A statistical framework for nutriomics data analysis

XIANGNAN XU

Supervisor: Jean Yang and Samuel Muller

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
The University of Sydney
Australia

17 February 2023

Statement of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature

Abstract

The introduction of diverse high throughput omics technologies into nutrition research have given rise to the discipline of nutriomics, defined as the science that researches how nutrition comprehensively relates to health and wellbeing. Nutriomics holds promise in understanding many aspects of how nutrition impacts health. From a data analysis viewpoint, there are currently many open challenges due to the inherent complex and heterogenous nature of the available data in nutriomics.

In this thesis, we introduce nutriomics, the data to be analysed and frame the statistical challenges in Chapter 1. We then devise statistical modelling and machine learning methods for nutriomics data analysis and address three aspects of these challenges: non-linearity in the statistical relationship between covariates and the health response, high dimensionality, as well as the complex heterogeneity of the available data. Specifically,

(1) Due to the complex interaction among nutrients, the relationship between nutrition and omics features such as gene expression would be non-linear. In Chapter 2, we propose a statistical framework, that we coin LC-N2G, to test whether the association between nutrition intake and omics features of interest are significantly different from being unrelated. We use public data as an example to show LC-N2G's ability to discover non-linear associations between nutrition and gene expression. Part of this chapter is published in Xu et al. [184].

(2) The high throughput technology provides high dimensional omics features that show diverse responses to nutrition intake as well as other experimental conditions. This poses a challenge in understanding how nutrition acts on the omics features. In Chapter 3, we propose a statistical method, coined eNODAL, to cluster high-dimensional omics features based on how they respond to nutrition intake. The application of eNODAL to a mouse proteomics nutrition study shows that eNODAL can identify interpretable clusters of proteins with similar responses to diet and drug

treatment. Part of this chapter will be submitted for publication after completion of the thesis.

(3) The associations among nutrition omics and health outcomes are heterogeneous: people with different diets could show different relationships between omics features and health outcomes. In Chapter 4, we propose a statistical model, which we call NEMoE, to uncover the heterogeneous interplay among diet, omics, and health outcomes. We use a microbiome Parkinson's disease (PD) study to illustrate the method and show that NEMoE is able to identify diet-specific microbial signatures of PD. Part of this chapter is under review and a preprint is available in Xu et al. [185].

The thesis concludes in Chapter 5 with a discussion that specifically provides possible future extensions based on the research in this thesis.

Publications

Some of the methods, concepts, analyses and results in this thesis have led to publications as listed in the following.

(1) **Xiangnan Xu**, Samantha M. Solon-Biet, Alistair Senior, David Raubenheimer, Stephen J. Simpson, Luigi Fontana, Samuel Muller, and Jean YH Yang. LC-N2G: a local consistency approach for nutrigenomics data analysis. *BMC Bioinformatics* 21(1): 1-14, 2020.

(2) Michal Lubomski[*], **Xiangnan Xu**[*], Andrew J. Holmes, Jean YH Yang, Carolyn M. Sue, and Ryan L. Davis. The impact of device-assisted therapies on the gut microbiome in Parkinson's disease. *Journal of Neurology* 269(2): 780-795, 2022.

(3) Michal Lubomski[*], **Xiangnan Xu**[*], Andrew J. Holmes, Samuel Muller, Jean YH Yang, Ryan L. Davis, and Carolyn M. Sue. Nutritional intake and gut microbiome composition predict Parkinson's disease. *Frontiers in Aging Neuroscience* 14: 1-16, 2022.

(4) Michal Lubomski[*], **Xiangnan Xu**[*], Andrew J. Holmes, Samuel Muller, Jean YH Yang, Ryan L. Davis, and Carolyn M. Sue. The Gut Microbiome in Parkinson's Disease: A Longitudinal Study of the Impacts on Disease Progression and the Use of Device-Assisted Therapies. *Frontiers in Aging Neuroscience* 14: 1-24, 2022.

(5) **Xiangnan Xu**, Michal Lubomski, Andrew J. Holmes, Carolyn M. Sue, Ryan L. Davis, Samuel Muller, and Jean YH Yang. NEMoE: A nutrition aware regularized mixture of experts model addressing diet-cohort heterogeneity of gut microbiota in Parkinson's disease. *Under Review* (2021).

There are other collaboration projects that I have contributed to which resulted in co-first authored and peer-reviewed publications. These works also eventuated from this thesis and have further implications. All the results of the three Lubomski et al projects [99–101] are

[*]Equal contribution

from analysing data from a large cohort longitudinal microbiome-Parkinson's disease (PD) project. This dataset is also used in Chapter 4. In all of these publications, I did the processing from raw sequence files to ASV tables and also provided all the statistical model building.

These three papers focus on different research questions: Lubomski et al. [101] aims at profiling acute gut microbial community alterations in response to PD device-assisted therapies initiation. Also, Lubomski et al. [101] is a pilot study of the microbial profile. Lubomski et al. [99] focuses on assessing the utility of gut microbiome compositional changes combined with macronutrient intake to develop a predictive model of PD. It is a crucial motivation of Chapter 4. Lubomski et al [100] is a longitudinal study that investigated the temporal stability of gut microbiome profiles from PD patients on standard therapies and those initiating DAT and define multivariate models of disease and progression. This study is a longitudinal analysis of the cohort and could serve as the basis for a potential extension of the methodology developed in Chapter 4.

Signature

Contents

CHAPTER 1

Introduction

Nutrition and diet play an important role in human health. As the German philosopher Ludwig Feuerbach famously phrased it in the 19th century: "We are what we eat". A balanced diet contributes to the quality of life, helps maintain a healthy body weight, and reduces the risk of several diseases such as diabetes [130] and inflammatory bowel disease [111, 160]. Health and wellness-oriented eating is becoming a big trend in food science with much research devoted to the selection of food for healthy living [154].

The thriving of high throughput technology gives rise to a variety of "omics" research, aiming at collective characterisation and quantification of pools of biological molecules that translate into the structure, function, and dynamics of organisms [92]. Such technology provides nutrition scientists insight into further understanding the interaction between the physical body and nutrients at unprecedented resolution. In 2011, Kato and colleagues [71] used the term "nutriomics" as the study of nutrient-gene interactions and their effects on health. Since then, a large number of studies have extended the scope of nutriomics to analyse the interactions between nutrition and other molecules such as proteins and metabolites [154].

Besides the "omics" introduced above, the microbiome, which refers to the totality of all microbes in and on the human body [90], has recognised its importance in health [130]. With the discovery from The Human Microbiome Project Consortium [50] that the composition of the microbiome varies based on diet, health and environment, investigation of the microbiome and how it contributes to human health has gained much attention. This, in turn, has generated much interest in the development of novel statistical and computational methods for the analysis of microbiome data [90, 182], and in the last few years, the integration of microbiome with other omics data.

In this thesis, we will refer to "nutriomics" in a broad sense as the study of nutrients and any molecular level processes (e.g. protein, metabolite) in the body as well as the microbial symbionts and their interactions on health. Nutriomics studies provide a better understanding of the processes and mechanisms that define the relationships between nutrition, omics, and physiology and further hold great promise for well-tailored prevention and intervention plans for populations, cohorts, and individuals in precision health [141].

