp (332). The loss function used in RMSProp here is the mean squared error (MSE) plus regularization of model weights **w** as in Equation (4.18). We run RMSProp with learning rate of 0.001 until convergence. The training is stopped whenever MSE has less than 0.0001 improvement in the last 100 epochs.

$$loss_w = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda_1 \parallel \mathbf{w} \parallel + \lambda_2 \parallel \mathbf{w} \parallel_2^2 \qquad (4.18)$$

Linear GNN (LinGNN): Regulatory network connections can be incorporated into a linear model for GE prediction, given TF expression level as it has demonstrated in previous models (333,334). To evaluate the performance of a linear model with the proposed architecture, we developed LinGNN, which has the same GNN framework, but a linear function in Equation (4.19) is used for

node activation, instead of the nonlinear activation in Equation (4.1). Here $b \in \mathbb{R}$ is the bias term, $\mathbf{a} \in \mathbb{R}^d$ vector consists of additive coefficients and $M \in \mathbb{R}^{d \times d}$ consists of multiplicative coefficients.

$$f_{b,\mathbf{a},M}(\mathbf{x}) = b + \sum_{i=1}^{d} \left(a_i x_i + \sum_{j=1}^{d} M_{i,j}\, x_i x_j \right) \tag{4.19}$$

For training the LinGNN, the same layer-wise training strategy is used. However, given the linear function, we used the OLS for parameter fitting to solve Equation (4.4) (instead of **Algorithm 1.**).

4.4.1 Evaluation Metrics

To evaluate the model performance given observed GE values $\mathbf{y}$ and corresponding predicted GE values $\hat{\mathbf{y}}$, we use the Pearson Correlation Coefficient (PCC) and Mean Absolute Error (MAE) (335).

4.5 Empirical Results

We designed two sets of experiments to assess the impact of network complexity and data availability (second and third aforementioned challenges) on predictive power.

4.5.1 Network complexity impact

First, we constructed the full TRN of *E. coli* curated by RegulonDB v9.4 (323). Second, from this full TRN, we extracted 33 network modules each containing between 10 to 1000 genes. This was done using the greedy module extraction method in GeneNetWeaver software (336) also used in number of DREAM challenges (337,338). Third, for each extracted TRN module, we identified the MR genes by selecting genes that are not regulated by any other gene within that network.

Fourth, for each TRN module, we performed thousands of steady state thermodynamic simulation experiments using GeneNetWeaver with added microarray noise. Each simulation run requires kinetic parameters for each gene which are randomly initialized by the simulation software (and unknown to GNN and other predictive models). These *in silico* experiments consist of random multi-factorial perturbations for MR genes and single gene knockouts for other genes giving rise to GE dataset with thousands GE profiles. Fifth, for each GE dataset we identified 10 dissimilar GE profiles . To identify these dissimilar GE profiles in a dataset, we performed hierarchical clustering (339) with cluster size of 10 and selected one GE profile from each cluster randomly. Finally, we performed 5-fold cross-validation (CR) for the task of GE prediction given MR expression levels and knockout information in each dataset. Results in **Figure 4.3 A, C**, show that GNN method outperforms other methods on datasets with network sizes ranging between 10 and 1000 when compared based on MAE and PCC metrics in 5-fold CR setting. Our results show that the GNN has a smaller error (average MAE 0.09 ± 0.01; PCC 0.86 ± 0.03) than LinGNN (average MAE 0.11 ± 0.01; PCC 0.80 ± 0.05), Lasso (average MAE 0.19 ± 0.04; PCC 0.55 ± 0.09), MLP (average MAE 0.15 ± 0.03; PCC 0.63 ± 0.10), RNN (average MAE 0.15 ± 0.03; PCC 0.63 ± 0.08) and BiRNN (average MAE 0.15 ± 0.03; PCC 0.65 ± 0.08). **Figure 4.3 B** depicts the overall performance on all datasets.

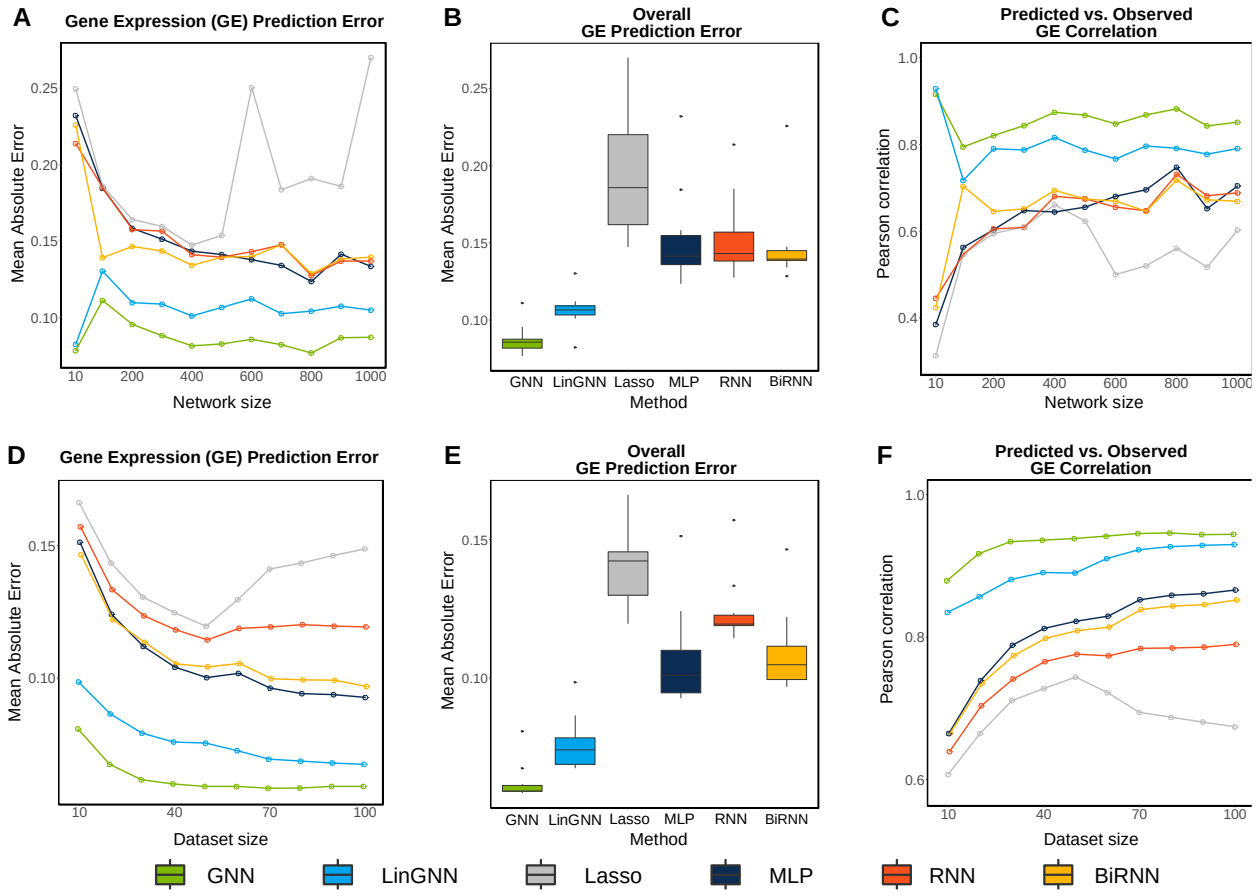

Figure 4.3 5-fold cross validation performance evaluation using data acquired through thermodynamic simulation. **A, B, C** show performance of all methods given randomly selected network modules for *E. coli* transcription network. Better performance of GNN can be seen in all cases. The BiRNN architecture outperforms other conventional ANN architectures in some cases. D, E, F show the prediction performance for TRN module of chemotaxis (61 genes) in *E. coli*. Results suggest that GNN performs in call cases particularly on smaller datasets.

4.5.2 Effect of data availability

In order to assess predictive power of various architectures based on data availability, we picked the transcription network of chemotaxis since it is one of the most well-studied signaling pathways (340).

First, we constructed the TRN of chemotaxis. To do so, we used KEGG (341) to get list of 57 genes involved in chemotaxis signal transduction pathway of bacterium *E. coli*. We then added to this list, the genes directly involved in transcription of chemotaxis genes. We then removed genes

that are not involved in any transcriptional regulation (i.e. genes that have no reported TF listed) hence 61 genes with 84 TF-Gene relationships remained. For TF-Gene relationships we used all regulatory relationships that are based on experimental evidence curated in RegulonDB v9.4. Second, we identify MR genes in the chemotaxis TRN as the list of genes with no TF within the chemotaxis TRN. Third, we performed thermodynamic simulations (with same method mentioned in 4.5.1) 100 times. Each time we used different set of network parameters (TF binding affinities, degradation rates, etc.) for the chemotaxis network using GeneNetWeaver. This gave us 100 different datasets each with thousands of GE profiles and their corresponding knockout information. Fourth, from each GE dataset we extract 10 stratified datasets with varying sizes (10 to 100 GE profiles each). To generate a stratified dataset of size K, we perform hierarchical clustering (339) with number of clusters set to K and randomly pick one profile from each cluster. Finally, to evaluate performance of GNN and competing methods we perform 5-fold CR validation for the task of GE prediction given expression level of MR genes and knockout information. Results in **Figure 4.3 D, F**, show that GNN method outperforms other methods on stratified datasets with sizes ranging between 10 and 100 profiles each when compared based on MAE and PCC metrics in 5-fold CR setting. **Figure 4.3 E** shows the overall performance on all datasets. Note that the gap between GNN and other methods is larger on smaller dataset sizes.

4.5.3 In vivo experiments

For in vivo evaluation, we used Affymetrix gene expression data set of bacterium *E. coli* (compiled and made available by (338) also known as DREAM5 challenge data). The dataset had been normalized already using Robust Multichip Averaging (RMA) (342). The compendium's GE data corresponds to genetic and environmental perturbation experiments on various strains. We only used profiles from the wild-type strain (MG1655) for our evaluations. There are 427 wild type GE

profiles. For replicates, we use the mean GE value resulting in 227 GE profiles corresponding to unique experimental settings.

4.5.3.1 Transcription Network

To identify transcription network, we used GENIE3 (343) which performed best for transcription network inference in DREAM5 challenge (338). The network inference method GENIE3 takes GE data as input and produces a list of TF-Gene relationships ordered based on confidence level we call $edge\_candidates$.

4.5.3.2 Master Regulators

The set of transcription factors on top of the regulatory hierarchy are referred to as master regulators (344). To define this more concretely, we use directional graph $G = \{V, E\}$ to represent a TRN where gene x is represented by vertex $v_x \in V$. An edge $(v_y, v_x) \in E$ represents transcriptional regulation of gene x by the product of gene y. The estimated confidence for edges are stored in matrix W where $W_{x,y} \in \mathbb{R}_{\geq 0}$ represents the reported confidence from network inference for edge $(v_y, v_x) \in E$. Note that $W_{x,y} = 0$ if there is no edge between vertices v_x, v_y. Here an MR gene is considered to be a TF gene that is not regulated by any other TF. Additionally, in cases where genes inside a regulatory cycle are non-reachable using any MR gene, the gene inside the cycle with maximum $impact$ will be selected as MR among them. The $impact(x)$ for given gene x is calculated as in Equation (4.20) where d_x is the number of genes regulated by the product of gene x:

$$impact(x) = \frac{1}{d_x} \sum_{j \in V} W_{x,j} \tag{4.20}$$

4.5.3.3 In vivo evaluation pipeline and results

Sub-sampling and network inference: We generate 10 datasets using stratified sampling (each containing 11 GE profiles). For each of the 10 datasets, we perform network inference using the remaining samples by GENIE3. This provides 10 networks, each with a dataset that was not used to infer the network. From each network, we extract 11 TRN modules with number of genes ranging from 10 to 1000 generating 110 TRN modules in total. To extract a module with N number of genes, we start with an empty network $G = \{V, E\}$. First, we add edges to this network starting from highest confidence (from $edge\_candidates$ list produced by GENIE3), until $|V| = N$. Second we add 20% more edges from $edge\_candidates$. Finally we run the greedy module extraction method using GeneNetWeaver (336) with the desired network size N to extract a TRN network module.

Dataset construction: For each network module, MR genes are identified using method explained in 4.5.3.2. Module's corresponding dataset (which consist of GE profiles not used in inferring the parent network) is then partitioned into input (GE of MR genes) and output (GE of non-MR genes). Stratified sampling and 5-fold cross validation is performed same as explained in 4.5.1.

The role of TRN information: We performed a separate experiment to evaluate the role of TRN information on predictive performance of methods. In this experiment, we randomly shuffle the gene names in the GE output data after the network inference step (this is same shuffling the node names in the network while preserving MR gene names). We then perform same 5-fold cross validation as explained before. This simulates a situation where the network information used by the model is random. Corresponding results are reported as GNN-rnd and LinGNN-rnd in **Figure 4.4** using dashed lines.

Figure 4.4 summarizes the results indicating better overall prediction performance for GNN (average MAE 0.58 ± 0.02; PCC 0.81 ± 0.04) compared to LinGNN (average MAE 0.60 ± 0.01; PCC 0.79 ± 0.02), Lasso (average MAE 0.81 ± 0.04; PCC 0.70 ± 0.02), MLP (average MAE 0.78 ± 0.06; PCC 0.75 ± 0.06), RNN (average MAE 0.93 ± 0.12; PCC 0.68 ± 0.07) and BiRNN (average MAE 0.81 ± 0.03; PCC 0.74 ± 0.04). Note that in vivo GE values range from 3 to 15 while in silico GE values are normalized between 0 and one.

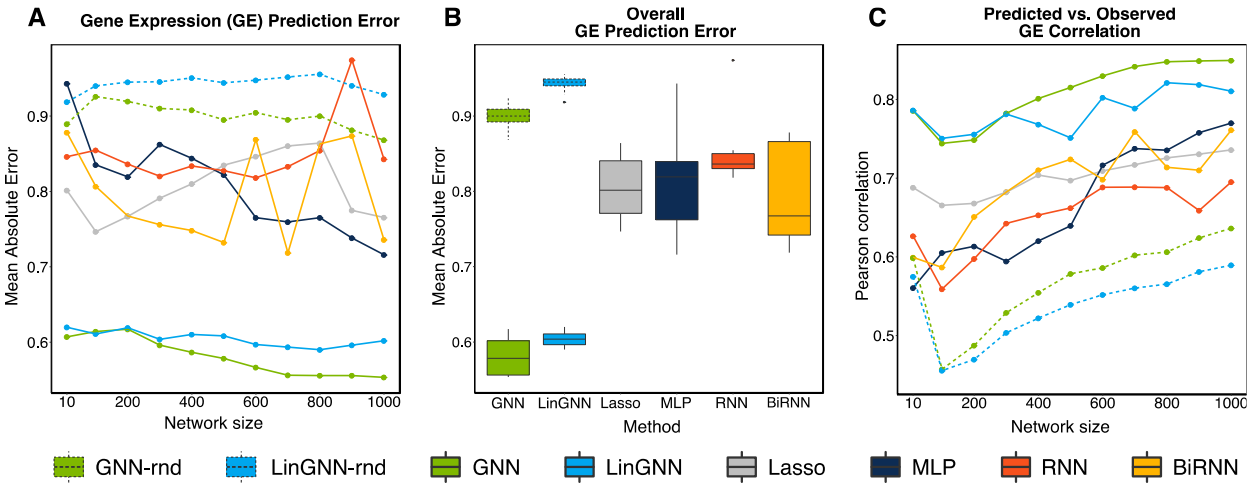

Figure 4.4 5-fold cross validation performance evaluation using in vivo microarray data. **A, B, C** show performance of all methods on 110 randomly selected network modules on inferred transcription network of *E. coli*. GNN shows better overall performance. GNN-rnd and LinGNN-rnd show the performance of TRN based methods when inferred TRN is randomized.

4.5.4 Runtime Comparison

For runtime comparison of methods, we used a dataset with size 10 for a network of 1000 genes and evaluated the training time. As in **Table 4.1**, GNN is lacking in terms of runtime compared to other methods. LinGNN performs best due to fast OLS operations on small datasets. Other methods require hyper-parameter optimization adding to their runtime (e.g., BiRNN is slower than when 50 hyper-parameter combinations are used). Note that the training procedure (described in section 4.3.2) is inherently parallel. Therefore, a parallel implementation can make the training approximately n times faster where n is the number of cores used.