The combination of nutrition and omics technology on the one hand sheds light on the processes and mechanisms of how nutrition helps to keep us healthy, on the other hand, such integration also imposes statistical challenges in extracting useful information from complex multi-modality data. Currently, even without further data collection, much remains to be discovered from existing data sets - the limitation is not the availability or makeup of the data, but there is a pressing gap of adequate statistical and computational approaches that can extract meaningful information from such complex data [75].

In this thesis, we consider four types of nutriomics data:

- nutrition data from cohort studies and surveys,
- matched omics data,
- microbiome data from 16S sequencing platform,
- physiological outcomes from clinical observations or other measurements.

In addition to the various data analysis challenges within each of these four data modalities, the key challenges of a joint investigation of various nutriomics data modalities can be summarised as follows.

- **Non-linearity.** The relationship among nutrients and other dietary constituents as well as the host phenotype could be non-linear (in terms of modelling the expected value of the response variable) and ignoring such association might risk drawing erroneous conclusions about the relationship of nutrition and physiological outcomes [121].

- **High-dimensionality.** High throughput experiments generate high dimensional omics features, therefore imposing challenges in the analysis of their relationship with nutrition intake. Due to a large number of variables typically present in omics data, it might not be possible to investigate how every single omics feature responds to different amounts of nutrients. This makes it challenging to comprehensively study and understand how nutrition affects the omics profile and how these high dimensional omics features in turn affects health [141].
- **Heterogeneity in cohort studies.** In cohort studies, people can have very different diets, therefore the information about the nutrients they eat can greatly vary. In nutriomics studies, this means that the samples differ from one another. People who eat differently may have a different microenvironment and different relationships between the omics features and the phenotypes. Discovering subcohorts with similar diets is beneficial for building more accurate models serve for specific subcohort.

In addition to the usual challenges associated with integrating multi-omics data, the dynamic nature of the microbiome data and the added taxonomic structure as its features characteristic [182] creates additional challenges in identifying features of interest with high association with omics data. A recent review by Qian and Ho [123] pointed out that there is a critical need to utilise prior biological knowledge and existing published data sets for the discovery of molecular mechanisms underlying the association between diet, microbial compositions and human health. Such challenges arise from different types of nutriomics studies. **In this thesis, we will propose novel statistical methods to deal with these challenges.**

1.1 Nutriomics data modalities

Omics data from sequencing platforms, nutrition data from cohort studies and surveys, and the different matched multi-omics platforms are all important types of nutriomics data. In this thesis, four different nutriomics data modalities are included, which consist of four matrices containing different information from the same n samples corresponding to their nutrient

intake $W_{n\times q}$, omics features $Z_{n\times s}$, microbiome features $X_{n\times p}$ and physiological outcomes $Y_{n\times r}$ (Figure 1.1).

1.1.1 Nutrition data

Nutrition data represents the dietary information of each sample. A typical nutrition data contains information about the nutrient intake profile of each sample. It could be derived from either experimental designs including quantity and quality of food intake for each mouse in animal studies or food frequency questionnaires (FFQs) from cohort studies in human studies.

In this thesis, we will denote nutrition data using a nutrition matrix $W_{n\times q}$, where each column represents an intake of a nutrient, like macro-nutrients (protein, carbohydrate and fat) or micronutrients such as vitamins. Each row represents a subject in a cohort study or a mouse in an experimental study. It is important to note that in cohort studies, the nutrition data is often a matrix containing already processed information, especially from FFQs, which typically contain some measurement error [18] due to the inherent imprecision in self-reporting. In this thesis, we consider the matrix $W_{n\times q}$ as given, that is assuming no measurement error. Including the additional challenge of measurement errors in the data we analysed here, and the statistical methodology we developed, is outside the scope of this thesis and could be tackled in future work.

1.1.2 Omics data

"Omics data" refers to data generated from fields of biological study whose names end with "omics". These typically refer to genomics, proteomics, metabolomics, interactomics, and others. In this thesis, we only considered one specific type of omics for each research question. We will define a given omics matrix as $Z_{n\times s}$, where each column represents an omics feature such as a particular gene expression or abundance of a protein and each row represents sample information similar to the nutrition matrix. For example, each column of $Z_{n\times s}$ could be the expression of a particular gene when we want to investigate how different diets shape

the profile of gene expression. It also could be the concentration of one metabolite or the abundance of a protein when we focus on metabolomics or proteomics profiles.

1.1.3 Microbiome data

Microbiome data refers to data generated by 16S rRNA sequencing and shotgun metagenomic sequencing, representing the abundance or composition of the gut microbiota of each sample [183]. We use a matrix $X_{n\times p}$ to represent such microbial data. Similar to the omics matrix $Z_{n\times s}$, each row in $X_{n\times p}$ corresponds to a sample, and each column represents a relative or absolute abundance of taxa, which we also call microbial features. Besides these similarities, microbial features also have their own unique characteristics, including a taxonomic table providing a hierarchical structure of the microbiome [183] and a phylogenetic tree reflecting the evolutionary information among the microbiota [24].

1.1.4 Physiological or outcome data

In this thesis, physiological and outcome data include disease state, weight, body mass index (BMI), and others. We denote this data as a matrix $Y_{n\times r}$, representing the physiological outcomes for each sample. The dimension of r could be flexible depending on the research question. For example, in a disease study, Y is a vector indicating whether each individual has the disease or not. In a nutriomics study, Y could also be a matrix involving more phenotypes such as body weight, BMI, blood pressure, and others. Such physiological outcomes are usually considered as the results or consequences of different underlying biological processes of body system [173] and could be considered as a response variable or vector for X, Z, and W.

Figure 1.1 provides a summary of these notations and throughout the thesis, we keep the meaning and use of the matrices W, X, Z, and Y to represent a nutrition, microbiome, omics, and physicological outcome matrix, respectively. However, we adapt the size n, p, q and r within the context of each chapter.

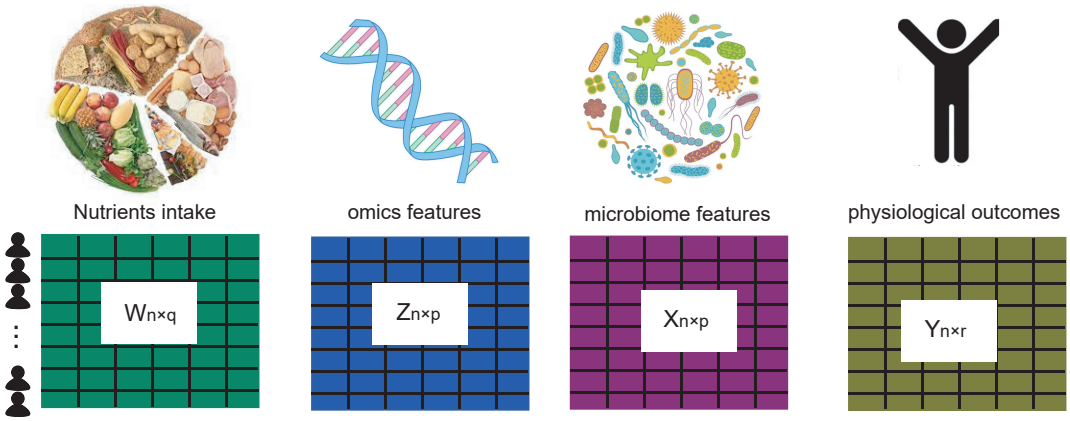

FIGURE 1.1. **Illustration of datasets in nutriomics studies.**

1.2 Nutriomics studies

A substantial body of research has been devoted to generating such nutriomics data. Broadly, we can categorise the various studies used in this thesis into mouse experimental studies and cohort studies and discuss these in more detail in the next two sections.