Architecture	GNN	LinGNN	Lasso	MLP	RNN	BiRNN
Runtime (min:sec)	24:26	0:01	0:12	0:15	0:27	1:16

Table 4.1 Training time for dataset with 1000 genes and 10 GE profiles.

4.6 Conclusion

We presented GNN, an artificial neural network that incorporates gene regulatory network into it's architecture to predict GE in novel conditions given minimal training data. A trained GNN takes the expression level of MR genes and information about knockout experiments and predicts the expression of the rest of genes in the given transcription network. We compared GNN with three common neural network architectures, linear regression with ℓ_1 regularization and a network based linear model. Our comparison benchmarks include in vivo micro array data and thermodynamic simulation data for real biological network (e.g., chemotaxis). In our evaluations GNN showed considerably higher prediction performance when tested on hundreds of real TRNs extracted from *E. coli*'s full TRN. This was in spite of the fact that GNN did not enjoy the hyper-parameter optimization used for competing methods. The prediction performance gap was particularly higher on smaller data sets. Although this is not the first time ANNs were employed for GE prediction, this is a novel architecture with a custom designed node that is tailored for gene expression prediction. Concomitantly, to best of our knowledge, this is the first time that TRN, expression of MR genes and gene knockout information are used together for this task.

One limitation of the GNN architecture that we described in this paper is that our implementation cannot take into account feedback loops, as it is based on a feed-forward network. Given the prevalence of cycles in biological networks (345), such limitation is expected to negatively impact predictive power. A natural extension would be to apply the GNN cell to recurrent neural networks (RNNs), which have the capacity to connect through time multiple instances of acyclic network

maps, by feeding to the hidden layer of the next time slice, the hidden layer output of the previous time slice. Although individual GNN models are acyclic, together they have potential to model dynamics that arise in biological cycles. It would be also good to test the performance of this method in larger networks with tens of thousands of nodes. For that to happen with the non-linear GNNs, we need to take advantage of parallelism for the training algorithm, as training time is a consideration (**Table 4.1**). It would also help if the activation function is modified to one that can formulate a convex problem and its optimization in layer-wise training. Given that the linear GNN is performing quite well, despite its simplicity and more than an order of magnitude faster performance, a system that runs the LinGNN for very large networks (>5000 nodes) and non-linear GNN otherwise, would score high in prediction, runtime performance and scalability.

Extensions to this work include the integration of contextual information, such as gene sequence, experimental settings and metabolic pathways. More thorough validation in large compendia (e.g. see (295)) and multiple pathways may further pinpoint the limitations of this approach. Although our focus here is bacterial regulation, this work can be extended to organisms of higher complexity, albeit with modifications that would rectify the large discrepancy in the number of genes (from 182 in bacterium *Carsonella ruddii* (346) to more than 20,000 in humans) and more complex modes of regulation.

Chapter 5: Algorithmic lifestyle optimization

5.1 Abstract

Motivation. A vast amount of medical and nutrition research is focused on identifying the most effective treatment for adverse health conditions. In some chronic health conditions however, the effectiveness of several lifestyle interventions is identified one at the time for a given individual, which can be exhaustive when the number of candidate lifestyle interventions is high. For example, in irritable bowel syndrome (IBS), and in food allergies, the standard elimination diet (SED) is used where the effectiveness of removing individual foods from the diet for symptom relief is identified 1-by-1.

Results. We have developed algorithmic lifestyle optimization (ALO), for rapid identification of effective lifestyle interventions (LIs) in individuals, using a novel group testing algorithm called constrained adaptive group testing (CAGT). CAGT works by determining the lack of efficacy for multiple LIs in a group by following them simultaneously (instead of 1-by-1). The group of LIs that CAGT suggests for the next round is informed by the individual's response to the groups of LIs that have been followed by the individual in prior rounds, as well as the minimum and maximum number of effective LIs in a given group estimated by ALO. ALO has three modules where in the first two modules, the configuration in which CAGT achieves its optimal performance for a given set of candidate LIs and their effectiveness probabilities are identified, and in the third module CAGT is used to identify the effectiveness (i.e., 0|1 potency) of LIs in each individual. Our evaluations on synthetic and real data shows that ALO is robust to noise, data size and data heterogeneity, is between 58.9% to 68.4% more efficient compared to SED for identification of

effective LIs in IBS and food allergies, and better than alternative state of the art group testing method for this application. ALO provides a novel approach for rapid discovery of effective interventions in nutrition and medicine, and can lead into substantial improvements in the status que.

5.2 Introduction

The variance of responses to lifestyle interventions (LI) has been a major challenge in the fields of nutrition and medicine throughout history (347,348). Typical LI include changes in diet (348), exercise (349), administrating drugs (350), stress management (351), smoking cessation (352), assisted sleep methods (353), and fasting (354), among others. It has been shown that biomarkers can be used to predict an individual's response to a given lifestyle intervention (LI) (2,355,356), however, such biomarkers are often difficult to acquire and carry limited information when in isolation (357,358). Therefore, trial-and-error remains as the main alternative in which target health metrics are evaluated following an LI regimen to determine an individual's response hence labeling each LI as "potent" or "impotent" (359,360).

Previous studies have proposed and measured the adoption of systematic strategies for particular interventions including diet and physical activity. The standard elimination diet (SED) is used for identification of food allergies in serous otitis (361), atopic dermatitis (eczema) (359), as well as food intolerances in irritable bowel syndrome (362), esophagitis (363), and ADHD (364) among others. In SED, series of oral food challenges are used in which target symptoms are evaluated following dietary elimination and subsequent introduction of each food for 2-3 days at the time. More recently, N-of-1 trials have emerged for systematic personalization of medical treatments in cases were the individualized potency of alternative treatment strategies need to be determined

(360). They involve trial periods during which alternative treatments are followed one after the other and treatment outcomes are measured in order identify the treatment with the best statistical support. N-of-1 trials are used for dietary intervention in inflammatory bowel disease (365), determining the impact of dietary macronutrients on postprandial glucose response (366), and personalized goal setting strategies to increase physical activity (367) among others. These trial-and-error approaches commonly involve a single LI at a time which is not time efficient when there is a large number of non-interacting candidate LIs, many of which are impotent for most people. Therefore, the number of candidate LIs that can be evaluated by an individual will be limited given the time that they can spend for determining LI responses.

To address this issue in a faster, less invasive, and more efficient way, we propose a systematic approach that we call "algorithmic lifestyle optimization (ALO)", a heuristic approach for identifying the individualized binary labels (i.e. potent or impotent) of the candidate LIs, based on heterogeneous data, including biomarker information. In ALO, the required time for discovering candidate LI potencies in an individual is minimized using a group testing strategy. In its core, ALO uses an adaptive group testing strategy and involves multiple rounds of LIs for each individual. In each round, a set of LIs are provided to the individual to follow. These LIs are chosen by ALO based on **(a)** the individual's health score (0|1) in response to each set of LIs in prior rounds, and **(b)** the probability of a positive health score for each LI in a population. These probabilities may also be calibrated based on a biomarker when available. In ALO, we strive to identify the individual's response to each LI in minimal number of rounds and provide guarantees for both average and worst-case scenarios. The ALO methodology is fully described under the Methods section and illustrated in **Figure 5.1**.

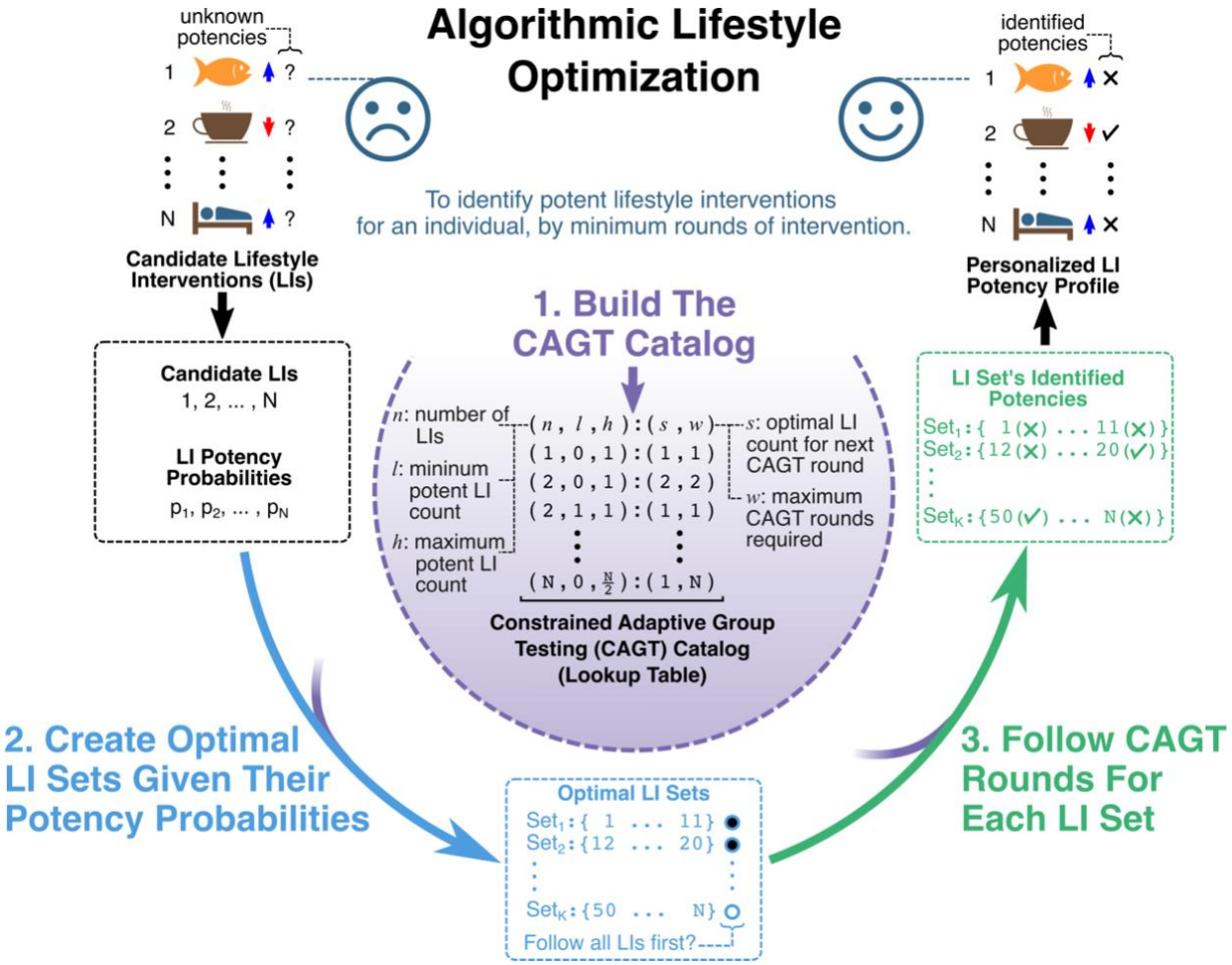

Figure 5.1 Algorithmic Lifestyle Optimization (ALO). ALO is designed to guide individuals in rapid discovery of lifestyle interventions (LIs) that are effective (potent) for them amongst many candidate LIs, for achieving a target health outcome. First, it builds the constrained adaptive group testing (CAGT) catalog, which is a lookup table for finding the maximum number of rounds needed by the CAGT algorithm for identifying between l and h number of potent LIs amongst n candidate LIs. Second, it partitions the LIs into disjoint sets given the potency probability of each LI, and determines whether the first step of the CAGT algorithm involves following all the LIs in a given set. These probabilities can be estimated from population wide studies that report the percentage of individuals that achieve the target health outcome following each LI. Third, the suggested LIs by the CAGT algorithm is followed by the individual in subsequent rounds. The CAGT algorithm stops once the potency of the LIs in each set is identified.

5.3 Methods

5.3.1 Main Algorithm

ALO has three major modules, all of which rely on the constrained adaptive group testing (CAGT) method that we have developed. Briefly in adaptive group testing (368), groups of available objects

are selected in sequential rounds for testing, with the goal of discovering the target objects (e.g. defective light bulbs, SARS-CoV-2 positive nasal swabs, or the potent LIs that we discuss here) amongst many, in minimum number of rounds. Group testing is applicable in cases where objects are noninteracting. This means that if multiple objects are tested together in a group, a positive test result is indicative of one or more target objects in the group (e.g. at least one defective light bulb in the group), while a negative test result indicates that the group is void of any target object (e.g. no defective light bulbs in the group). Note that, in this article, a "potent LI" corresponds to a "target object" that is subject to group testing while in the literature the "defective lightbulb" terminology is commonly used.

ALO is applicable in cases where **(a)** the individual is concerned about a single binary target health score such as having a symptom-free digestive state (0|1), **(b)** each LI is binary such as drinking coffee in the morning (yes/no), **(c)** it takes the same amount of time (e.g. 3-days) to see the impact of each LI on health score, **(d)** multiple LIs are independent hence can be followed together simultaneously, and **(e)** multiple LIs are noninteracting. Noninteracting here means that if a set of LIs together are determined to be "impotent" (i.e., not leading to a positive health score), we can conclude that each LI is also "impotent". However, when a given LI is "potent" (i.e., leading to a positive health score), it will remain as "potent" when combined with other LIs.

Constrained adaptive group testing (CAGT). ALO, in its core, relies on the CAGT algorithm that aims to identify the minimal number of adaptive group testing rounds needed to identify the potent LIs amongst the candidate LIs (LI) for a given individual, by solving the optimization problem in Equation (5.1). Here $R_i \subseteq LI$ represents the group of LIs that will be followed simultaneously by the individual in round i during which the potency of R_i will be determined as represented by $r_i \in \{0{:}\text{impotent}, 1{:}\text{potent}\}$. V_1 and V_0 represent the sets of potent LIs and

impotent LIs respectively which can be fully identified by a function f given $LI, R, \boldsymbol{r}$ as well as the l:low and h:high bounds for the number of potent LIs.

$$\mathrm{R}^{*} = \underset{R}{\textbf{argmin}} \quad |R| \tag{5.1}$$

$$\textbf{subject to} \quad R = \left[R_1 \dots R_{|R|}\right], R_i \subseteq LI, i = 1, \dots, |R|$$

$$\boldsymbol{r} = \left[r_1 \dots r_{|R|}\right], r_i \in \{0, 1\}, i = 1, \dots, |R|$$

$$[V_0, V_1] = f(LI, R, \boldsymbol{r}, l, h)$$

$$V_0 \cup V_1 = LI$$

$$l \leq |V_1| \leq h$$

In CAGT, we solve Equation (5.1) following **Algorithm 1.** with three major steps in each round, using the $CAGT\_Model$ that captures l and h bounds for subsets of LIs that are generated in each round. In step1, a non-nested subset of LIs (R_i) that is expected to minimize the final $|R|$ is identified given the model. In step2, the potency r_i of R_i is determined by the individual. In step3, the model is updated (given R_i and r_i), and the sets of impotent and potent LIs (V_0 and V_1) that can be determined using the updated model are identified. These three steps are repeated until the potency of all LIs are identified. See **Appendix Section A.1** that describes the $CAGT\_Model$ and its relevant functions in detail.

Algorithm 1. Solve the optimization problem in Equation (5.1) using the CAGT algorithm.

Inputs: The set of candidate LIs (LI). The low and high thresholds (l and h) that bound the number of potent LIs.

Outputs: The set of impotent LIs V_0 and potent LIs V_1 identified by the algorithm.

1: $V_0 \leftarrow \{\}; V_1 \leftarrow \{\}$
2: $model \leftarrow CAGT\_Model(LI, l, h)$
3: **do:**
4: $R_i \leftarrow model.next\_round()$ // step1
5: $r_i \leftarrow get\_potency(R_i)$ // step2
6: $(V_0, V_1) \leftarrow model.f(R_i, r_i)$ // step3
7: **while** $|LI \setminus V_0 \setminus V_1| > 0$
8: **return** (V_0, V_1)

ALO Module-1 (build the CAGT catalog). In the first module, we build the CAGT catalog which is a lookup table that the step1 of **Algorithm 1.** relies on. This lookup table determines the tuple (s, w) for a given tuple (n, l, h) where w is the maximum number of rounds that the algorithm needs for identifying the potencies of n LIs when there are between l and h potent LIs amongst them. The value of s, determines the number of LIs to be used in the first round of **Algorithm 1.** in order to achieve w for the given (n, l, h) tuple. A dynamic programming strategy is used for building the CAGT catalog based on the fact that in each round of **Algorithm 1.,** the $CAGT\_Model$ gets updated and existing LI subsets within the model are split into smaller subsets. Therefore, in this module, we populate the catalog starting from tuples with $n = 1$ for which the optimal (s, w) are known, and iteratively populate the catalog by tuples with larger n values given the catalog itself. See **Appendix Section A.2.1** for further details.