1.2.1 Mouse experimental study

A typical mouse experimental study considers experiments where different litters of mice have an equal chance to receive any of the treatments under study, and generate comparable intervention groups, which are alike in all the important aspects except for the intervention each group receives [149]. In such experimental studies, mice are randomised with well-controlled experimental conditions and thus less susceptible to confounding [162]. The data generated from such experiments could be useful when analysing the association between experimental conditions and their impact on high-dimensional omics features. In a nutriomics context, such randomized mice could receive varied diets with different quantities and quality.

In Chapter 3, we consider such experimental nutriomics data. We use a mouse nutrition-proteomics study published in Le Couteur et al. [87], where the experiment was performed on 128 mice with 10 different diets, five different ratios of protein, carbohydrates, and fat intake (protein: carbohydrate: fat = 10:70:20, 14:57:29, 19:63:18, 33:20:47 or 55:25:20 as

% of kJ/g) at either of two different energy contents (standard diet 14.8 kJ/g, low energy 8 kJ/g), and 4 different drug intakes, resulting in a total of 40 treatment conditions. Besides the intake of diet and treatment information, the proteomics of the mouse liver is also measured to analyse the effect of diet and drug on the liver proteome which results in 4,987 protein abundance. Furthermore, for each mouse, metabolic phenotypes such as body weight, BMI, glucose-insulin level and other outcomes were measured.

Based on the complex experimental design, this data potentially provide evidence on how diet and drug intake affect the liver proteome. Through correlation analysis, this research discovered an overall strong negative correlation between dietary energy and the spliceosome. Furthermore, it was pointed out that the intake of three kinds of drugs mostly dampened the correlation coefficients between energy intake and protein abundance. These findings indicate that diet, drug as well as their interaction could have an impact on liver proteins.

1.2.2 Cohort study

A cohort study is an observational study design in which groups of subjects are identified based on their exposure to a particular risk factor and then compared to a group that have not been exposed to that same factor [157]. In the context of nutriomics, such cohort studies generate data from a population, thus providing knowledge and recommendations that are relevant for humans [64].

In Chapter 4, we use a dataset from a microbiome-PD study, as published and analysed in Lubomski et al. [100], Lubomski et al. [101], and Lubomski et al. [99]. This study includes 103 PD patients and 81 healthy controls (HC) who tracked their dietary habits through a comprehensive Food Frequency Questionnaire. The gut microbiome profile of the cohort was measured using 16S rRNA sequencing resulting in high dimensional microbial features. The corresponding clinical variables such as the measurement of constipation severity and gut motility, and physical activity are also included in the data.

We analyzed this dataset to uncover relationship between the microbiome and PD and the related results were published in Lubomski et al. [101] and in Lubomski et al. [99]. We

used differential analysis as well as correlation analysis to identify significant compositional differences in the gut microbial profiles between PD and HC. In addition, it was noticed that incorporating nutritional information into the prediction model could improve predictive performance [99]. These findings support the combination of the gut microbiome and nutritional data as a potentially useful biomarker of PD to improve diagnosis and guide clinical management [99].

1.3 Statistical challenge of the nutriomics data analysis and contribution of this thesis

The characteristics and structures of nutriomics datasets give rise to some statistical challenges in their analysis. Several methods have been proposed to deal with such challenges. In this thesis, we focus on important challenges in nutriomics data that originate from

- non-linearity of the associations between nutrients and omics features,
- the complex response of omics features to varying nutrition intakes in high-dimensional experimental nutriomics studies,
- diet induced heterogeneity in observational nutriomics studies.

Figure 1.3 gives a visual overview of how the thesis is organised.

1.3.1 Non-linear response between nutrition and omics features

For the non-linearity problem among nutrition intake and molecular features, which arises in most of the nutriomics studies. A popular approach to deal with this problem is to first fit a nonlinear model such as a generalised additive model (GAM) [175, 177, 179], where the fitted surfaces are projected onto a two-dimensional space spanned by two nutrition intake features. This method was termed as the nutritional geometry framework in series of papers [87, 115, 128, 138, 140, 141].

The nutritional geometry framework (NGF) provides a conceptual tool that helps dissect nutrient-nutrient interactions. It considers nutritional problems in an q-dimensional nutritional space. By assessing the effects of nutrients intakes across this space on an outcome of interest, one is able to parse the main and interaction effects of the q-nutrients considered. An illustration of a NGF is shown in Figure 1.2 using a public dataset from Solon-Biet et al. [145]. In this figure, the x-axis represents the intake of carbohydrate, the y-axis shows the intake of protein in kJ per day and the surface shows the expression of gene *Fgf21*, which plays an important role in the maintenance of energy homeostasis [76, 147]. Because of this ability to give insight into main and interaction effects, the NGF has been used to design experiments in model organisms, plan the analyses of nutritional epidemiology datasets, and even define the treatment groups in randomised controlled trials [138].

The NGF method gives a good visualisation of the non-linear relationship between nutrients intake and a response variable. However, how to best evaluate an NGF is not yet well studied. Here we refer to "evaluation of NGF" as testing whether the relationship visualised by the NGF is significantly different from a randomly permuted response variable. We also consider related questions including how well the NGF represents the nutrition data, how is the difference between two NGFs, and how to choose an informative combination of nutrition variables when multiple nutrients intake were measured [128, 141].

To address this, in Chapter 2, we develop the Local Consistency Nutrition to Gene (LC-N2G) method, which builds on the NGF. We focus on the discovery of non-linear associations between nutrition and genomics (Figure 1.3 a). In Chapter 2 we aim to determine whether a particular omics feature of interest shows a significant association with selected nutrition features. The motivating data for this chapter is a mouse nutrition-genomics dataset [145] that includes the gene expression of 46 mice as well as their nutrients intake. We apply LC-N2G to this dataset and discover potential non-linear relationships between nutrition variables and gene expressions.

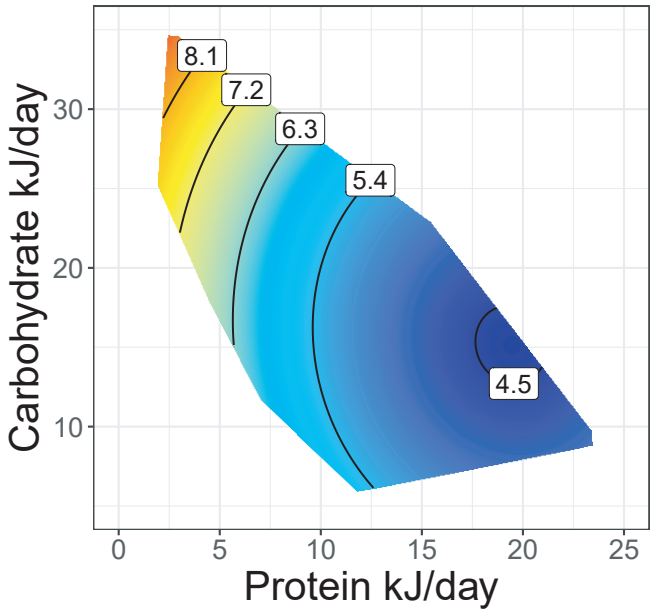

FIGURE 1.2. **Illustration of the NGF.** x-axis represents the intake of carbohydrate, y-axis shows the intake of protein in kJ per day and the surface shows the expression of gene *Fgf21*

1.3.2 Complex associations in high-dimensional experimental studies

For experimental nutriomics studies, as illustrated in the previous section, researchers seek to uncover the associations among nutrition features, molecular features, and physiological outcomes. Modern high throughput 'omics' experiments are able to create high dimensional features, often in the thousands. Such high dimensionality gives rise to the challenge of understanding how nutrition shapes these omics features: typically, the NGF approach introduced in the previous section, which considers one omics feature at a time, is not geared towards handling such multivariate outcomes. Especially, the non-linearity in nutriomics data makes interpreting such relationships difficult (see Section 1.3.1).

A popular way to discover such patterns among high-dimensional omics features is to first cluster omics features based on their similarity to then analyse how each group responds to its experimental settings. There are several statistical and machine learning methods that cluster such features. These methods include ClusterofVar [21], WGCNA [82] and other clustering methods as partially listed in Murphy [114]. However, in the nutriomics context,

such clustering methods only use the information from the omics dataset, while ignoring the relationship with nutrition. An integrated clustering method between nutrition information, omics features, and phenotype outcomes could shed light on how nutrition acts on the physical body.

In Chapter 3, we propose a new approach which we call the experiment guided nutriomics data clustering (eNODAL) framework. The eNODAL approach deals with the challenges mentioned above. This new method aims to identify clusters of omics features in experimental nutriomics data that have similar nutrition responses. Chapter 3 builds on Chapter 2: eNODAL further groups the high-dimensional omics features based on their response to nutrition and drug intake. The motivating data set for this chapter is from a mouse nutrition-proteomics experiment [87], which we introduced above in Section 1.2.1. Application of eNODAL on this dataset identifies biologically meaningful subclusters with similar responses to nutrition and drug treatment.

1.3.3 Heterogeneity in cohort studies

In cohort studies, the collection of samples can be heterogeneous because of very different conditions, such as the vastly different lifestyles of people, and their natural and physical environment [59]. The high variability in these data imposes challenges in many applications such as in the prediction of health outcomes using molecular features and in the identification of disease-related molecular signatures. In nutriomics cohort studies, such heterogeneity gives rise to the concept of personalised nutrition [154]: the influence of nutrition will not be universal for all people. Thus, the recommendation of diet should be individualised for better health [191].

Divide-and-conquer, a widely used strategy [67] to deal with such cohort heterogeneity, first partitions the population into several subcohorts that have similar properties and then performs downstream analysis such as building a classification model focused on samples within each subcohort. Many existing methods use such a strategy that partitions the population based on their similarity. Examples include latent class analysis [103] and mixture models [44,

108]. These methods work well in the identification of clusters in the population. However, in the context of cohort nutriomics data, the interplay between nutrition, "omics" and health outcomes can be complex. Clustering within each dataset may not fully reflect such complex relationships. Methods that jointly model nutrition information, omics features, and phenotype outcomes may better suit the underlying complex interactions and thus have the potential to make important contributions to personalised nutrition.

In Chapter 4 we focus on the heterogeneity in nutriomics cohort studies: people with different diets may experience very different relationships between omics features and their respective health outcomes. This is conceptually visualised as a high-level overview in Figure 1.3 c). In Chapter 4 we consider this population heterogeneity problem in a microbiome Parkinson's disease context using data from previously published studies [98–101] (See Section 1.2.2). We propose the Nutrition-Ecotype Mixture of Experts (NEMoE) model, to identify diet subcohorts that experience different microbiome and disease state relationships. Application of NEMoE on both simulated data and real-world microbiome data shows that NEMoE can identify meaningful diet subcohorts as well as their PD-related diet-specific microbial signatures.

1.4 Organisation of this thesis

This thesis proposes several approaches to address various challenges in the analysis of nutriomics data originating from different types of studies aiming to answer a range of nutriomics questions. Some of these methods, concepts, analyses, and results have already been published or are currently under review. Figure 1.3 summarises the topics of each chapter and also provides references and publications that originated from these chapters.

The thesis is organised as follows. Chapter 2 focuses on the non-linear relationship between nutrition and omics features where we create a statistic called local consistency (LC) statistic and use LC to test whether particular non-linear nutrition-omics relationships are significant. A mouse nutrition-genomics dataset is used to evaluate this method, which led to the publication in Xu et al. [184]. Chapter 3 focuses on experimental nutriomics data in mouse experimental