ALO Module-2 (create optimal LI sets). In this module, we use the LI potency probabilities (estimated apriori) to create an optimal LI partition (i.e., disjoint LI sets), such that the expected total number of rounds needed for identifying LI potencies is minimized while the maximum total

number of rounds is kept at bay. This is done by (a) ordering the LIs by their potency probabilities, (b) estimating "h" for a given LI set, from the corresponding potency probabilities, the Poisson binomial distribution, and a confidence threshold t, (c) using the CAGT catalog to determine "w" for a given set of LIs with an estimated "h", and (d) allowing "ex" more rounds compared to the maximum total rounds needed, for decreasing the expected total rounds needed by introducing rounds that involve all LIs in a set. This module, as described in **Appendix Section A.2.2**, provides the LI sets that are used in separate runs of **Algorithm 1.** for the next module and identifies the runs that start will an initial round that involve all LIs.

ALO Module-3 (follow CAGT rounds for LI sets). Lastly for each individual, we perform independent runs of the **Algorithm 1.** where in each run a disjoint set of LIs (determined by Module-2) is used which leads into determining the potency of each LI after all runs are completed. See **Appendix Section A.2.3** for further details.

5.3.2 Evaluation

5.3.2.1 Datasets

In our evaluations, we relied on synthetic data for robustness and sensitivity analysis, and on real data for food intolerance and allergy identification applications.

Synthetic data. We initiated the data generation from three sets of LI potency probabilities each with 50 values that follow beta distributions with three different shapes (Dataset-1: α=0.5, β=5.0, Dataset-2: α=2.0, β=6.0, and Dataset-3: α=0.1, β=0.1). Next, we generated 200 values for each LI potency probability of the prior step following Bernoulli distributions parametrized by each probability value. This provided us with three datasets that each consists of a 200x50 matrix that represent the LI potencies for 200 individuals, along with the set of LI potency probabilities that were used to generate each. Finally, for each set of LI potency probabilities in a dataset, we

generated nine sets of noisy LI potency probabilities by adding different levels of white noise with standard deviation (SD) values that ranged from 0.05 to 0.5. These noisy LI potency probabilities were clamped in the 0-1 range (i.e., set to 0 if less than 0, and set to 1 if greater than 1).

Real data. We defined two sets of LIs, one for management of food intolerances in IBS and another for management of allergic food reactions. In both LI sets, an LI corresponds to the elimination of a particular food from the patient's diet, and the LI's potency probability corresponds to the percentage of individuals in which a given food triggers adverse symptoms. First, we extracted the LIs and their potency probabilities from published studies of IBS (369) and food allergies (370) separately. Second, we used the Poisson distribution parametrized by the average number of potent LIs from each study (reported as 7 in the IBS study and estimated as 1.43 for the food allergy study given their reported statistics), in order to generate one-thousand integers for each study, where each integer corresponds to the number potent LIs in a given individual. Finally, we randomly assigned individual potency values (0|1) for the LIs in each patient given the number of potent LIs in each, and the potency probability of each LI that was extracted from the corresponding study. This provided us with an IBS dataset with 56 LIs, and a food allergy dataset with 19 LIs, each with the corresponding potency probabilities, and one thousand LI potency profiles that adhere to the reported summary statistics.

5.3.2.2 Evaluation metrics

We used the average and median number of rounds needed for identifying the LI potencies individuals for our method evaluations. For each dataset, we first identified the optimal hyper-parameters using grid search on half of the dataset, then performed our evaluations on the remaining records. In each case, a maximum of fifty pair of hyper-parameter values were examined for ex and t in the ALO method, while for the SPIV method, a maximum of hundred hyper-parameter value pairs were examined for its epsilon, and t parameters.

5.4 Results

5.4.1 Robustness and sensitivity analysis

We generated three datasets with various levels of homogeneity (**Figure 5.2 A**, **D** & **G**), in order to evaluate the sensitivity of each method to the number of LIs (**Figure 5.2 B**, **E** & **H**), and to the noise in LI potency probabilities (**Figure 5.2 C**, **F** & **I**). In all cases the average rounds needed for identifying the LI potencies increased linearly while the ALO method had the lowest increase, followed by SPIV, and the baseline (**Figure 5.2 B**, **E** & **H**). The largest reduction in average rounds needed for ALO compared to the baseline was observed for Dataset-1 and Dataset-3 in which a large portion of LIs have low potency probabilities (**Figure 2. B** & **E**). This reduction was much lower for Dataset-2 in which a lower proportion of LIs have low potency probabilities (**Figure 5.2 E**). The addition of white noise to LI potency probabilities increased the average rounds needed by each method (see **Figure 5.2 C**, **F** & **I** where methods were evaluated on all 50 LIs while white noise with varying standard deviations (SD) were added to the LI potency probabilities). For example, white noise with SD of 0.5 increased the average rounds needed in Dataset-1 by ALO from 18.2 to 26.5 (45.6%), and by SPIV from 26.9 to 34.7 (29.0%) (**Figure 5.2 C**).

5.4.2 Rapid food intolerance and allergy identification with ALO

The gold standard method used in the clinic for identifying foods that cause intolerance or allergic reactions, is the standard elimination diet (SED) during which food challenges are performed. A food challenge is a lifestyle intervention (LI) during which target health symptoms are monitored while a given food item is introduced to the individual's diet for 3 days, then subsequently removed from the diet for another 3 days (the number of days may vary). We compared ALO with SED as well as a state of the art group testing method called spatial inference vertex cover (SPIV) (371) for identification of food intolerances and food allergies as described next.

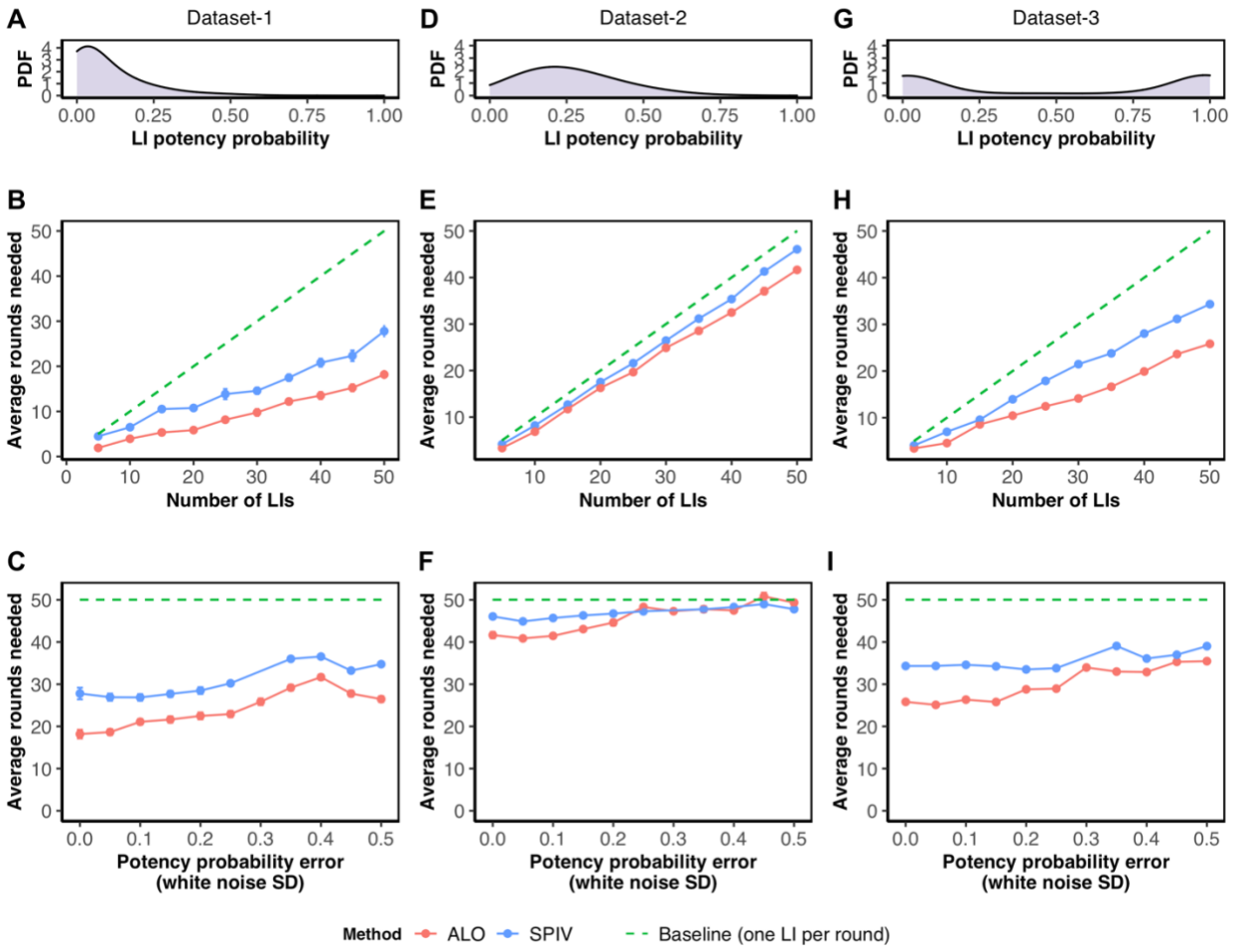

Figure 5.2 Robustness and Sensitivity analysis. Three synthetic datasets of potency probabilities relating to fifty (50) LIs were sampled independently from heterogenous beta distributions A, D & G, and subsequently used to generate synthetic datasets 1-3 each representing the LI potencies (0|1) for a hundred (100) individuals. B, E & H (relating to datasets 1-3) illustrate the average number of rounds needed by each method to identify the potent LIs in hundred individuals for LI subsets having 5 to 50 LIs each. C, F & I (relating to datasets 1-3) illustrate the method's robustness to the standard deviation (SD) of the added white noise that was added to LI potency probabilities. The error bars represent the standard error.

Food Intolerance in IBS Case Study. IBS is a chronic gastrointestinal disease with 11% prevalence in adults (251). One of the most effective symptom management strategies of IBS is to identify their food intolerances (i.e., food items that exacerbate IBS symptoms such as bloating, constipation, diarrhea and abdominal pain) and eliminate them from the patient's diet. We used ALO for discovery of food intolerances based on realistic synthetic data of 500 IBS patients given self-reported intolerance statistics of 56 food items (369) and compared the performance of ALO

with the standard elimination diet (SED) involving a constant 56 of LI rounds. The results are shown in **Figure 5.3 A** where ALO reduced the median number of LI rounds by 58.9% (33/56), while the SPIV method reduced the number of LI rounds by 32.1% (18/56). Our results suggest that both ALO and SPIV can replace the SED method in the clinic, however our novel ALO method showed 26.8% advantage compared to the SPIV.

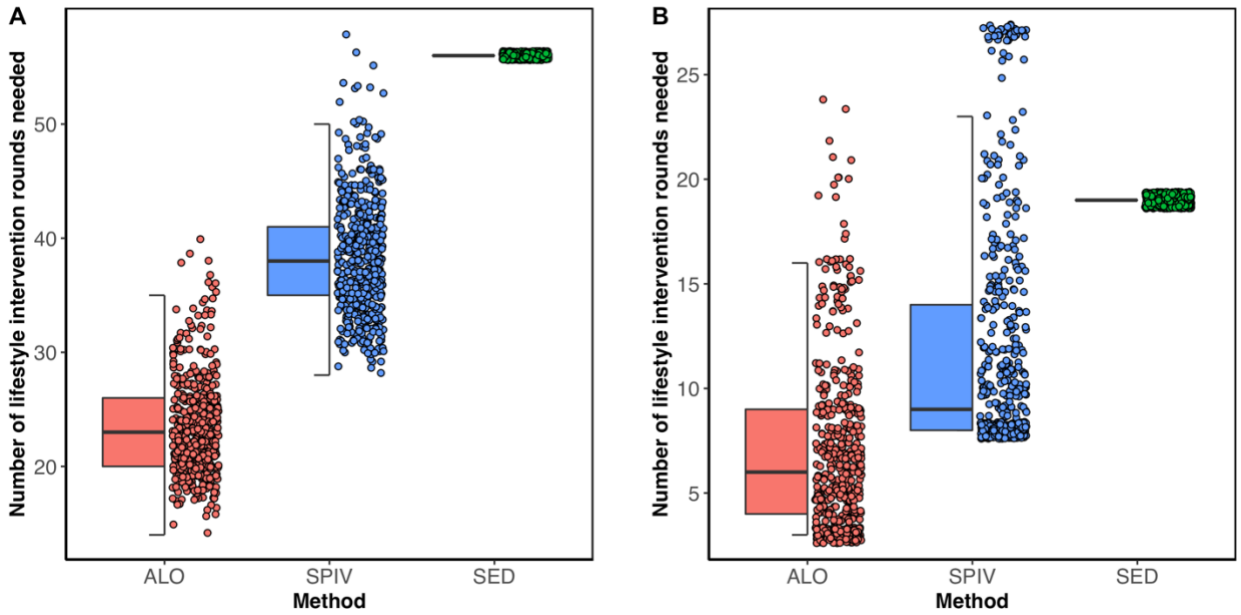

Figure 5.3 Rapid IBS food intolerance and allergy identification. Various methods were used for discovery of food intolerances in IBS and food allergies (**A**&**B**). **(A)** ALO and SPIV methods lead into 58.9% and 32.1% reduction in median number of lifestyle intervention (LI) rounds needed compared to SED, for discovering the foods that exacerbate IBS symptoms amongst 56 foods in 500 IBS patients. **(B)** The median number of LI rounds needed compared to SED was reduced by 68.4% using ALO, and by 52.6% using SPIV, for identifying the foods that trigger food allergies in 500 patients.

Food Allergy Case Study. Food allergy is an immune response from food exposure, and has a prevalence between 5.3% to 9.1% in the United States' adults (370). Food allergy can be managed by strict avoidance of trigger foods that can be identified using SED. We simulated ALO for food trigger identification based on realistic synthetic data from 500 individuals given medical doctor diagnosed food trigger statistics of 19 foods (370), and compared the performance of ALO with SED and SPIV. The results are illustrated in **Figure 5.3 B** where ALO reduced the median number

of LI rounds by 68.4% (13/19), while the SPIV method resulted in 52.6% (10/19) reduction compared to SED. Both ALO and SPIV showed considerable performance advantage over SED while ALO method was 15.8% more efficient than SPIV.

5.5 Discussion

We developed algorithmic lifestyle optimization (ALO) for rapid identification of lifestyle interventions (LIs) that a given individual needs for achieving a target health goal such as a symptom-free digestive state. ALO relies on estimated LI potency probabilities that can come from population wide studies that report the percentage of population in which a given LI is potent for achieving the target health goal. ALO uses a group testing method that we have developed called constrained adaptive group testing (CAGT) for identifying the group of LIs that a given individual needs to include in their lifestyle in each round, given their health state (0|1) in response to LI groups followed in prior rounds, as well as the minimum and maximum number of potent LIs in the set of candidate LIs. The first ALO module builds a lookup table named "CAGT catalog" that CAGT relies on for optimal performance. The second ALO module, create an optimal LI partition (disjoint sets of LIs) that leads to minimum total rounds when CAGT is followed on each set separately. The third ALO module involves LIs that are suggested by the CAGT algorithm given each LI set until the potency of all LIs are determined.

We evaluated ALO, for rapid food intolerance and allergy identification, and compared to the standard elimination diet (SED) method that is commonly used (**Figure 5.3**). Our results indicate between 58.9% and 68.4% reduction in number of LI rounds needed by ALO when compared to SED. The ALO method is robust to noise and works for different sets of LIs with varying homogeneity properties (**Figure 5.2**). This is the first time that a group testing is suggested for this

application. Furthermore, our experiments show that our newly developed method is better than a state of the art group testing method called SPIV (371) for this application.

Future research should focus on evaluating group testing methods such as ALO in practice for personalized lifestyle intervention, in order to improve the efficiency of existing methods such as SED and N-of-1 trials and identify application specific considerations that need to be made in the group testing method to minimize the associated risks and maximize its practical efficiency. We anticipate that future algorithmic improvements using active machine learning, and optimal experimental design that are shown to speedup biological discoveries (372), will lead into further performance improvements, and guide us into a new era of personalized nutrition.

Chapter 6: Conclusion

Computation is a major resource for enhancing our ability to optimize a given biological organism for achieving a desired phenotype which is the fundamental goal in many areas of science including medicine, nutrition and agriculture. Despite the progress in development of computational methods for optimizing biological organisms (Chapter 2), several challenges remain to be addressed sufficiently: (i) data heterogeneity and disconnect between the available data and existing computational methods, (ii) integration of domain knowledge into uninterpretable data-driven models, and (iii) the plethora of factors that can perturb a biological phenotype, and limited time to examine each. These challenges are addressed in Chapter 3, Chapter 4 and Chapter 5 for several important applications. The first application is the development of a microbiome-based optimal dietary strategy for management of IBS as presented in Chapter 3. The second application is the prediction of genome-wide transcriptome given gene knockouts and master regulator perturbations in bacterium *E. coli*. The third application is the rapid identification of effective lifestyle interventions in individuals, such as dietary food eliminations for management of IBS and food allergies.

The presented computational methods can be adopted for many applications. The data integration, and data analysis methods that are used in Chapter 3, can be applied for identifying the biomarkers of response to various treatment strategies. The genetic neural network architecture presented in Chapter 4, can be leveraged in product yield maximization of bacteria, genetic engineering of crops, and multi-target drug discovery in animal and human diseases. The algorithmic lifestyle optimization method presented in Chapter 5 can be used to improve the efficiency of N-of-1 trials in animal and human diseases.

In conclusion, three strategies are presented here for addressing aforementioned challenges in development of computational methods for optimizing biological systems. First, is to gather and homogenize heterogenous biological datasets and analyze them through careful application of existing computational methods. Second, is to integrate the domain knowledge (such as gene regulatory relationships) into data-driven predictive models in order to improve their accuracy with minimum amount of data and make them interpretable. Third, is to use constrained adaptive group testing for evaluating the effect of multiple factors simultaneously on a given phenotype, using minimum number of experiments. The developed methods that employ these strategies can be advanced further as described at the end of each chapter. The presented results here show that each strategy leads into substantial performance improvements compared to the status quo. Furthermore, combining these strategies together, as well as with active learning and reinforcement learning strategies, can give rise to novel methods that will enhance our ability to optimize biological organisms by several orders of magnitude. Provided with sufficient resources and commitment in the relevant research, we can see a future in which biological organisms are healthier and more capable towards higher peace and prosperity.

References

1. Eetemadi A, Rai N, Pereira BMP, Kim M, Schmitz H, Tagkopoulos I. The Computational Diet: A Review of Computational Methods Across Diet, Microbiome, and Health. Frontiers in Microbiology. 2020;11:393.

2. Eetemadi A, Tagkopoulos I. Methane and fatty acid metabolism pathways are predictive of Low-FODMAP diet efficacy for patients with irritable bowel syndrome. Clinical Nutrition. 2021;

3. Eetemadi A, Tagkopoulos I. Genetic Neural Networks: An artificial neural network architecture for capturing gene expression relationships. Bioinformatics. 2018 Nov 19;bty945–bty945.

4. Lush JL, others. Animal breeding plans. Animal breeding plans. 1943;(Edn 2).

5. Allard RW. Principles of plant breeding. John Wiley & Sons; 1999.

6. Castiglioni A. A history of medicine. Routledge; 2019.

7. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. Science translational medicine. 2012;4(158):158rv11-158rv11.

8. Barry AN, Starkenburg SR, Sayre RT. Strategies for optimizing algal biology for enhanced biomass production. Frontiers in Energy Research. 2015;3:1.

9. Reali F, Priami C, Marchetti L. Optimization algorithms for computational systems biology. Frontiers in Applied Mathematics and Statistics. 2017;3:6.

10. Markowetz F. All biology is computational biology. PLoS biology. 2017;15(3):e2002050.

11. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Current opinion in gastroenterology. 2015 Jan;31(1):69–75.

12. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host microbiome as an ecosystem on a leash. Nature. 2017;548(7665):43.

13. Barratt MJ, Lebrilla C, Shapiro H-Y, Gordon JI. The gut microbiota, food science, and human nutrition: a timely marriage. Cell host & microbe. 2017;22(2):134–41.

14. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery. 2016;15(7):473.

15. Taroncher-Oldenburg G, Jones S, Blaser M, Bonneau R, Christey P, Clemente JC, et al. Translating microbiome futures. Nature Publishing Group 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA; 2018.

16. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018;1.

17. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature medicine. 2018;24(4):392.

18. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. Cell. 2015;163(5):1079–94.

19. Korem T, Zeevi D, Zmora N, Weissbrod O, Bar N, Lotan-Pompan M, et al. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. Cell metabolism. 2017;25(6):1243–53.

20. Thaiss CA, Itav S, Rothschild D, Meijer MT, Levy M, Moresi C, et al. Persistent microbiome alterations modulate the rate of post-dieting weight regain. Nature. 2016;540(7634):544.

21. Bauer E, Thiele I. From metagenomic data to personalized in silico microbiotas: predicting dietary supplements for Crohn's disease. NPJ systems biology and applications. 2018;4(1):27.

22. Baldini F, Heinken AK, Heirendt L, Magnusdottir S, Fleming R, Thiele I. The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. Bioinformatics. 2018;

23. Greenhalgh K, Ramiro-Garcia J, Heinken A, Ullmann P, Bintener T, Pacheco MP, et al. Integrated in Vitro and in Silico Modelling Delineates the Molecular Effects of a Symbiotic Regimen on Colorectal Cancer-Derived Cells. Available at SSRN 3287784. 2018;

24. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, et al. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell metabolism. 2015;22(2):320–31.

25. Sherwin E, Dinan TG, Cryan JF. Recent developments in understanding the role of the gut microbiota in brain health and disease. Annals of the New York Academy of Sciences. 2018;1420(1):5–25.

26. Zmora N, Suez J, Elinav E. You are what you eat: diet, health and the gut microbiota. Nature reviews Gastroenterology & hepatology. 2019;16(1):35–56.

27. Pereira J, Rea K, Nolan Y, O'Leary O, Dinan T, Cryan J. Depression's Unholy Trinity: Dysregulated Stress, Immunity, and the Microbiome. Annual review of psychology. 2019;

28. Stinson LF, Boyce MC, Payne MS, Keelan JA. The not-so-sterile womb: Evidence that the human fetus is exposed to bacteria prior to birth. Frontiers in microbiology. 2019;10:1124.

29. Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature. 2019;574(7776):117–21.

30. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell host & microbe. 2018;24(1):133–45.

31. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proceedings of the National Academy of Sciences. 2010;107(26):11971.

32. Rodríguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, et al. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Health Dis [Internet]. 2015 Feb 2 [cited 2018 Dec 28];26. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315782/

33. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015 May 13;17(5):690–703.

34. Wang M, Karlsson C, Olsson C, Adlerberth I, Wold AE, Strachan DP, et al. Reduced diversity in the early fecal microbiota of infants with atopic eczema. Journal of Allergy and Clinical Immunology. 2008;121(1):129–34.

35. Abrahamsson TR, Jakobsson HE, Andersson AF, Björkstén B, Engstrand L, Jenmalm MC. Low diversity of the gut microbiota in infants with atopic eczema. Journal of Allergy and Clinical Immunology. 2012 Feb 1;129(2):434-440.e2.

36. Zheng H, Liang H, Wang Y, Miao M, Shi T, Yang F, et al. Altered Gut Microbiota Composition Associated with Eczema in Infants. PLoS ONE. 2016;11(11):e0166026.

37. Yuan C, Gaskins AJ, Blaine AI, Zhang C, Gillman MW, Missmer SA, et al. Cesarean birth and risk of offspring obesity in childhood, adolescence and early adulthood. JAMA pediatrics. 2016;170(11):e162385–e162385.

38. Thavagnanam S, Fleming J, Bromley A, Shields MD, Cardwell CR. A meta-analysis of the association between Caesarean section and childhood asthma. Clin Exp Allergy. 2008;38(4):629–33.

39. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Current opinion in gastroenterology. 2015;31(1):69.

40. Conlon M, Bird A. The impact of diet and lifestyle on gut microbiota and human health. Nutrients. 2015;7(1):17–44.

41. Willing BP, Russell SL, Finlay BB. Shifting the balance: antibiotic effects on host–microbiota mutualism. Nature Reviews Microbiology. 2011;9(4):233.

42. Mathew S, Smatti MK, Al Ansari K, Nasrallah GK, Al Thani AA, Yassine HM. Mixed Viral-Bacterial Infections and Their Effects on Gut Microbiota and Clinical Illnesses in Children. Scientific reports. 2019;9(1):865.

43. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55(2):205–11.

44. Scher JU, Ubeda C, Artacho A, Attur M, Isaac S, Reddy SM, et al. Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. Arthritis & rheumatology. 2015;67(1):128–39.

45. de Goffau MC, Luopajärvi K, Knip M, Ilonen J, Ruohtula T, Härkönen T, et al. Fecal Microbiota Composition Differs Between Children With β-Cell Autoimmunity and Those Without. Diabetes. 2013;62(4):1238–44.

46. Schippa S, Iebba V, Barbato M, Di Nardo G, Totino V, Checchi MP, et al. A distinctive "microbial signature" in celiac pediatric patients. BMC Microbiol. 2010;10(175):1471–2180.

47. Castaner O, Goday A, Park Y-M, Lee S-H, Magkos F, Shiow S-ATE, et al. The Gut Microbiome Profile in Obesity: A Systematic Review. International Journal of Endocrinology. 2018;2018:9.

48. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490:55.

49. Menni C, Lin C, Cecelja M, Mangino M, Matey-Hernandez ML, Keehn L, et al. Gut microbial diversity is associated with lower arterial stiffness in women. European heart journal. 2018;

50. Goldenberg JZ, Yap C, Lytvyn L, Lo CK-F, Beardsley J, Mertz D, et al. Probiotics for the prevention of Clostridium difficile-associated diarrhea in adults and children. Cochrane Database Syst Rev. 2017 19;12:CD006095.

51. Hao Q, Dong BR, Wu T. Probiotics for preventing acute upper respiratory tract infections. Cochrane Database Syst Rev. 2015 Feb 3;(2):CD006895.

52. Mansfield JA, Bergin SW, Cooper JR, Olsen CH. Comparative probiotic strain efficacy in the prevention of eczema in infants and children: a systematic review and meta-analysis. Mil Med. 2014 Jun;179(6):580–92.

53. Saez-Lara MJ, Gomez-Llorente C, Plaza-Diaz J, Gil A. The role of probiotic lactic acid bacteria and bifidobacteria in the prevention and treatment of inflammatory bowel disease and other related diseases: a systematic review of randomized human clinical trials. Biomed Res Int. 2015;2015:505878.

54. Delzenne NM, Olivares M, Neyrinck AM, Beaumont M, Kjølbæk L, Larsen TM, et al. Nutritional interest of dietary fiber and prebiotics in obesity: Lessons from the MyNewGut consortium. Clinical Nutrition. 2019;

55. Anderson J, Edney R, Whelan K. Systematic review: faecal microbiota transplantation in the management of inflammatory bowel disease. Alimentary pharmacology & therapeutics. 2012;36(6):503–16.

56. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS computational biology. 2010;6(2):e1000667.

57. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Briefings in bioinformatics. 2017;

58. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature biotechnology. 2017;35(9):833.

59. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, Vehik K, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. Nature. 2018;562(7728):589.

60. Brown CT, Davis-Richardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, et al. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. PloS one. 2011;6(10):e25792.

61. Walker A, Pfitzner B, Neschen S, Kahle M, Harir M, Lucio M, et al. Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. The ISME journal. 2014;8(12):2380.

62. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, et al. Shotgun metaproteomics of the human distal gut microbiota. The ISME journal. 2009;3(2):179.

63. Zhang X, Deeke SA, Ning Z, Starr AE, Butcher J, Li J, et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. Nature communications. 2018;9(1):2873.

64. Amann RI, Ludwig W, Schleifer K-H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Mol Biol Rev. 1995;59(1):143–69.

65. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal. 2012;6(3):610.

66. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2012;41(D1):D590–6.

67. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology. 2013;31(9):814.

68. Martiny AC, Martiny JBH, Weihe C, Field A, Ellis J. Functional metagenomics reveals previously unrecognized diversity of antibiotic resistance genes in gulls. Frontiers in microbiology. 2011;2:238.

69. Ranjan R, Rani A, Finn PW, Perkins DL. Evaluating bacterial and functional diversity of human gut microbiota by complementary metagenomics and metatranscriptomics. bioRxiv. 2018;363200.

70. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220.

71. Heintz-Buschart A, Wilmes P. Human gut microbiome: function matters. Trends in microbiology. 2017;

72. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207.

73. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. nature. 2009;457(7228):480.

74. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: networks, competition, and stability. Science. 2015;350(6261):663–6.

75. Relman DA. The human microbiome: ecosystem resilience and health. Nutrition reviews. 2012;70(suppl_1):S2–9.

76. Mehta RS, Abu-Ali GS, Drew DA, Lloyd-Price J, Subramanian A, Lochhead P, et al. Stability of the human faecal microbiome in a cohort of adult men. Nature microbiology. 2018;3(3):347.

77. Smith EP, van Belle G. Nonparametric estimation of species richness. Biometrics. 1984;119–29.

78. NIH Human Microbiome Project - Home [Internet]. [cited 2019 Feb 11]. Available from: https://hmpdacc.org/hmp/

79. NIH Human Microbiome Project - Home [Internet]. [cited 2019 Feb 11]. Available from: https://hmpdacc.org/ihmp/

80. American Gut – What's in your gut? [Internet]. [cited 2019 Feb 11]. Available from: http://americangut.org/

81. Data – The Harvard Personal Genome Project (PGP) [Internet]. [cited 2019 Feb 11]. Available from: https://pgp.med.harvard.edu/data

82. TwinsUK – The biggest twin registry in the UK for the study of ageing related diseases [Internet]. [cited 2019 Feb 11]. Available from: http://twinsuk.ac.uk/

83. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. nature. 2012;486(7402):222.

84. CARDIOBIOME - Era7 Bioinformatics [Internet]. [cited 2019 Feb 11]. Available from: https://era7bioinformatics.com/en/page.cfm?id=2620&title=cardiobiome

85. Pediatric Metabolism and Microbiome Repository - Full Text View - ClinicalTrials.gov [Internet]. [cited 2019 Feb 11]. Available from: https://clinicaltrials.gov/ct2/show/NCT02959034

86. BioLINCC: The Lung HIV Microbiome Project (LHMP) [Internet]. [cited 2019 Feb 11]. Available from: https://biolincc.nhlbi.nih.gov/studies/lhmp/

87. NASA - Study of the Impact of Long-Term Space Travel on the Astronauts' Microbiome [Internet]. [cited 2019 Feb 11]. Available from: https://www.nasa.gov/mission_pages/station/research/experiments/1010.html

88. The Michigan Microbiome Project | University of Michigan | Center for Microbial Systems [Internet]. [cited 2019 Feb 11]. Available from: https://microbe.med.umich.edu/research/michigan-microbiome-project

89. Our Science - uBiome - uBiome Board [Internet]. [cited 2019 Feb 11]. Available from: https://ubiome.com/science/

90. HOMD :: Human Oral Microbiome Database [Internet]. [cited 2019 Feb 11]. Available from: http://www.homd.org/index.php

91. HPMCD: Human Pan Microbial Communities Database [Internet]. [cited 2019 Feb 11]. Available from: http://www.hpmcd.org/

92. curatedMetagenomicData | curatedMetagenomicData [Internet]. [cited 2019 Feb 11]. Available from: http://waldronlab.io/curatedMetagenomicData/

93. European Nucleotide Archive < EMBL-EBI [Internet]. [cited 2019 Feb 11]. Available from: https://www.ebi.ac.uk/ena

94. EMBL-EBI Mg. MGnify home page > EMBL-EBI [Internet]. MGnify. [cited 2019 Feb 11]. Available from: https://ebi.ac.uk/metagenomics/beta/

95. MG-RAST [Internet]. [cited 2019 Feb 11]. Available from: http://www.mg-rast.org/

96. Quinn RA, Navas-Molina JA, Hyde ER, Song SJ, Vázquez-Baeza Y, Humphrey G, et al. From sample to multi-omics conclusions in under 48 hours. MSystems. 2016;1(2):e00038-16.