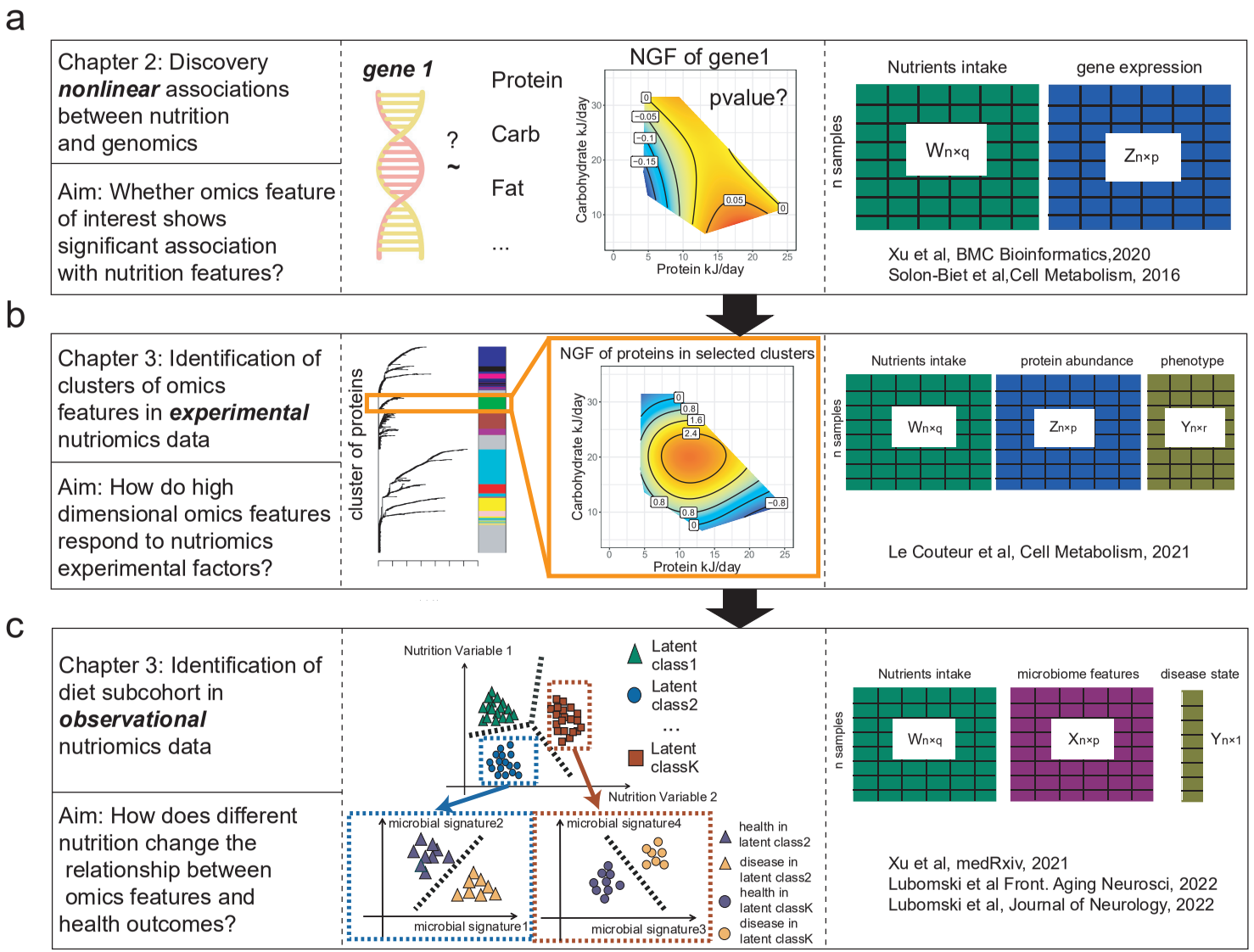

FIGURE 1.3. **Organization of the thesis**. a) Chapter 2 focuses on the discovery of nonlinear associations between nutrition and omics. The motivating dataset is Solon-Biet et al [145] and the corresponding method was also published in Xu et al[184]. b) Chapter 3 focuses on the identification of clusters of omics features in experimental nutriomics data. The motivating dataset is Le Couteur et al [87] c) Chapter 4 focuses on the identification of diet subcohort in observational nutriomics data. The motivating dataset is Lubomski et al [99, 101] and can also be seen at Xu et al [185]

nutritional studies with high dimensional proteomics features. We develop an experiment guided nutriomics data clustering (eNODAL) method to group the high-dimensional omics features based on their response to nutrition. We use a mouse nutrition-proteomics dataset to evaluate the proposed method. Chapter 4 focuses on observational microbiome and nutrition studies. Here, we propose the Nutrition-Ecotype Mixture of Experts (NEMoE) model, to identify diet subcohorts with different microbiome and disease state relationships. We conclude the thesis in Chapter 5. A summary of all the data and key contributions of each chapter is shown in Figure 1.3. This thesis integrates statistical methodologies and nutritional considerations to create a framework that contributes to research in nutriomics. Through developments of three flexible statistical approaches, we thus realise the potential to unravel the complex relationship among nutrition information, omics features as well as health outcomes to further precision nutrition.

CHAPTER 2

Discovery of nonlinear associations between nutrition and omics

---

Diet and health are closely related [128, 154]. Interestingly, several studies have pointed out that the statistical relationship between nutrients and features related to health outcomes could be nonlinear due to interactions occurring between nutrients and other dietary constituents [141]. Several studies suggested that ignoring such non-linearity may risk drawing contradictory conclusions [5, 121]. A widely used multidimensional visualisation of nutrition, called nutritional geometry framework (NGF)[141], which aims to represent the nonlinear relationship between nutrition and response of interest, was introduced (section 1). The NGF provides a unified perspective for the effect of nutrition on health outcomes and could be efficient in the analysis and visualisation of many nutriomics studies [27, 53, 121, 141].

The NGF provides a way to represent how nutrition acts on the response. There are still several open questions as pointed out in Simpson et al. [141]. One of them concerns the lack of quantitative measurement in the NGF. For example, the interpretation of the NGF is mainly based on a visual examination of the fitted surfaces as we described in Chapter 1. Whether the relationship shown in the NGF is significantly different from noise is unclear [141]. Such drawbacks could be even more pronounced as the number of measured nutrient features is growing. Eventually, for such high-dimensional scenarios, visually examining each fitted curve or determining which combination of nutrients best reflects the association becomes infeasible.

---

*Most of this work was published in: **Xiangnan Xu**, Samantha M. Solon-Biet, Alistair Senior, David Raubenheimer, Stephen J. Simpson, Luigi Fontana, Samuel Mueller, and Jean YH Yang. LC-N2G: a local consistency approach for nutrigenomics data analysis. *BMC bioinformatics* 21(1): 1-14, 2020.

In Chapter 1, we introduced the currently widely used NGF [128, 140] in the context of nutriomics data analysis. In this chapter, we propose a novel statistical method, called **L**ocal-**C**onsistency-**N2G** (LC-N2G), to quantitatively discover the relationship between nutrition intake and the expression of a gene of interest. LC-N2G achieves this using a novel Local Consistency Statistic (LC-Stat, section 2.2.1), by optimising combinations of nutrients having the smallest LC-Stat value. LC-N2G identifies informative nutrients that act on the response. Then a significance test, which we coin the LC-Test, further informs whether the relationship between the gene expression and a combination of nutrients is significantly different from randomly permuted gene expression. Finally, these informative combinations can be visualised through the NGF. The effectiveness of LC-N2G is validated through simulation studies in sections 2.3.1 and 2.3.2 and real data analysis in section 2.3.3. It is shown that LC-N2G can discover the nonlinear relationship between nutrition intake and gene expression. In this chapter, we focus on nutrition and gene expression data. However, we also point out that the method developed is not limited to gene expression data but could also be easily adapted to other nutriomics data.

2.1 Background

Nutrients are simple organic compounds involved in biochemical reactions that produce energy or are constituents of cellular biomass [39]. Nutrigenomics, the combination of nutrition and genomics research, which aims to shed light on and describe, characterise, and integrate the interactions between nutritional compounds and genome-wide gene expression [34], has been thriving after the completion of the Human Genome Project 15 years ago [28, 81]. This research led to a better understanding of the processes and mechanisms that define the relationships between nutrition, genetics, and physiology. Nutrigenomics holds great promise to provide individualised plans to improve health (precision health)[109].

The relationship between the mechanisms of nutrients and genes is complex. For example, there are many potential non-linear interactions which is a challenge for any univariate

measure of association [89]. One recent approach that addresses this challenge is the Nutritional Geometry Framework (NGF) proposed by Simpson and Raubenheimer [140] and Raubenheimer and Simpson [128], where the nutritional requirements and their response is represented graphically in a pre-specified k-dimensional space. Each of these dimensions is a nutrient or another dietary constituent. For example, the NGF shown in Figure 2.1(c) is a two-dimensional colour graph and visualises the relationship of two nutrients, Protein and Carbohydrate, on the expression level of a particular gene, *Fgf21*. These two nutrients are represented on the x and y-axis, respectively. The gene expression level is highlighted through a surface using a colour scale [140]. This framework makes available tools to interpret how nutrients and other dietary constituents, directly or through their interactions influence a given phenotype. The success of NGF has been demonstrated by much recent research [30, 31, 88, 121, 141, 145]. However, this NGF framework requires the manual selection of informative combinations of nutrients prior to visualisation. This often requires deep domain knowledge and when the number of nutrients q is large, considering all pairwise combinations $q(q-1)/2 = \binom{q}{2}$ becomes time-consuming.