97. De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Storia A, Laghi L, et al. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut. 2016;65(11):1812–21.

98. Shim J-S, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiology and health. 2014;36.

99. Agriculture USD of. USDA National nutrient database for standard reference, release 28. Agricultural Research Service; 2010.

100. Canada H. Canadian nutrient file. Government of Canada Ottawa,, Canada; 2010.

101. Sánchez B, Delgado S, Blanco-Míguez A, Lourenço A, Gueimonde M, Margolles A. Probiotics, gut microbiota, and their influence on host health and disease. Molecular nutrition & food research. 2017;61(1):1600240.

102. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018;359(6380):1151–6.

103. Barabási A-L, Menichetti G, Loscalzo J. The unmapped chemical complexity of our diet. Nature Food. 2019;1–5.

104. Hall AB, Tolonen AC, Xavier RJ. Human genetic variation and the gut microbiome in disease. Nature Reviews Genetics. 2017;18(11):690.

105. Thaiss CA, Zeevi D, Levy M, Zilberman-Schapira G, Suez J, Tengeler AC, et al. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. Cell. 2014;159(3):514–29.

106. Cox LM, Blaser MJ. Antibiotics in early life and obesity. Nature Reviews Endocrinology. 2015;11(3):182.

107. Gopalakrishnan V, Spencer C, Nezi L, Reuben A, Andrews M, Karpinets T, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science. 2018;359(6371):97–103.

108. Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, et al. Association of host genome with intestinal microbial composition in a large healthy cohort. Nature genetics. 2016;48(11):1413.

109. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, et al. The effect of host genetics on the gut microbiome. Nature genetics. 2016;48(11):1407.

110. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5(6):e1000529.

111. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic determinants of the gut microbiome in UK twins. Cell host & microbe. 2016;19(5):731–43.

112. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555(7695):210.

113. Schwartz S, Friedberg I, Ivanov IV, Davidson LA, Goldsby JS, Dahl DB, et al. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. Genome biology. 2012;13(4):r32.

114. de Steenhuijsen Piters WA, Heinonen S, Hasrat R, Bunsow E, Smith B, Suarez-Arrabal M-C, et al. Nasopharyngeal microbiota, host transcriptome, and disease severity in children with respiratory syncytial virus infection. American journal of respiratory and critical care medicine. 2016;194(9):1104–15.

115. Thaiss CA, Levy M, Korem T, Dohnalová L, Shapiro H, Jaitin DA, et al. Microbiota diurnal rhythmicity programs host transcriptome oscillations. Cell. 2016;167(6):1495–510.

116. Pan W-H, Sommer F, Falk-Paulsen M, Ulas T, Best P, Fazio A, et al. Exposure to the gut microbiota drives distinct methylome and transcriptome changes in intestinal epithelial cells during postnatal development. Genome medicine. 2018;10(1):27.

117. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015;47(9):1091.

118. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic acids research. 2017;46(D1):D633–9.

119. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2011;40(D1):D109–14.

120. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic acids research. 2017;46(D1):D649–55.

121. Antonazzo G, Attrill H, Brown N, Marygold SJ, McQuilton P, Ponting L, et al. Expansion of the Gene Ontology knowledgebase and resources. 2017;

122. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic variation impacts microbiome composition across human body sites. Genome biology. 2015;16(1):191.

123. Dobson AJ, Chaston JM, Newell PD, Donahue L, Hermann SL, Sannino DR, et al. Host genetic determinants of microbiota-dependent nutrition revealed by genome-wide analysis of Drosophila melanogaster. Nature communications. 2015;6:6312.

124. Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. Genome-wide association studies of the human gut microbiota. PLoS One. 2015;10(11):e0140301.

125. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of epidemiology. 2016;26(5):330–5.

126. Tyler AD, Smith MI, Silverberg MS. Analyzing the human microbiome: a “how to” guide for physicians. The American journal of gastroenterology. 2014;109(7):983.

127. Zhou Y-H, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in Genetics. 2019;10:579.

128. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of Machine Learning in Microbiology. Frontiers in microbiology. 2019;10.

129. Topçuoğlu BD, Lesniak NA, Ruffin M, Wiens J, Schloss PD. Effective application of machine learning to microbiome-based classification problems. BioRxiv. 2019;816090.

130. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335.

131. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology. 2009;75(23):7537–41.

132. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods. 2013;10(10):996.

133. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology. 2019;37(8):852–7.

134. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):R46.

135. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. PLoS computational biology. 2016;12(6):e1004957.

136. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature methods. 2015;12(10):902.

137. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS computational biology. 2012;8(6):e1002358.

138. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

139. Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson GW. Fast, accurate error-correction of amplicon pyrosequences using Acacia. Nature methods. 2012;9(5):425.

140. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv. 2016;081257.

141. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16):2194–200.

142. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;2(2):e00191-16.

143. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581.

144. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357.

145. Agarwala R, Morgulis A. BMTagger: Best Match Tagger for Removing Human Reads from Metagenomics Datasets [Internet]. 2011. Available from: ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/

146. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS one. 2011;6(3):e17288.

147. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6.

148. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772–80.

149. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.

150. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017;gr-213959.

151. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2014;12(1):59.

152. DeSantis T, Hugenholtz P, Keller K, Brodie E, Larsen N, Piceno Y, et al. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic acids research. 2006;34(suppl_2):W394–9.

153. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

154. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and environmental microbiology. 2005;71(3):1501–6.

155. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016;32(16):2520–3.

156. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nature methods. 2017;14(11):1023.

157. Wang J, Kurilshikov A, Radjabzadeh D, Turpin W, Croitoru K, Bonder MJ, et al. Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. BioMed Central; 2018.

158. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS computational biology. 2016;12(7):e1004977.

159. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature communications. 2017;8(1):1784.

160. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nature medicine. 2019;1.

161. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature medicine. 2019;1.

162. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017;8:2224.

163. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology. 2014;10(4):e1003531.

164. Willis AD. Rarefaction, alpha diversity, and statistics. Frontiers in microbiology. 2019;10:2407.

165. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics and Its Application. 2015;2:73–94.

166. Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985;125(1):1–15.

167. Zhu X. Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison. 2(3):4.

168. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010;10(22):1345–59.

169. Prehn-Kristensen A, Zimmermann A, Tittmann L, Lieb W, Schreiber S, Baving L, et al. Reduced microbiome alpha diversity in young patients with ADHD. PloS one. 2018;13(7):e0200728.

170. Davenport ER, Mizrahi-Man O, Michelini K, Barreiro LB, Ober C, Gilad Y. Seasonal variation in human gut microbiome composition. PloS one. 2014;9(3):e90731.

171. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995;289–300.

172. Rice WR. Analyzing tables of statistical tests. Evolution. 1989;43(1):223–5.

173. Johnson RA, Wichern DW. Multivariate analysis. Encyclopedia of Statistical Sciences. 2004;8.

174. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral ecology. 2001;26(1):32–46.

175. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Australian journal of ecology. 1993;18(1):117–43.

176. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proceedings of the National Academy of Sciences. 2010;107(33):14691–6.

177. Zhang C, Yin A, Li H, Wang R, Wu G, Shen J, et al. Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. EBioMedicine. 2015;2(8):968–84.

178. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, et al. A microbial signature for Crohn's disease. Gut. 2017;gutjnl-2016.

179. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecological monographs. 1957;27(4):325–49.

180. Spearman C. The proof and measurement of association between two things. The American journal of psychology. 1904;15(1):72–101.

181. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal. 2016;10(7):1669.

182. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012;8(9):e1002687.

183. Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics. 2006;22(20):2532–8.

184. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014;159(4):789–99.

185. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 2017;6:e21887.

186. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016;32(17):2611–7.

187. Ospina R, Ferrari SL. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis. 2012;56(6):1609–23.

188. Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. Journal of Computational Biology. 2016;23(2):102–10.

189. Washburne AD, Morton JT, Sanders J, McDonald D, Zhu Q, Oliverio AM, et al. Methods for phylogenetic analysis of microbiome data. Nature Microbiology. 2018;3(6):652.

190. Bradley PH, Nayfach S, Pollard KS. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. PLoS computational biology. 2018;14(8):e1006242.

191. Hotelling H. Relations between two sets of variates. In: Breakthroughs in statistics. Springer; 1992. p. 162–90.

192. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34.

193. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell host & microbe. 2015;17(2):260–73.

194. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

195. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters. 1999;9(3):293–300.

196. Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2001.

197. Kim M, Eetemadi A, Tagkopoulos I. DeepPep: Deep proteome inference from peptide profiles. PLoS computational biology. 2017;13(9):e1005661.

198. Singh R, Lanchantin J, Robins G, Qi Y. Deepchrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–48.

199. Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nature communications. 2016;7:13090.

200. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436.

201. Shavitt I, Segal E. Regularization Learning Networks: Deep Learning for Tabular Datasets. In: Advances in Neural Information Processing Systems. 2018. p. 1384–94.

202. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nature biotechnology. 2014;32(8):834.

203. Rahman SF, Olm MR, Morowitz MJ, Banfield JF. Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. MSystems. 2018;3(1):e00123-17.

204. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdisciplinary reviews: computational statistics. 2013;5(2):149–79.

205. Robertson RC, Kaliannan K, Strain CR, Ross RP, Stanton C, Kang JX. Maternal omega-3 fatty acids regulate offspring obesity through persistent modulation of gut microbiota. Microbiome. 2018;6(1):95.

206. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. The ISME journal. 2016;10(3):707.

207. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology. 2012;13(9):R79.

208. Yong AG, Pearce S. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in quantitative methods for psychology. 2013;9(2):79–94.

209. Sordillo JE, Korrick S, Laranjo N, Carey V, Weinstock GM, Gold DR, et al. Association of the Infant Gut Microbiome With Early Childhood Neurodevelopmental Outcomes: An Ancillary Study to the VDAART Randomized Clinical Trial. JAMA network open. 2019;2(3):e190905–e190905.

210. Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of educational research. 2006;99(6):323–38.

211. Moossavi S, Sepehri S, Robertson B, Bode L, Goruk S, Field CJ, et al. Composition and variation of the human milk microbiota are influenced by maternal and early-life factors. Cell host & microbe. 2019;25(2):324–35.

212. Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). Journal of statistical software. 2012;48(2):1–36.

213. Lê S, Josse J, Husson F, others. FactoMineR: an R package for multivariate analysis. Journal of statistical software. 2008;25(1):1–18.

214. IBM Corp N. IBM SPSS statistics for windows. Version 220. 2013;

215. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27.

216. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature methods. 2013;10(12):1200.

217. Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. Biostatistics. 2006;8(1):86–100.

218. Bartenhagen C, Klein H-U, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC bioinformatics. 2010;11(1):567.

219. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, et al. Microbiome interactions shape host fitness. Proceedings of the National Academy of Sciences. 2018;115(51):E11951–60.

220. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. nature. 2011;473(7346):174.

221. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS computational biology. 2013;9(1):e1002863.

222. Kaufman L, Rousseeuw P. Clustering by means of medoids. North-Holland; 1987.

223. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(1):86–97.

224. Tibshirani R, Walther G. Cluster validation by prediction strength. Journal of Computational and Graphical Statistics. 2005;14(3):511–28.

225. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987;20:53–65.

226. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods. 1974;3(1):1–27.

227. Hildebrand F, Nguyen TLA, Brinkman B, Yunta RG, Cauwe B, Vandenabeele P, et al. Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. Genome biology. 2013;14(1):R4.

228. Costea PI, Hildebrand F, Manimozhiyan A, Bäckhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nature microbiology. 2018;3(1):8.

229. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;19(5):299.

230. Kim M, Tagkopoulos I. Data integration and predictive modeling methods for multi-omics datasets. Molecular omics. 2018;14(1):8–25.

231. Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. Frontiers in genetics. 2019;10.

232. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Frontiers in genetics. 2017;8:84.

233. Grice EA, Segre JA. The human microbiome: our second genome. Annual review of genomics and human genetics. 2012;13:151–70.

234. Gentile CL, Weir TL. The gut microbiota at the intersection of diet and human health. Science. 2018;362(6416):776–80.

235. Tran TNT, Atas M, Felfernig A, Stettinger M. An overview of recommender systems in the healthy food domain. Journal of Intelligent Information Systems. 2017;1–26.

236. Suphavilai C, Bertrand D, Nagarajan N. Predicting Cancer Drug Response Using a Recommender System. Bioinformatics.

237. Burke R. Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction. 2002;12(4):331–70.

238. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, et al. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic acids research. 2018;47(D1):D614–24.

239. Bauer E, Zimmermann J, Baldini F, Thiele I, Kaleta C. BacArena: individual-based metabolic modeling of heterogeneous microbes in complex communities. PLoS computational biology. 2017;13(5):e1005544.

240. Magnúsdóttir S, Thiele I. Modeling metabolism of the human gut microbiome. Current opinion in biotechnology. 2018;51:90–6.

241. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature biotechnology. 2019;37(2):179.

242. Tramontano M, Andrejev S, Pruteanu M, Klünemann M, Kuhn M, Galardini M, et al. Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. Nature microbiology. 2018;3(4):514.

243. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature biotechnology. 2017;35(1):81.

244. Lops P, De Gemmis M, Semeraro G. Content-based recommender systems: State of the art and trends. In: Recommender systems handbook. Springer; 2011. p. 73–105.

245. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Advances in artificial intelligence. 2009;2009.

246. Ekstrand MD, Riedl JT, Konstan JA, others. Collaborative filtering recommender systems. Foundations and Trends® in Human–Computer Interaction. 2011;4(2):81–173.

247. Fankhauser M, Moser C, Nyfeler T. Patents as early indicators of technology and investment trends: analyzing the microbiome space as a case study. Frontiers in bioengineering and biotechnology. 2018;6:84.

248. Green JM, Barratt MJ, Kinch M, Gordon JI. Food and microbiota in the FDA regulatory framework. Science. 2017;357(6346):39–40.

249. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32(24):3823–5.

250. Hitch TC, Creevey CJ. Spherical: an iterative workflow for assembling metagenomic datasets. BMC bioinformatics. 2018;19(1):20.

251. Lovell RM, Ford AC. Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. Clinical gastroenterology and hepatology. 2012;10(7):712–21.

252. Chey WD, Kurlander J, Eswaran S. Irritable bowel syndrome: a clinical review. Jama. 2015;313(9):949–58.

253. Ford AC, Moayyedi P, Chey WD, Harris LA, Lacy BE, Saito YA, et al. American College of Gastroenterology monograph on management of irritable bowel syndrome. American Journal of Gastroenterology. 2018;113:1–18.

254. Moayyedi P, Andrews CN, MacQueen G, Korownyk C, Marsiglio M, Graff L, et al. Canadian Association of Gastroenterology clinical practice guideline for the management of irritable bowel syndrome (IBS). Journal of the Canadian Association of Gastroenterology. 2019;2(1):6–29.

255. Staudacher HM, Whelan K. The low FODMAP diet: recent advances in understanding its mechanisms and efficacy in IBS. Gut. 2017;66(8):1517–27.

256. Bennet SM, Böhn L, Störsrud S, Liljebo T, Collin L, Lindfors P, et al. Multivariate modelling of faecal bacterial profiles of patients with IBS predicts responsiveness to a diet low in FODMAPs. Gut. 2018;67(5):872–81.

257. Rossi M, Aggio R, Staudacher HM, Lomer MC, Lindsay JO, Irving P, et al. Volatile organic compounds in feces associate with response to dietary intervention in patients with irritable bowel syndrome. Clinical Gastroenterology and Hepatology. 2018;16(3):385–91.

258. Wilder-Smith C, Olesen SS, Materna A, Drewes A. Predictors of response to a low-FODMAP diet in patients with functional gastrointestinal disorders and lactose or fructose intolerance. Alimentary pharmacology & therapeutics. 2017;45(8):1094–106.

259. Kalantar-Zadeh K, Berean KJ, Burgell RE, Muir JG, Gibson PR. Intestinal gases: influence on gut disorders and the role of dietary manipulations. Nature Reviews Gastroenterology & Hepatology. 2019;1–15.

260. Casen C, Vebø H, Sekelja M, Hegge F, Karlsson M, Ciemniejewska E, et al. Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD. Alimentary pharmacology & therapeutics. 2015;42(1):71–83.

261. Chumpitazi BP, Hollister EB, Oezguen N, Tsai CM, McMeans AR, Luna RA, et al. Gut microbiota influences low fermentable substrate diet efficacy in children with irritable bowel syndrome. Gut microbes. 2014;5(2):165–75.

262. Chumpitazi BP, Cope JL, Hollister EB, Tsai CM, McMeans AR, Luna RA, et al. Randomised clinical trial: gut microbiome biomarkers are associated with clinical response to a low FODMAP diet in children with the irritable bowel syndrome. Alimentary pharmacology & therapeutics. 2015;42(4):418–27.

263. Harvie RM, Chisholm AW, Bisanz JE, Burton JP, Herbison P, Schultz K, et al. Long-term irritable bowel syndrome symptom control with reintroduction of selected FODMAPs. World journal of gastroenterology. 2017;23(25):4632.

264. McIntosh K, Reed DE, Schneider T, Dang F, Keshteli AH, De Palma G, et al. FODMAPs alter symptoms and the metabolome of patients with IBS: a randomised controlled trial. Gut. 2017;66(7):1241–51.

265. Staudacher HM, Lomer MC, Farquharson FM, Louis P, Fava F, Franciosi E, et al. A diet low in FODMAPs reduces symptoms in patients with irritable bowel syndrome and a probiotic restores bifidobacterium species: a randomized controlled trial. Gastroenterology. 2017;153(4):936–47.

266. Sloan TJ, Jalanka J, Major GA, Krishnasamy S, Pritchard S, Abdelrazig S, et al. A low FODMAP diet is associated with changes in the microbiota and reduction in breath hydrogen but not colonic volume in healthy subjects. PloS one. 2018;13(7):e0201410.

267. Schumann D, Langhorst J, Dobos G, Cramer H. Randomised clinical trial: yoga vs a low-FODMAP diet in patients with irritable bowel syndrome. Alimentary pharmacology & therapeutics. 2018;47(2):203–11.

268. Valeur J, Sm\aastuen MC, Knudsen T, Lied GA, Røseth AG. Exploring gut microbiota composition as an indicator of clinical response to dietary FODMAP restriction in patients with irritable bowel syndrome. Digestive Diseases and Sciences. 2018;63(2):429–36.

269. Lyra A, Hillilä M, Huttunen T, Männikkö S, Taalikka M, Tennilä J, et al. Irritable bowel syndrome symptom severity improves equally with probiotic and placebo. World journal of gastroenterology. 2016;22(48):10631.

270. Kaptchuk TJ, Friedlander E, Kelley JM, Sanchez MN, Kokkotou E, Singer JP, et al. Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PloS one. 2010;5(12):e15591.

271. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581–3.

272. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2012;41(D1):D590–6.

273. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal. 2012;6(3):610.

274. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017;8:2224.

275. Parthasarathy G, Chen J, Chen X, Chia N, O'Connor HM, Wolf PG, et al. Relationship between microbiota of the colonic mucosa vs feces and symptoms, colonic transit, and methane production in female patients with chronic constipation. Gastroenterology. 2016;150(2):367–79.

276. Sigg C. nsprcomp v0. 5.1-2.

277. Sigg CD, Buhmann JM. Expectation-maximization for sparse and non-negative PCA. In: Proceedings of the 25th international conference on Machine learning. 2008. p. 960–7.

278. Triantafyllou K, Chang C, Pimentel M. Methanogens, methane and gastrointestinal motility. Journal of neurogastroenterology and motility. 2014;20(1):31.

279. Levitt MD. Production and excretion of hydrogen gas in man. New England Journal of Medicine. 1969;281(3):122–7.

280. Thiele JH, Zeikus JG. Control of interspecies electron flow during anaerobic digestion: significance of formate transfer versus hydrogen transfer during syntrophic methanogenesis in flocs. Applied and environmental Microbiology. 1988;54(1):20–9.

281. Yamamura R, Nakamura K, Kitada N, Aizawa T, Shimizu Y, Nakamura K, et al. Associations of gut microbiota, dietary intake, and serum short-chain fatty acids with fecal short-chain fatty acids. Bioscience of microbiota, food and health. 2019;19–010.

282. Kim G, Deepinder F, Morales W, Hwang L, Weitsman S, Chang C, et al. Methanobrevibacter smithii is the predominant methanogen in patients with constipation-predominant IBS and methane on breath. Digestive diseases and sciences. 2012;57(12):3213–8.

283. Gill P, Van Zelm M, Muir J, Gibson P. Short chain fatty acids as potential therapeutic agents in human gastrointestinal and inflammatory disorders. Alimentary pharmacology & therapeutics. 2018;48(1):15–34.

284. Kolodziejczyk AA, Zheng D, Elinav E. Diet–microbiota interactions and personalized nutrition. Nature Reviews Microbiology. 2019;1–12.

285. Smith NW, Shorten PR, Altermann EH, Roy NC, McNabb WC. Hydrogen cross-feeders of the human gastrointestinal tract. Gut Microbes. 2019;10(3):270–88.

286. Maslowski KM, Vieira AT, Ng A, Kranich J, Sierro F, Yu D, et al. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature. 2009;461(7268):1282–6.

287. Pozuelo M, Panda S, Santiago A, Mendez S, Accarino A, Santos J, et al. Reduction of butyrate-and methane-producing microorganisms in patients with Irritable Bowel Syndrome. Scientific reports. 2015;5:12693.

288. Nagpal R, Wang S, Ahmadi S, Hayes J, Gagliano J, Subashchandrabose S, et al. Human-origin probiotic cocktail increases short-chain fatty acid production via modulation of mice and human gut microbiome. Scientific reports. 2018;8(1):1–15.

289. Pedersen N, Andersen NN, Végh Z, Jensen L, Ankersen DV, Felding M, et al. Ehealth: low FODMAP diet vs Lactobacillus rhamnosus GG in irritable bowel syndrome. World Journal of Gastroenterology: WJG. 2014;20(43):16215.

290. Momčilović S, Cantacessi C, Arsić-Arsenijević V, Otranto D, Tasić-Otašević S. Rapid diagnosis of parasitic diseases: current scenario and future needs. Clinical Microbiology and Infection. 2019;25(3):290–309.

291. Yao CK, Tuck CJ. The clinical value of breath hydrogen testing. Journal of Gastroenterology and Hepatology. 2017;32:20–2.

292. Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food. 2018;2(1):1–10.

293. Youn J, Naravane T, Tagkopoulos I. Using Word Embeddings to Learn a Better Food Ontology. Frontiers in Artificial Intelligence. 2020;3:93.

294. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. Molecular systems biology. 2014;10(7):735.

295. Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nature communications. 2016;7:13090.

296. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. Cell. 2015;161(5):971–87.

297. Abhyankar W, Stelder S, de Koning L, de Koster C, Brul S. 'Omics' for microbial food stability: Proteomics for the development of predictive models for bacterial spore stress survival and outgrowth. International journal of food microbiology. 2017;240:11–8.

298. Aucoin HR, Gardner J, Boyle NR. Omics in Chlamydomonas for biofuel production. In: Lipids in plant and algae development. Springer; 2016. p. 447–69.

299. Gonzalez de Castro D, Clarke P, Al-Lazikani B, Workman P. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. Clinical Pharmacology & Therapeutics. 2013;93(3):252–9.

300. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery. 2016;15(7):473.

301. Tachibana C. Transcriptomics today: Microarrays, RNA-seq, and more. Science. 2015;349(6247):544–6.

302. Ay A, Arnosti DN. Mathematical modeling of gene expression: a guide for the perplexed biologist. Critical reviews in biochemistry and molecular biology. 2011;46(2):137–51.

303. Dragosits M, Nicklas D, Tagkopoulos I. A synthetic biology approach to self-regulatory recombinant protein production in Escherichia coli. Journal of biological engineering. 2012;6(1):2.

304. Mahalik S, Sharma AK, Mukherjee KJ. Genome engineering for improved recombinant protein expression in Escherichia coli. Microbial cell factories. 2014;13(1):177.

305. Riglar DT, Silver PA. Engineering bacteria for diagnostic and therapeutic applications. Nature Reviews Microbiology. 2018;16(4):214.

306. Milne CB, Kim P-J, Eddy JA, Price ND. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnology Journal: Healthcare Nutrition Technology. 2009;4(12):1653–70.

307. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436.

308. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436.

309. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics. 2017;

310. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. bioRxiv. 2018;142760.

311. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional regulatory circuits: predicting numbers from alphabets. Science. 2009;325(5939):429–32.

312. VOHRADSKÝ J. Neural network model of gene expression. the FASEB journal. 2001;15(3):846–54.

313. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nature methods. 2018;15(4):290.

314. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;32(12):1832–9.

315. Singh R, Lanchantin J, Robins G, Qi Y. Deepchrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–48.

316. Kim M, Eetemadi A, Tagkopoulos I. DeepPep: Deep proteome inference from peptide profiles. PLoS computational biology. 2017;13(9):e1005661.

317. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–105.

318. Watters N, Tacchetti A, Weber T, Pascanu R, Battaglia P, Zoran D. Visual interaction networks. arXiv preprint arXiv:170601433. 2017;

319. Kansky K, Silver T, Mély DA, Eldawy M, Lázaro-Gredilla M, Lou X, et al. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. arXiv preprint arXiv:170604317. 2017;

320. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome biology. 2006;7(5):R36.

321. Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, et al. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. Proceedings of the National Academy of Sciences. 2017;114(38):10286–91.

322. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: Computer vision and pattern recognition, 2009 CVPR 2009 IEEE conference on. 2009. p. 248–55.

323. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic acids research. 2015;44(D1):D133–43.

324. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. science. 2005;307(5717):1962–5.

325. Møller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural networks. 1993;6(4):525–33.

326. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;267–88.

327. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural networks. 1989;2(5):359–66.

328. Pineda FJ. Generalization of back-propagation to recurrent neural networks. Physical review letters. 1987;59(19):2229.

329. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 1997;45(11):2673–81.

330. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012;13(Feb):281–305.

331. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural computation. 1989;1(2):270–80.

332. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. 2012;4(2):26–31.

333. Carrera J, Rodrigo G, Jaramillo A. Model-based redesign of global transcription regulation. Nucleic acids research. 2009;37(5):e38–e38.

334. Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, Sweet L, et al. The Mycobacterium tuberculosis regulatory network and hypoxia. Nature. 2013;499(7457):178.

335. Kvålseth TO. Cautionary note about r 2. The American Statistician. 1985;39(4):279–85.

336. Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics. 2011;27(16):2263–70.

337. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the national academy of sciences. 2010;107(14):6286–91.

338. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nature methods. 2012;9(8):796.

339. Day WH, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification. 1984;1(1):7–24.

340. Long Z, Quaife B, Salman H, Oltvai ZN. Cell-cell communication enhances bacterial chemotaxis toward external attractants. Scientific reports. 2017;7(1):12855.

341. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research. 2016;45(D1):D353–61.

342. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.

343. Irrthum A, Wehenkel L, Geurts P, others. Inferring regulatory networks from expression data using tree-based methods. PloS one. 2010;5(9):e12776.

344. Chan SS-K, Kyba M. What is a master regulator? Journal of stem cell research & therapy. 2013;3.

345. Brandman O, Meyer T. Feedback loops shape cellular signals in space and time. Science. 2008;322(5900):390–5.

346. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, et al. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science. 2006;314(5797):267–267.

347. Roden DM, George Jr AL. The genetic basis of variability in drug responses. Nature reviews Drug discovery. 2002;1(1):37–44.

348. Garcia-Perez I, Posma JM, Chambers ES, Mathers JC, Draper J, Beckmann M, et al. Dietary metabotype modelling predicts individual responses to dietary interventions. Nature Food. 2020;1(6):355–64.

349. Davis CL, Tomporowski PD, McDowell JE, Austin BP, Miller PH, Yanasak NE, et al. Exercise improves executive function and achievement and alters brain activation in overweight children: a randomized, controlled trial. Health psychology. 2011;30(1):91.

350. Law M, Morris J, Wald N. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. Bmj. 2009;338.

351. Lackner JM, Mesmer C, Morley S, Dowzer C, Hamilton S. Psychological treatments for irritable bowel syndrome: a systematic review and meta-analysis. Journal of consulting and clinical psychology. 2004;72(6):1100.

352. Berkowitz L, Schultz BM, Salazar GA, Pardo-Roa C, Sebastián VP, Álvarez-Lobos MM, et al. Impact of cigarette smoking on the gastrointestinal tract inflammation: opposing effects in Crohn's disease and ulcerative colitis. Frontiers in immunology. 2018;9:74.

353. Arora T, Taheri S. Sleep optimization and diabetes control: a review of the literature. Diabetes Therapy. 2015;6(4):425–68.

354. Longo VD, Mattson MP. Fasting: molecular mechanisms and clinical applications. Cell metabolism. 2014;19(2):181–92.

355. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nature reviews Drug discovery. 2003;2(7):566–80.

356. Hampel H, Frank R, Broich K, Teipel SJ, Katz RG, Hardy J, et al. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. Nature reviews Drug discovery. 2010;9(7):560–74.

357. Baker M. In biomarkers we trust? Nature biotechnology. 2005;23(3):297–304.

358. Ray P, Manach YL, Riou B, Houle TT, Warner DS. Statistical evaluation of a biomarker. The Journal of the American Society of Anesthesiologists. 2010;112(4):1023–40.

359. Nowak-Węgrzyn A, Assa'ad AH, Bahna SL, Bock SA, Sicherer SH, Teuber SS, et al. Work Group report: oral food challenge testing. Journal of Allergy and Clinical Immunology. 2009;123(6):S365–83.

360. Kravitz R, Duan N, Eslick I, Gabler N, Kaplan H, Larson E, et al. Design and implementation of N-of-1 trials: a user's guide. Agency for healthcare research and quality, US Department of Health and Human Services. 2014;

361. Nsouli T, Nsouli S, Linde R, O'mara F, Scanlon R, Bellanti J. Role of food allergy in serous otitis media. Annals of allergy. 1994;73(3):215–9.

362. Drisko J, Bischoff B, Hall M, McCallum R. Treating irritable bowel syndrome with a food elimination diet followed by food challenge and probiotics. Journal of the American College of Nutrition. 2006;25(6):514–22.

363. Kagalwalla AF, Sentongo TA, Ritz S, Hess T, Nelson SP, Emerick KM, et al. Effect of six-food elimination diet on clinical and histologic outcomes in eosinophilic esophagitis. Clinical gastroenterology and hepatology. 2006;4(9):1097–102.

364. Pelsser LM, Frankena K, Toorman J, Savelkoul HF, Pereira RR, Buitelaar JK. A randomised controlled trial into the effects of food on ADHD. European child & adolescent psychiatry. 2009;18(1):12–9.

365. Kaplan HC, Opipari-Arrigan L, Schmid CH, Schuler CL, Saeed S, Braly KL, et al. Evaluating the Comparative Effectiveness of Two Diets in Pediatric Inflammatory Bowel Disease: A Study Protocol for a Series of N-of-1 Trials. In: Healthcare. Multidisciplinary Digital Publishing Institute; 2019. p. 129.

366. Tian Y, Ma Y, Fu Y, Zheng J-S. Application of n-of-1 clinical trials in personalized nutrition research: a trial protocol for Westlake N-of-1 Trials for Macronutrient Intake (WE-MACNUTR). Current Developments in Nutrition. 2020;4(9):nzaa143.

367. Chevance G, Baretta D, Golaszewski N, Takemoto M, Shrestha S, Jain S, et al. Goal setting and achievement for walking: A series of N-of-1 digital interventions. Health Psychology. 2020;

368. Aldridge M, Johnson O, Scarlett J. Group Testing: An Information Theory Perspective. Foundations and Trends® in Communications and Information Theory. 2019;15(3–4):196–392.

369. Böhn L, Störsrud S, Törnblom H, Bengtsson U, Simrén M. Self-reported food-related gastrointestinal symptoms in IBS are common and associated with more severe symptoms and reduced quality of life. American Journal of Gastroenterology. 2013;108(5):634–41.

370. Vierk KA, Koehler KM, Fein SB, Street DA. Prevalence of self-reported food allergy in American adults and use of food labels. Journal of allergy and clinical immunology. 2007;119(6):1504–10.