In this chapter, we will address this challenge by proposing the **L**ocal-**C**onsistency-**N2G** (LC-N2G) method, where N2G stands for both, identifying nutrition variables and using these to inform on changes in genes, that is achieving "nutrition to graphics", visualising multivariate nutrition-gene relation. LC-N2G first constructs a Local Consistency statistic (LC-Stat) that measures the smoothness of a given gene expression surface relative to a set of axes that represents a combination of nutrients. Then some optimisation methods can be used for choosing combinations with the smallest LC-Stat (LC-Opt). Next, a significance test using the LC-Stat is performed to test if the relationship between the gene expression and a combination of nutrients is significantly different from randomly permuted gene expression (LC-Test). Finally, NGF is used to visualise selected combinations with gene expression. We use both simulated and real data to validate our proposed LC-N2G method and the results showed that it can correctly identify the relationship between genes and nutrients. The overall workflow is illustrated in Figure 2.1.

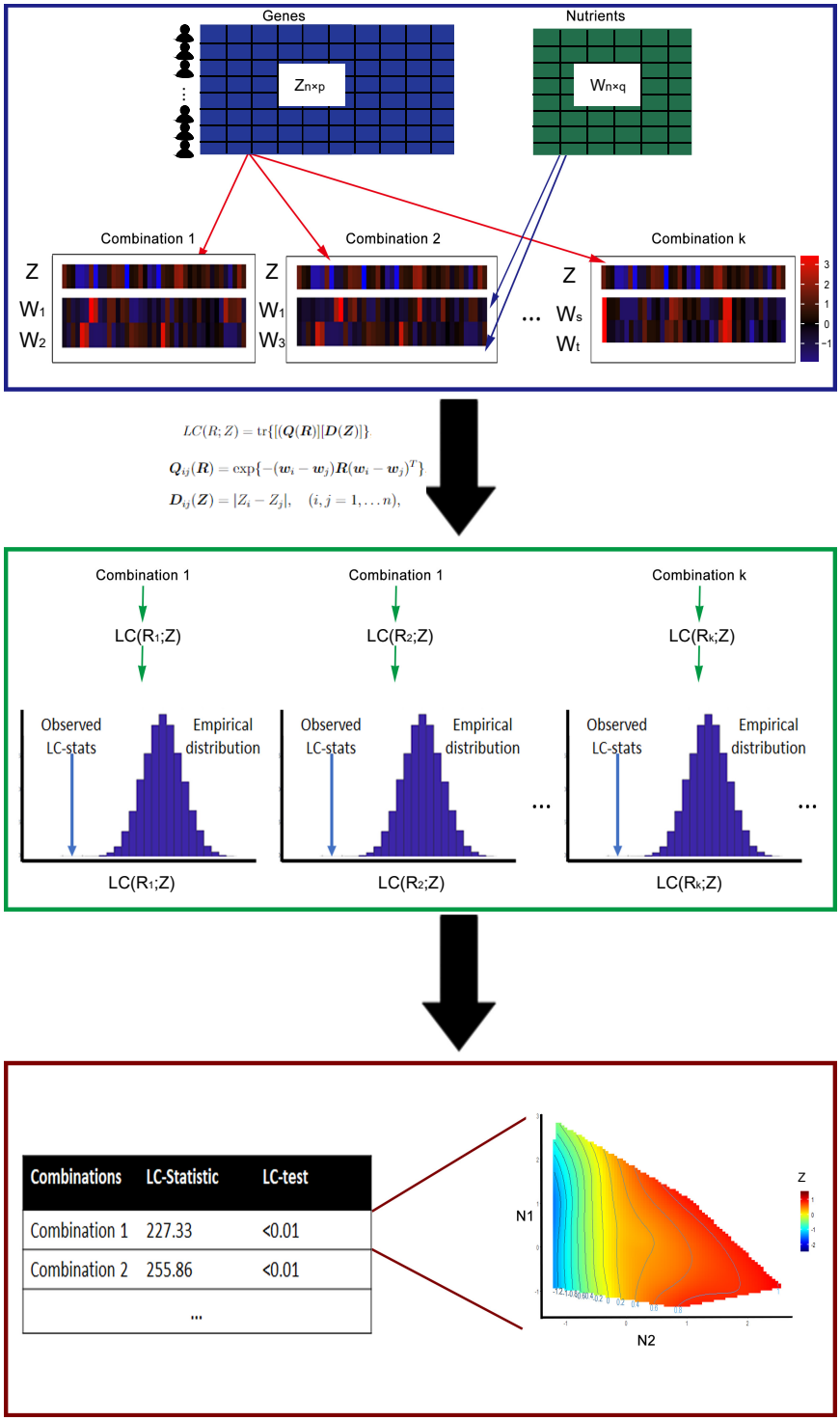

FIGURE 2.1. **Overall workflow of LC-N2G.** a) Data input: Gene expression data and Nutrients data. b) LC-N2G to rank and identify combinations of informative nutrition variables. c)-d) NGF for identified nutrition variables with a gene of interest (*Fgf21*).

2.2 Methods

2.2.1 Local Consistency Statistics

Let matrix $\boldsymbol{W}_{n\times q}$ denote the nutrition matrix where n is the number of rows (observations/samples), q the number of columns (nutrients/features), $\boldsymbol{w}_i, i = 1, \ldots, n$ denote the rows of the nutrition matrix and $\boldsymbol{Z}_{n\times 1}$ denotes the response, i.e. one particular gene expression vector. We define the LC-Stat as follows

$$LC(R; Z) = \mathrm{tr}\{[(\boldsymbol{Q}(\boldsymbol{R})][\boldsymbol{D}(\boldsymbol{Z})]\}, \tag{2.1}$$

where tr denotes the trace of a matrix and the elements of the matrices $\boldsymbol{Q}(\boldsymbol{R})$ and $\boldsymbol{D}(\boldsymbol{Z})$ are given by

$$\boldsymbol{Q}_{ij}(\boldsymbol{R}) = \exp\{-(\boldsymbol{w}_i - \boldsymbol{w}_j)\boldsymbol{R}(\boldsymbol{w}_i - \boldsymbol{w}_j)^T\}, \text{ and} \tag{2.2}$$

$$\boldsymbol{D}_{ij}(\boldsymbol{Z}) = |Z_i - Z_j|, \quad (i, j = 1, \ldots n), \tag{2.3}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are the i^{th} and j^{th} row of $\boldsymbol{W}_{n\times q}$, Z_i and Z_j are the i^{th} and j^{th} element of $\boldsymbol{Z}_{n\times 1}$, and $\boldsymbol{R}$ is a diagonal matrix consisting of 1's and 0's on the diagonal only. For example, if the first two nutrients are selected and $q = 3$ then

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, such a matrix $\boldsymbol{R}$ represents an indicator of subspace. The LC-Stat is a function of both the indicator matrix $\boldsymbol{R}$ and the given gene expression $\boldsymbol{Z}$. When $\boldsymbol{Z}$ is fixed, we can optimise $\boldsymbol{R}$ to find the combination of nutrition variables with minimal LC-Stat (see section 2.2.2). On the other hand, for a given combination of nutrition information, a permutation test can be performed to evaluate the significance of the relationship between this particular combination and $\boldsymbol{Z}$ (see section 2.3.2).