371. Coja-Oghlan A, Gebhard O, Hahn-Klimroth M, Loick P. Optimal group testing. In: Conference on Learning Theory. PMLR; 2020. p. 1374–88.

372. Wang X, Rai N, Merchel Piovesan Pereira B, Eetemadi A, Tagkopoulos I. Accelerated knowledge discovery from omics data by optimal experimental design. Nature Communications. 2020 Oct 6;11(1):5026.

Appendix A

A.1. Constrained Adaptive Group Testing

The core of the CAGT algorithm which is implemented by the CAGT model, is illustrated by an example in **Figure A.1,** and fully described in **Algorithm A.1** The CAGT model, is encoded by a binary tree connecting a set of nodes denoted by g_i. Each g_i, represents a set of LIs with unknown potency, as well as l and h integers that bound the number of potent LIs in the set. The tree represents nested sets of LIs where the LIs of a non-leaf node are comprised of the LIs of its children, and sibling nodes are disjoint (i.e. have no shared LI). In step1, the $next\_round$ function iterates through the leaf nodes and simulates the next round using each node leading into alternative trees. The leaf node g that leads into tree pairs (two trees per g due to two possible potencies) with best potential (i.e. minimum $|R|$), is chosen for the next round. The optimal number of LIs for next round ($|R_i|$) is then identified using the CAGT catalog given $|g|$ and the corresponding l and h from g. Any subset of size $|R_i|$ from g can be selected as R_i for the next round, however we used the first $|R_i|$ LIs in g for simplicity. In step2, the potency of R_i is determined as r_i by the individual. In step3, the f function uses R_i and r_i to split the g node and subsequently revise the tree which can lead into updated V_0 and V_1. Revisions are made in the tree using **Table A.1**, as long as there is a node that meets a revision criterion. Revising one node, can trigger revisions across the tree using the "trigger revision" column of **Table A.1,** which names the nodes that should be subsequently verified against the criteria enabling an efficient method for finding all the nodes that need verification (See $revise\_tree$ function in **Algorithm A.1**).

Algorithm A.1 The CAGT model and functions.

Inputs: The CAGT model is initialized by the set of candidate LIs (LI), the low and high thresholds (l and h) that bound the number of potent LIs. The model also relies on a prebuilt CAGT catalog. The f function takes $R_i \subseteq LI$ that represent the LIs followed by the individual in the last round, and $r_i \in \{0,1\}$ that represent the corresponding potency.
Outputs: The $next\_round$ function returns $R_i \subseteq LI$ to be followed by the individual. The f function returns set of impotent LIs V_0 and potent LIs V_1 identified by the Algorithm A.o far.

```
class NGroup
  // A node g in the CAGT model tree which contains:
  //   V: a set of LIs, l: minimum number of potent LIs,
  //   h: maximum number of potent LIs, parent: an NGroup, and
  //   children: a set of NGroup
  function next_round(g):
    optim_LI_count ← CATALOG.optim_size(|g.V|, g.l, g.h)
    return g.V[1: optim_LI_count]

  function max_rounds(g):
    return CATALOG.max_rounds(|g.V|, g.l, g.h)

class CAGT_Model:
  function __init__(model, LI, l, h):
    model.root ← NGroup(LI, l, h)
    V0 ← {}; V1 ← {};

  function next_round(model):
    best_g ← None
    for g in model.leaves():
      if model.max_rounds_if(g) < model.max_rounds_if(best_g):
        best_g ← g
    model.last_g ← best_g
    return best_g.next_round()

  function f(model, Ri, ri):
    model.split_g(Ri, ri, model.last_g)
```

```
    model.revise_tree()
    return [model.V_0, model.V_1]

  function split_g(model, R_i, r_i, g):
    if r_i is 0:
      model.V_0 ← model.V_0 ∪ R_i
      // Update g:
      g.V ← g.V \ R_i
      g.h ← min(|g.V|, g.V.h)
      g.l ← min(g.l, g.h)
    else:
      // Create new NGroups under gl:
      g1 ← NGroup(parent ← g, V ← R_i)
      g2 ← NGroup(parent ← g, V ← g.V \ R_i)
      g.children ← [g1, g2]
      // Update bounds:
      g1.h ← min(g.h, |R_i|)
      g1.l ← min(|g1.V|, max(1, g.l − |g2.V|))
      g2.h ← g.h − g1.l
      g2.l ← min(|g2.V|, max(0, g.l − g1.h))

  function max_rounds_if(model, g):
    n_if_potent ← model.max_rounds_hypothetical(g, 1)
    n_if_impotent ← model.max_rounds_hypothetical(g, 0)
    return max(n_if_potent, n_if_impotent)

  function max_rounds_hypothetical(model, g, potency):
    // Calculate the maximum number of rounds assuming g
    // is used in the next round and the potency is
    // determined.
    h_model ← deepcopy(model)
    h_model.last_g ← g
    h_model.f(g.next_round(), potency)
    return h_model.max_rounds()
```

```
48:   function max_rounds(model):
49:     T ← model.leaves()
50:     return Σ_{g∈T} g.max_rounds()

51:   function revise_tree(model):
52:     G_revise ← {model.last_g}
53:     do: // Revise while still needed
54:       g ← pop(G_revise)
55:       G_revise ← G_revise ∪ model.revise_g(g)
56:     while |G_revise|>0

57:   function revise_g(model, g):
58:     G_revise ← {}
59:     revision ← model.identify_revision(g)
60:     while revision:
61:       G_revise ← G_revise ∪ revision.apply()
62:       revision ← model.identify_revision(g)
63:     return G_revise

64:   function identify_revision(model, g):
      // Identify a revision for g from the revision table (Table
   A.1).
```

Table A.1 Criteria used to revise the CAGT tree in each round. Once a node g satisfies a revision "criteria", the corresponding "revision[s]" are applied on g, also leading into revising the "trigger revision" nodes. The LIs under node g are represented by $g.V$, the number of potent LIs in $g.V$ is bounded by $g.l$ and $g.h$.

#	**criteria**	**revision[s]**	**trigger revision**
1	$g.l = \lvert g.V \rvert$	$V_1 \leftarrow V_1 \cup g.V$ $model.remove(g)$	$g.parent$
2	$g.h = 0$	$V_0 \leftarrow V_0 \cup g.V$ $remove(g)$	$g.parent$
3	$g.h \neq 0$ & $g.parent \neq None$ & $g.parent.h = g.sibling().l$	$g.l \leftarrow 0$ $g.h \leftarrow 0$	g
4	$g.l < \sum_{c \in g.children} c.l$	$g.l \leftarrow \sum_{c \in g.children} c.l$	g

A.2. ALO Modules

A.2.1. ALO Module-1: Build the CAGT catalog

The CAGT catalog acts as a lookup table that takes (n, l, h) as input and provides (s,w) as the output (see **Figure A.2 C**). A dynamic programming strategy is used to build the CAGT catalog based on the fact that in each round of CAGT, either the number of LIs in individual leaf nodes decreases, or their corresponding bounds (l and h) tighten. Therefore, to find the optimal (s_y, w_y) of new catalog record y: (n_y, l_y, h_y), if the catalog is built to contain (s_x, w_x) for all the records x:(n_x, l_x, h_x) where $n_x \leq n_y$ and (l_x, h_x) are tighter than (l_y, h_y), then we can run simulations of the CAGT algorithm using all valid s_y values for the initial $next\_round$ of a CAGT run in order to find the s_y that minimizes w_y (see **Algorithm A.2)**. Note that during each simulation, the s_y value is only used for the initial "$next\_round$" call when only the root node exists. All the other catalog lookups rely on optimal (s_x, w_x) values calculated beforehand. For the base cases of this dynamic programming algorithm, we rely on known optimal (s, w) values. When $0 \leq l$ & $h \leq 1$

the binary splitting algorithm is optimal using $s = 2^{\lfloor \log_2 \max(n-1,1) \rfloor}$ leading into $w = \lceil \log_2(n+1-l) \rceil$, while for $h \geq 1 + \frac{n}{2}$ using $s = 1$ is optimal which leads into $w = n$.

Algorithm A.2 Build the CAGT catalog that acts as a lookup table. The catalog takes the triplet (n: #of LIs, l: min #of potent LIs, h: max #of potent LIs) as input and provides the tuple (s: optimal #of LIs for the next round, w: maximum #of rounds) as the output.

Input: MAX_N representing the maximum value of n in the catalog records.
Output: The $CATALOG$ that contains records in the form of "(n, l, h): (s, w)".

```
for n ← 2 to MAX_N:
  for h ← 2 to n/2:
    for l ← h to 0:
      (s, w) ← optim_sw(n, l, h)
      CATALOG[(n, l, h)] ← (s, w)
return CATALOG

// Identify the optimal (s, w) given (n, l, h)
function optim_sw(n, l, h):
  (s_optim, w_optim) ← (1, n)
  for s← 2 to 2^⌊log2 max(n−1,1)⌋:
    // The value of w won't matter for the catalog here since it
    // is used in the initial 'next_round' call of the CAGT
    // Algorithm 1. only and will be replaced once
    // (s_optim, w_optim) is identified.
    w ← −1
    CATALOG[(n, l, h)] ← (s, w)
    for num_potent ← l to h:
      for LI_potencies in all_combinations(n, num_potent):
        num_rounds ← simulate_Algorithm1(LI_potencies)
        w ← max(w, num_rounds)
    if w ≤ w_optim:
      (s_optim, w_optim) ← (s, w)
  return (s_optim, w_optim)
```

```
19: function simulate_Algorithm1(LI_potencies):
      // Simulate Algorithm 1. given the binary vector of
      // LI_potencies and return the number of CAGT rounds.

20: function all_combinations(n, num_potent):
      // Return a list of all binary vectors that have length n and
      // 'num_potent' number of 1s (i.e., n − num_potent 0s).
```

A.2.2. ADO Module-2 (A&B): Create optimal LI sets given their potency probabilities

We use the potency probability of LIs, and the CAGT catalog, in order to create an optimal LI partition (disjoint sets of LIs) to minimize the expected number of CAGT rounds needed while the maximum number of CAGT rounds is also bounded. This is done in two steps (A&B). In step A of this module, the optimal LI partition is created only to minimize the maximum number of CAGT rounds needed given the LI potency probabilities and CAGT catalog. In step B, a new optimal LI partition is created in order to minimize the expected number of CAGT rounds while the maximum number of rounds used remains bounded bellow a user defined threshold.

ALO Module-2.A: Identify the maximum CAGT rounds needed

When the prevalence of potent LIs is high, individual testing will be more efficient than group testing (368). More generally, to achieve optimal performance, LIs can be partitioned into disjoint sets based on their potency probabilities such that group testing is performed independently in each set (and with different group testing parametrizations) (368). We achieve this in ALO *Module-2.A* using **Algorithm A.3** in order to find the optimal maximum number of CAGT rounds that are needed for discovering the potent LIs (w^*) by function $find\_wc^*$. First, we reorder the LIs based on LI potency probability vector $\boldsymbol{p}$ such that $0 \leq p_1 \leq p_2 \leq \cdots \leq p_N < 1$. Second, we calculate h

as the maximum number of potent LIs using the Poisson binomial distribution of $\boldsymbol{p}$ and a confidence threshold of $t \in [0,1]$ (e.g., 0.95). Third, we consider the $best\_partition$ to be all the LIs relating to $\boldsymbol{p}$ with a w^* that is returned from the CAGT catalog for $(|p|,0,h)$. Fourth, we run $find\_wc^*$ for $\boldsymbol{p}_{1\ldots j}$ and $\boldsymbol{p}_{j\ldots|p|}$ recursively $\forall j \in [1,|\boldsymbol{p}|]$ and update $best\_partition$ if the returned partition pairs by the two $find\_wc^*$ calls are better than the current $best\_partition$. To avoid duplicate runs of $find\_wc^*$, we check the input $\boldsymbol{p}$ against a lookup table (named $cache\_wc^*$) in the beginning and only calculate the answer on the first occurrence.

Algorithm A.3 Identify the optimal maximum number of CAGT rounds that are needed for discovering the potent LIs (w^*) using the $find\_wc^*$ function considering the LI potency probability vector $\boldsymbol{p} \in [0,1]^N$ where N is the number of LIs.

Inputs: LI potency probability vector $\boldsymbol{p}$ where $0 \leq p_1 \leq p_2 \leq \cdots \leq p_N < 1$ (i.e. assume that LIs are ordered by their potency probabilities without loss of generality), and a confidence threshold $t \in [0,1]$.
Outputs: Disjoint sets of LIs (amounting to partition called Q^+) such that subjecting individual sets Q to the CAGT algorithm leads into optimal maximum number of CAGT rounds that are needed for discovering the potent LIs (w^*). It will be in the following nested format due to recursion where w^* is the Q_3 from the top nested set, and the non-nested subsets amount to the disjoint sets of LIs.
Q: (Ql, Qr, w)
$Ql(Q_1)$: left subset of Q (if nested), or index of the first LI in the set (if not nested) $(Q/Integer)$
$Qr(Q_2)$: right subset of Q (if nested), or index of the last LI in the set (if not nested) $(Q/Integer)$
$w(Q_3)$: maximum number of rounds in this Q $(Integer)$

```
cache_wc* ← {} // A cache used to avoid duplicate runs of find_wc*.
// Note: In find_wc*, 'b' and 'e' are indices of the beginning and
// end LIs respectively with default values of '1' and '|p|'.
function find_wc*(p, t, b = 1, e = |p|):
  if (b, e) in cache_wc*:
    return cache_wc*[(b, e)]
  h ← max _potent_LIs(p_{b...e}, t)
  (s, w) ← CATALOG.lookup(e − b + 1, 0, h)
```

```
    best_partition ← (b, e, w)
    if e = b:
      return best_partition
    for j ← b to e − 1:
      Ql ← find_wc*(p, t, b, j)
      Qr ← find_wc*(p, t, j + 1, e)
      if Ql_3 + Qr_3 < best_partition_3:
        best_partition ← (Ql, Qr, Ql_3 + Qr_3)
    cache_wc*[(b, e)] ← best_partition
    return best_partition

// The 'max_potent_LIs' function calculates the maximum number of
// potent LIs with probability 't' given the potency
// probabilities 'p'.
function max_potent_LIs(p, t):
    h ←1
    // 'Pr_p(K ≤ h)': the probability that the number of potent LIs
    // is less than or equal to h using Poisson binomial
    // distribution of p.
    while h < |p| and Pr_p(K ≤ h) < t:
      h ← h + 1
    return h
```

ALO Module-2.B: Partition LIs into disjoint sets for minimizing the expected number of CAGT rounds while keeping its maximum at bay

We introduce a new parameter named ex (set by the user) representing the number of extra rounds allowed in addition to w^* summing up to $wex = w^* + ex$. In the beginning of the CAGT algorithm, we allow an extra round in which the LIs of a disjoint set will be followed simultaneously by the individual, as long as the total of the maximum number of CAGT rounds from all disjoint sets is not greater than wex. **Algorithm A.4** minimizes the expected number of

CAGT rounds by identifying the optimal partition (i.e. disjoint sets of LIs) along with the sets that will have an extra initial CAGT round. The expected number of rounds for a set with extra initial round can be calculated using the weighted average $p0 + (1 - p0) \times w1$ where $p0$ is $Pr_{\boldsymbol{p}}(K \leq 0)$ (probability that the initial extra round returns a '0'), and $w1$ is the maximum number of CAGT rounds for the set if the initial extra round returns a '1' (see '$partition\_with\_extra\_round$' function under **Algorithm A.4)**. This optimal partition (calculated by $find\_Q^*$) is identified by finding the sets in which the extra initial round provides the maximum benefit towards the objective. The algorithm relies on the fact that if there is only one extra initial round available ($ex = 1$), it should be used for a set who's LIs have lower potency probabilities.

Algorithm A.4 Identify the optimal partition along with the disjoint sets that will have an extra initial CAGT round using the $find\_Q^*$ function, such that the expected number of CAGT rounds is minimized while keeping its maximum at bay.

Inputs: LI potency probability vector $\boldsymbol{p}$ and confidence threshold t (same as in **Algorithm A.3)**, $cache\_wc^*$ (populated from **Algorithm A.3)** and total number of CAGT rounds allowed wex (sum of w^* calculated by **Algorithm A.3,** and the user defined extra rounds allowed ex).
Output: The optimal partition Q^* along with the disjoint sets that will have an extra initial CAGT round. It will be in the following nested format dur to recursion. The non-nested Qs amount to the disjoint sets of interest which will be used in Module-3.
Q: (Ql, Qr, w, ar, er)
Ql (Q_1): left subset of Q (if nested), or index of the first LI in the set (if not nested) $(Q/Integer)$
$Qr(Q_2)$: right subset of Q (if nested), or index of the last LI in the set (if not nested) $(Q/Integer)$
w (Q_3): maximum number of rounds in this Q $(Integer)$
$ar(Q_4)$: average number of rounds in this Q $(Float)$
$er(Q_5)$: whether extra round should be used in this Q if not nested $(True/False)$.

```
cache_wex ← {} // A cache used to avoid duplicate runs of find_Q*.
function find_Q*(p, t, cache_wc*, wex, b = 1, e = |p|):
    best_partition ← cache_wc*[(b, e)]
    if best_partition_3 ≥ wex:
        return best_partition
```

```
    if (b, e, wex) in cache_wex:
      return cache_wex[(b, e, wex)]
    partition_wer ← partition_with_extra_round(p, t, b, e)
    if partition_wer_4 < best_partition_3:
      best_partition ← partition_wer
    for j ← b to e − 1:
      Qrw ← cache_wc*[(j + 1, e)]
      Ql ← find_Q*(p, t, cache_wc*, wex − Qrw_3, b, j)
      if Qrw_3 + Ql_3 > wex:
        continue
      if Qrw_3 + Ql_3 = wex:
        Q ← (Ql, Qrw, Qrw_3 + Ql_3, Qrw_3 + Ql_4)
      if Qrw_3 + Ql_3 < wex:
        Qr ← find_Q*(p, t, cache_wc*, wex − Ql_3, j + 1, e)
        Q ← (Ql, Qr, Qr_3 + Ql_3, Qr_4 + Ql_4)
      if Q_4 < best_partition_4:
        best_partition ← Q
    cache_wex[(b, e, wex)] ← best_partition
    return best_partition

  // Create a non-nested Q relating to LIs between b and e in
  // which the extra initial CAGT round is used.
  function partition_with_extra_round(p, t, b, e):
    // Calculate average number of CAGT rounds if extra initial
    // round is used.
    p ← p_{b...e}
    p0 ← Pr_p(K ≤ 0)
    w0 ← 1
    p1 ← 1 − p0
    h1 ← max_potent_LIs(p, t)  // Same function as in Algorithm A.3
    (s1, w1) ← CATALOG.lookup(|p|, 1, h1) + 1
    ar ← p0 × w0 + p1 × w1  // Average number of rounds
    return (b, e, w1, ar, True)
```

A.2.3. ALO Module-3: Perform the CAGT rounds for each LI subset adaptively

Finally in the last ALO module, for each individual, **Algorithm A.5** is followed separately for all disjoint sets of LIs identified (i.e. Q*). **Algorithm A.5,** is similar to main CAGT **Algorithm 1.** with three extensions. First, the maximum number of potent LIs (h) in each disjoint set is first estimated using the max_potent_LIs function (see **Algorithm A.3)**. Second, in a disjoint set Q, if $Q_5 = True$, the initial CAGT round will involve all the LIs of that disjoint set. Third, the LIs that are identified as impotent due to criteria '#2' of the revision table (**Table A.1)** will be combined together from all disjoint sets and verified using the main CAGT **Algorithm 1** with an initial round that involves such LIs (referred to as $V_0^{'}$). This will be repeated until $|V_0^{'}| = 0$. Note that, criteria '#2' is only valid when our assumption about the maximum number of LIs (h) is correct. However, our assumption is correct only with the confidence t (e.g. 0.95 probability) hence requires additional verification.

Algorithm A.5 Identify potent LIs for an individual using one independent CAGT run for each disjoint set of LIs returned by **Algorithm A.4,** and a final CAGT run to identify potent LIs in $V_0^{'}$.

Input: Disjoint sets from the Q^* returned by **Algorithm A.4** referred to as QF^* (i.e. a flat list of non-nested Qs from Q^*), the LI potency probabilities $\boldsymbol{p}$ and confidence threshold t.
Output: The list of impotent and potent LIs referred to as V_0^* and V_1^* respectively.

$V_0^* \leftarrow \{\}; V_1^* \leftarrow \{\}; V_0' \leftarrow \{\}$
for Q **in** QF^*:
    LI $\leftarrow \{Q_1, Q_1 + 1, \ldots Q_2\}$
    $(V_0, V_1) \leftarrow run\_Algorithm1\_extended(LI, Q_5, \boldsymbol{p}, \boldsymbol{t})$
    $V_0^* \leftarrow V_0^* \cup V_0$
    $V_1^* \leftarrow V_1^* \cup V_1$
    $V_0' \leftarrow V_0' \cup (Q \backslash V_0 \backslash V_1)$
while $|V_0'| > 0$:

```
    (V_0, V_1) ← run_Algorithm1_extended(V_0', True, p, t)
    V_0* ← V_0* ∪ V_0
    V_1* ← V_1* ∪ V_1
    V_0' ← V_0' \V_0\ V_1
return (V_0*, V_1*)

function run_Algorithm1_extended(LI, extra, p, t):
    V_0 ← {}; V_1 ← {}
    if extra and get_potency(LI) = 0:
        V_0 ← V_0 ∪ Q
        return (V_0, V_1)
    l ← 1 if extra else 0
    h ← max_potent_LIs(p_{LI}, t)  // Same function as in Algorithm A.3
    model ← CAGT_Model(LI, 0, h)
    do:
        R_i ← model.next_round()
        r_i ← get_potency(R_i)
        (V_0, V_1, V_0') ← model.f_third_extention(R_i, r_i)
    while |LI \ V_0 \ V_1 \ V_0'| > 0
return (V_0, V_1)

class CAGT_Model(LI, l, h):
    // Same as the CAGT_Model used in Algorithm 1. except for the
    // function 'f' that is replaced by 'f_third_extention' which
    // ensures that LIs identified by criteria '#2' of the
    // revision table (Table A.1) are removed from V_0 before the
    // function return.
```

A.3. Spatial Inference Vertex Cover (SPIV)

It is recently shown that a two stage group testing algorithm called SPIV (371) is asymptotically optimal when the expected number of potent LIs is known. We estimated the expected number of potent LIs that are needed by the SPIV algorithm from the LI potency probabilities. Initially the

LIs with high potency probabilities (greater than a hyper-parameter threshold t used in the range of 0.2 to 0.4) are separated from the rest and followed one at a time. In *step1*, we devise overlapping sets of LIs in a ring according to the first stage of the SPIV algorithm (371). In *step2*, each devised set of LIs are followed by the individual simultaneously to determine the potency of each set. In *step3*, individual LI potencies are determined using two rules: (a) if an LI set is determined to be impotent by the individual, all LIs that participate that set will be marked as impotent, and (b) if an LI set is determined to be potent, and all except one of the LIs is determined to be impotent from the previous rule, the single remaining LI in that set will be marked as potent. In *step4*, the remaining LIs (with unknown potencies) are followed by the individual one-by-one to determine each LI potency, and the rules are applied after each until all potencies are determined (see **Figure A.2**). The second stage of the SPIV algorithm can be followed in *step4*, however we believe our updated step is appropriate due to the adaptive nature of the problem, its simplicity and practicality. One can integrate the second stage of SPIV into the last step to evaluate its potential value.

A.4. Appendix Figures

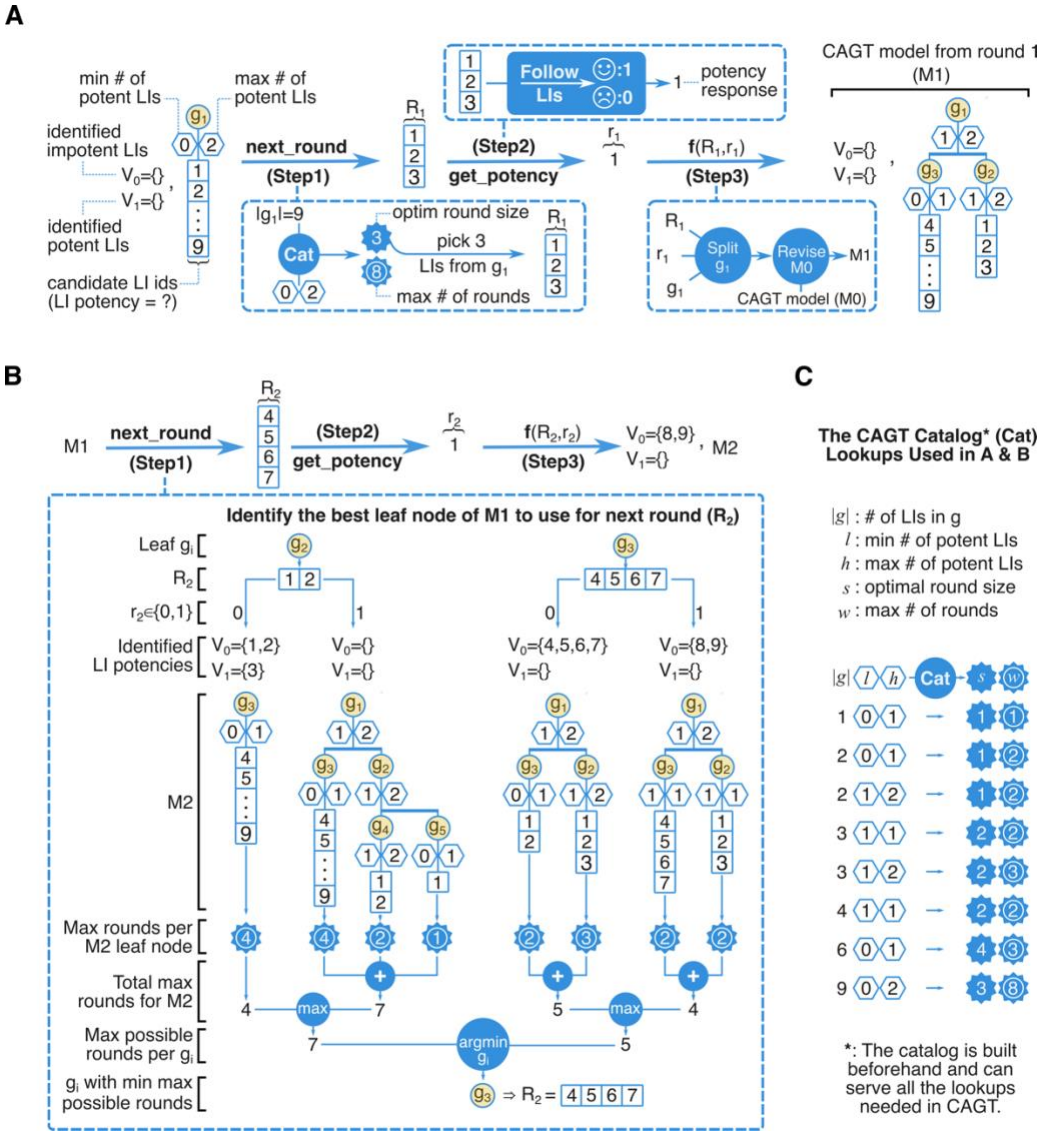

Figure A.1 CAGT algorithm illustration by an example in two rounds. The CAGT algorithm is used to identify a maximum of 2 potent LIs amongst 9 candidate LIs in two rounds **A** & **B**. Note that the "Cat" icon represents lookup from the CAGT catalog in order to find (s, w) for a given $|g|$ number of LIs with between l and h potent LIs as illustrated with examples in **C**. (**A**) The CAGT model is initiated with a single node that includes all LIs, and a range for the number of potent LIs to be identified. In Step1, a catalog lookup is made using (9, 0, 2) tuple which returns (3, 8) indicating that CAGT requires a maximum of 8 rounds, and 3 LIs should be picked for the next round R_1. In Step2, the individual follows the LIs 1, 2, and 3 simultaneously and achieves the target health outcome indicating that at least one LI amongst these three are potent. In Step3, the CAGT model is updated by (a) splitting the node g_1 into two nodes and revising the CAGT model

using **Table A.1** The revised CAGT model M1 at the end of this step has two leaf nodes for two disjoint sets of LIs, each with different limits on the number of potent LIs in them. (**B**) In Step1, each leaf node of the M1 model is examined to identify the best leaf node to pick the LIs from, for the next round. For g_2, a catalog lookup is performed using the (3,1,2) tuple which returns (2,3) indicating that 2 LIs should be used for the next round R_2 if the LIs are picked from g_2. Next the alternative updated models for potential responses to these LIs are illustrated below g_2 (and similarly for g_3). For each leaf node of the alternative models, a catalog lookup is performed to identify the total max number of rounds needed for the model and the maximum is identified. In the illustrated example, node g_3 is found to be better for picking the next LIs from since to max total rounds that CAGT will require is less, compared the case where g_2 is used. The Step2 and Step3 are followed similar to the first round leading into discovery of two LIs (8 and 9) that are determined to be impotent. This is due to the fact that in g_3, there maximum number of potent LIs is 1, therefore when the response r_2 to R_2=4,5,6,7 is 1, the remaining LIs in g_3 must be impotent.

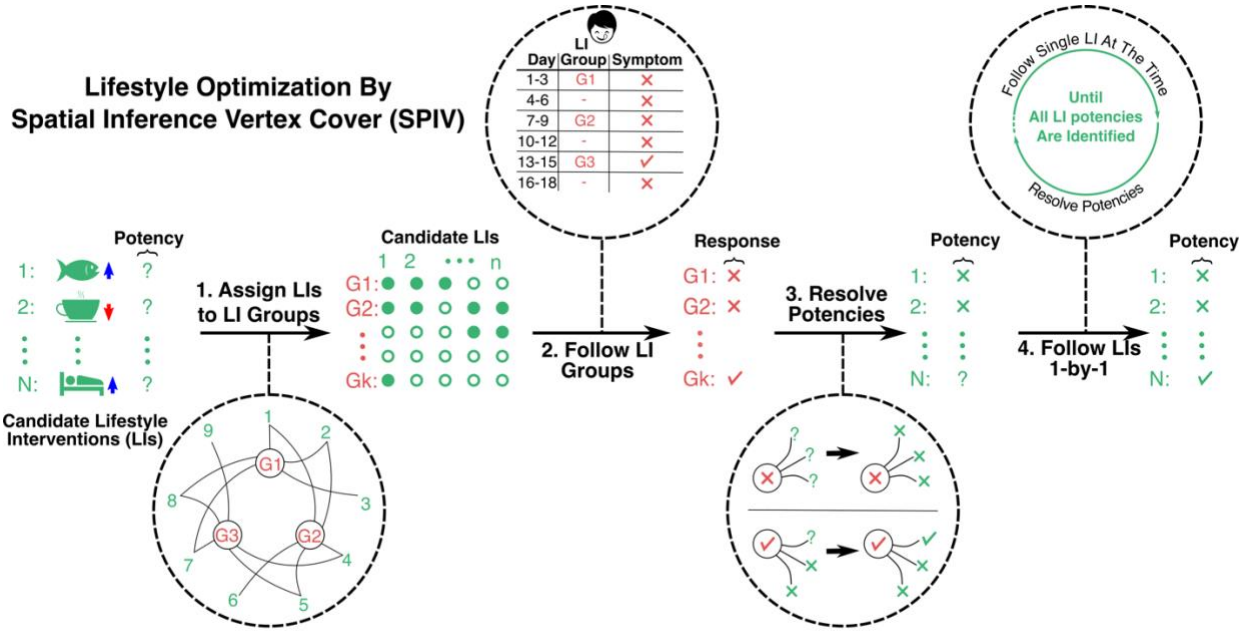

Figure A.2 Lifestyle Optimization By Spatial Inference Vertex Cover (SPIV). We adopted the asymptotically optimal two-stage group testing method called SPIV for lifestyle optimization. *First*, LIs are assigned to LI groups. *Second*, the LIs in each group are followed simultaneously by the individual. *Third*, the responses of the individual to LI groups (based on their symptoms) are used to resolve/identify the potency of each LI. In this step, (a) the LIs that were assigned to a group with no response are determined to be impotent, and (b) an LI is determined to be potent if all other LIs in that group are impotent while the individual had a positive response to the corresponding group. *Fourth*, the remining LIs with unknown potencies are followed 1-by-1, until all potencies are fully resolved.