The basic idea of the LC-Statistic (LC-Stat) is that given a gene of interest with respect to certain nutrition variables, the response is smooth within a neighborhood in the nutrient space, i.e. for nearby observed points in the nutrient space, the response varies little. From the expression of the LC-Stat and focusing on the term in Equation (2.2), we can see that if two points are distant in the nutrient subspace selected by $\boldsymbol{R}$, no matter their difference in response, its corresponding $\boldsymbol{Q}_{ij}$ term will contribute approximately zero to the total of the LC-Stat. However, when two points are close, their difference in response will contribute to with a substantial proportion of the value of the LC-Stat. Thus, for a given dataset, a smaller LC-Stat means that data points have a slowly changing response in their neighbourhood while a larger LC-Stat means that some data points have a dramatically changing response in their neighbourhood.

Figure 2.2 provides an illustration of how the LC statistic works. In this figure, a total of four normally distributed random variables, W_1, W_2, W_3, and W_4, are considered and the response, Z, is generated by $Z = \exp\{-(W_1^2 + W_2^2)\} + 0.1\epsilon$, where ϵ is a standard normally distributed noise term. Figure 2.2(a) shows the NGF of the response Z with respect to the two informative variables W_1 and W_2 while Figure 2.2(b) shows the same response Z with respect to the two randomly generated non-informative variables W_3 and W_4. We can see clearly that Figure 2.2(a) shows a smoother transition across this 2-dimensional space and does not show a surface with high variability, with a corresponding LC statistic is 43.79. In contrast, multiple peaks and troughs can be seen when choosing random nutrition variables such as in Figure 2.2(b). In this situation, we observe also a much larger LC statistic of 61.42.

2.2.2 Identify informative combination of variables using LC-Stat (LC-Opt)

In Equation (2.1), if we do not add any constraint on the diagonal matrix $\boldsymbol{R}$ for a given gene, the smallest LC is obtained when all entries in $\boldsymbol{R}$ equal 1, i.e. all variables are selected. Hence, we consider the following optimisation problem for finding the smallest local consistency in

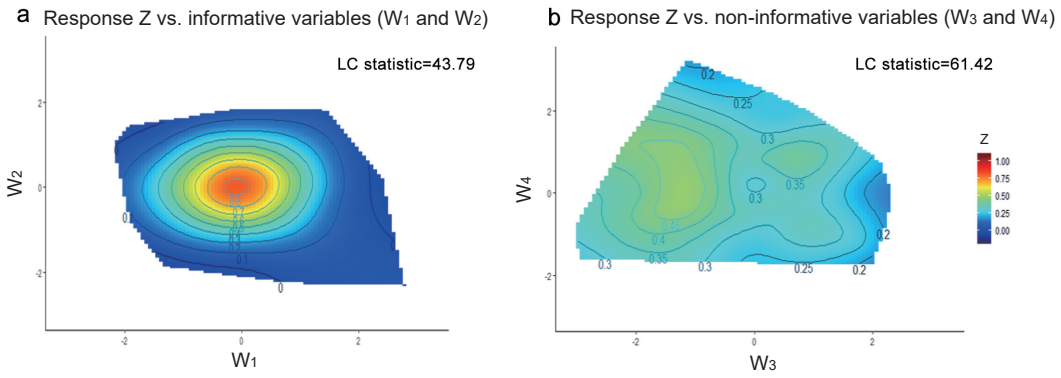

FIGURE 2.2. **Illustration of LC optimisation and LC-Test.** a) and b) are the NGFs of a simulated data which a) and b) use the same response Z while the covariate of a) is informative and b) is random.

k-dimensional space for a given gene:

$$\min_{\mathrm{tr}(\boldsymbol{R})=k} \quad LC(\boldsymbol{R}; \boldsymbol{Z}), \tag{2.4}$$

where $LC(\boldsymbol{R}; \boldsymbol{Z})$ is defined in Equation (2.1), k is the pre-specified number of covariates. All visualisation examples in this article consider $k = 2$ only.

The optimisation in Equation (2.4) is a binary optimisation and many methods exist to solve such a problem. In our implementation, we use an exhaustive search to calculate the LC-Stat for each considered combination of variables. This is computationally plausible when the number of combinations $\binom{q}{k}$ is not large. In situations where $\binom{q}{k}$ is large, we could instead use the Genetic Algorithm (GA) [51]. In the implementation of GA, $LC(\boldsymbol{R}; \boldsymbol{Z})$ is set to be the fitness function. In order to meet the constraints in Equation (2.4), a penalty term proportional to $|\,\mathrm{tr}(\boldsymbol{R}) - k|$ is added to the fitness function.

2.2.3 Evaluate significance of relationship between combination of variables and given gene expression using LC statistic (LC-Test)

In Equation (2.1), we can see that if there is a relationship between the gene expression and selected nutrition variables, the LC-Stat should be small. For given nutrition variables

$W_1, \ldots, W_k$ selected by the indicator matrix $\boldsymbol{R}$ and a gene expression $\boldsymbol{Z}$, we propose a LC-Test, a permutation test that uses the LC statistic to evaluate if the nutrition-gene relationship is significant. Under the alternative hypothesis of LC-Test, the response is modelled by $Z = f(W_1, \ldots, W_k) + \epsilon, \epsilon \sim N(0, s), s > 0$ and f is a non-constant smooth function, ϵ is random noise.

The procedure of LC-Test is described as follows: we first calculate $LC(\boldsymbol{R}; \boldsymbol{Z})$, the observed LC-Stat. Then we randomly permute the gene expression B times, and we denote the b^{th} permutation as $\boldsymbol{Z}_b$, $(b = 1, \ldots, B)$, and calculate corresponding $LC(\boldsymbol{R}; Z_b)$ statistic. The permutation p-value is then

$$p_{\text{perm}} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{LC(\boldsymbol{R}; \boldsymbol{Z}_b) < LC(\boldsymbol{R}; \boldsymbol{Z})\}, \tag{2.5}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function.

2.2.4 Simulation study

To evaluate the performance of LC-N2G, we consider three data generative models for the $(q + 1)$ nutrition variables and response, respectively:

$$\begin{aligned}
&W_1, ...W_q \sim N(0, 1),\ W_{q+1} = \frac{|W_1|}{|W_1| + |W_2|} + r\epsilon, \\
&Model\ 1:\ Z = \exp\{-(W_1^2 + W_{q+1}^2)\} + r\epsilon, \\
&Model\ 2:\ Z = \sin\{(\pi/2)(W_1^2 + W_{q+1}^2)\} + r\epsilon, \\
&Model\ 3:\ Z = \exp\{-(W_1^2 + |W_3|)/5\} + \exp\{-W_{q+1}^2/5\} + r\epsilon, \\
&Model\ 4:\ Z = \exp\{-(W_1^2 + W_2^2)/2\} + \exp\{-(W_3^2 + W_{q+1}^2)/2\} + r\epsilon,
\end{aligned}$$

where $\epsilon \sim N(0, 1)$ is random noise.

In these four models, we set the sample size n equal to 20, 50, 100 and the number of independent and identically distributed variables $(q - 1)$ equal to 9 and 19, respectively. Then we generate the q^{th} covariate being a non-linear function of W_1 and W_2. For this non-linear transformation, we choose a composition fraction as is common in nutrigenomics [140, 141,

145], e.g. a protein-carbohydrate proportion. The parameter r is used to model the level of noise and we vary r over 0.05, 0.1 and 0.2.

In Model 1, the response density has a single peak, which occurs regularly in real data, for example as seen in Solon-Biet et al. [145], Solon-Biet et al. [144], Le Couteur et al. [86] and Simpson et al. [141]. Many gene expression and other types of physiological variables with respect to the protein intake and carbohydrate intake have this kind of distribution. In Model 2, we try to consider a more complex response, that is with more than one peak with respect to the informative variables. In Model 3 and 4, more than 2 covariates are involved in explaining Z, we use all $k = 2, 3, 4$ to explore the results of LC optimisation for combination identification with different settings of k.

2.2.5 Mouse nutrition study

2.2.5.1 Data description

We apply our LC-N2G to a recent mouse nutrition study, which has been investigated in Solon-Biet et al. [145]. The dataset consists of 176 mice that were fed with different types of diet and water. Diets varied in protein (P), carbohydrate (C), fat (F), and energy (E) content. Energy manipulations were done through the addition of cellulose, allowing for low, medium, and high energy density diets (8, 13, and 17 kJ/g). Spearman correlation between nutrition variables are shown in Figure 2.3.

The liver microarray gene expression studies were performed on 48 livers across all diets with a total of 21,800 genes measured. Here, we consider 11 nutrients and all the genes. After excluding the samples with missing values, 42 samples remained for further consideration. In previous work [145], the gene *Fgf21* was investigated and has been found to be maximally elevated under low protein, high carbohydrate intakes via NGF.

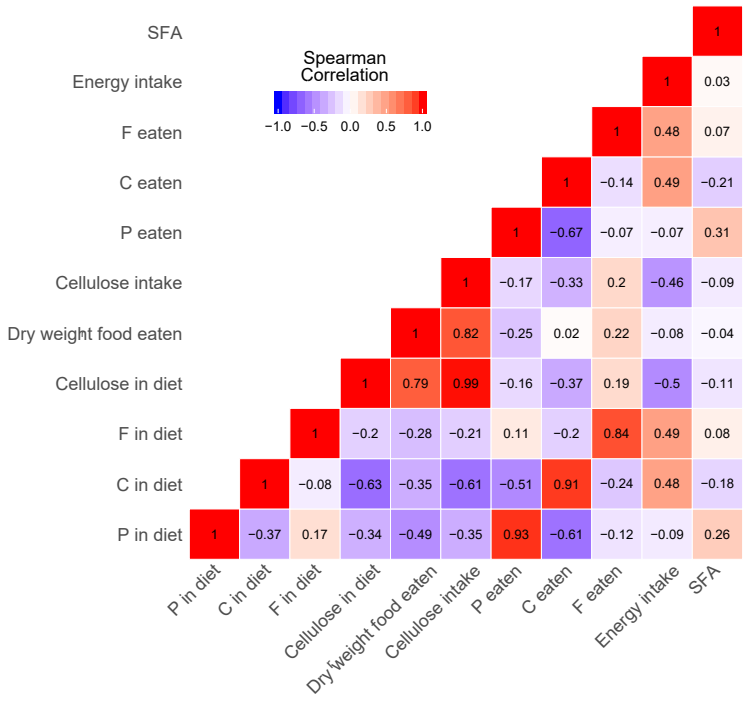

FIGURE 2.3. **Spearman correlation between nutrition variables in the mouse nutrition study.**

2.2.5.2 Selection of nutrition variables and genes of interest

To investigate the relationship of gene expression and nutrition variables, we select some representative genes in the dataset.

For simplicity, in this study 6 nutrition variables are excluded to avoid combinations with redundant information, such as (C, C eaten), (F, F eaten) etc. and the 5 nutrition variables: P eaten, C eaten, F eaten, Cellulose intake and SFA are selected. We perform LC-Test on these 5 nutrition variables on each gene. Then genes with no significant associations to any combination of nutrition variables, i.e. LC-Test p-value larger than 0.05, are removed. This results in 1,851 significant genes. Finally, genes are clustered into four modules using weighted gene co-expression network analysis (WGCNA) [193] and the corresponding hubs are *Ggcx*, *Slc27a5*, *Clec4d*, and *Adgrg2*. We use these selected genes and nutrition variables to analyse their potential relationship using NGF.

2.3 Results

2.3.1 Results on simulated data demonstrate that LC-Opt correctly identifies the most important covariates

To evaluate the ability of LC-Opt for identifying informative combinations of variables, we perform a simulation study. Three models in this study are described in section 2.3.1 and for each model we vary the number of independent nutrient variables $(q-1)$ over 9 and 19, with a 10^{th} and 20^{th} variable, respectively, which depends on the first two independent nutrient variables. The sample size n varies over 20, 50 and 100. We repeatedly generate data (100 times) for each of the model and for each combination of q and n. Then we calculate the LC-Stat for every combination of covariates. To summarise our empirical results, we group each combination of nutrient information according to how many of the informative variables it contains. For example, in Model 1 when $q = 10$, the informative combination is $(W_1,\ W_{10})$. We divide the combinations into 3 groups A_1, A_2 and A_3, say with $A_1 = \{$combinations include all informative variables$\} = \{(W_1, W_{10})\}$, $A_2 = \{$combinations include only one informative variable$\} = \{(W_1, W_2), \ldots, (W_1, W_9), (W_2, W_{10}), \ldots, (W_9, W_{10})\}$ and $A_3 = \{$combinations do not include any informative variable$\}$. The box plots of the LC-Stat for Model 1 and Model 3 ($k = 3$) with respect to each group are shown in Figure 2.4 and other settings are shown in Figure 2.5.

From Figure 2.4 and Figure 2.5, we can see that in all cases, where combinations include all informative variables, the LC-Stat is the smallest compared to the other combinations. The more informative variables are included in the combination, the smaller the LC-Stat. As expected, the larger the sample size n the smaller the LC-Stat (for all cases). When the number of samples decreases, the differences between groups with and without informative variables become smaller. Comparing the results with a different number of variables q, i.e. $q = 10$ against $q = 20$, we can see that the performance does not deteriorate as q becomes larger when n is held fixed. Results in these simulations show that our proposed LC-Opt has the ability to identify combinations of variables that have a true non-linear relationship with

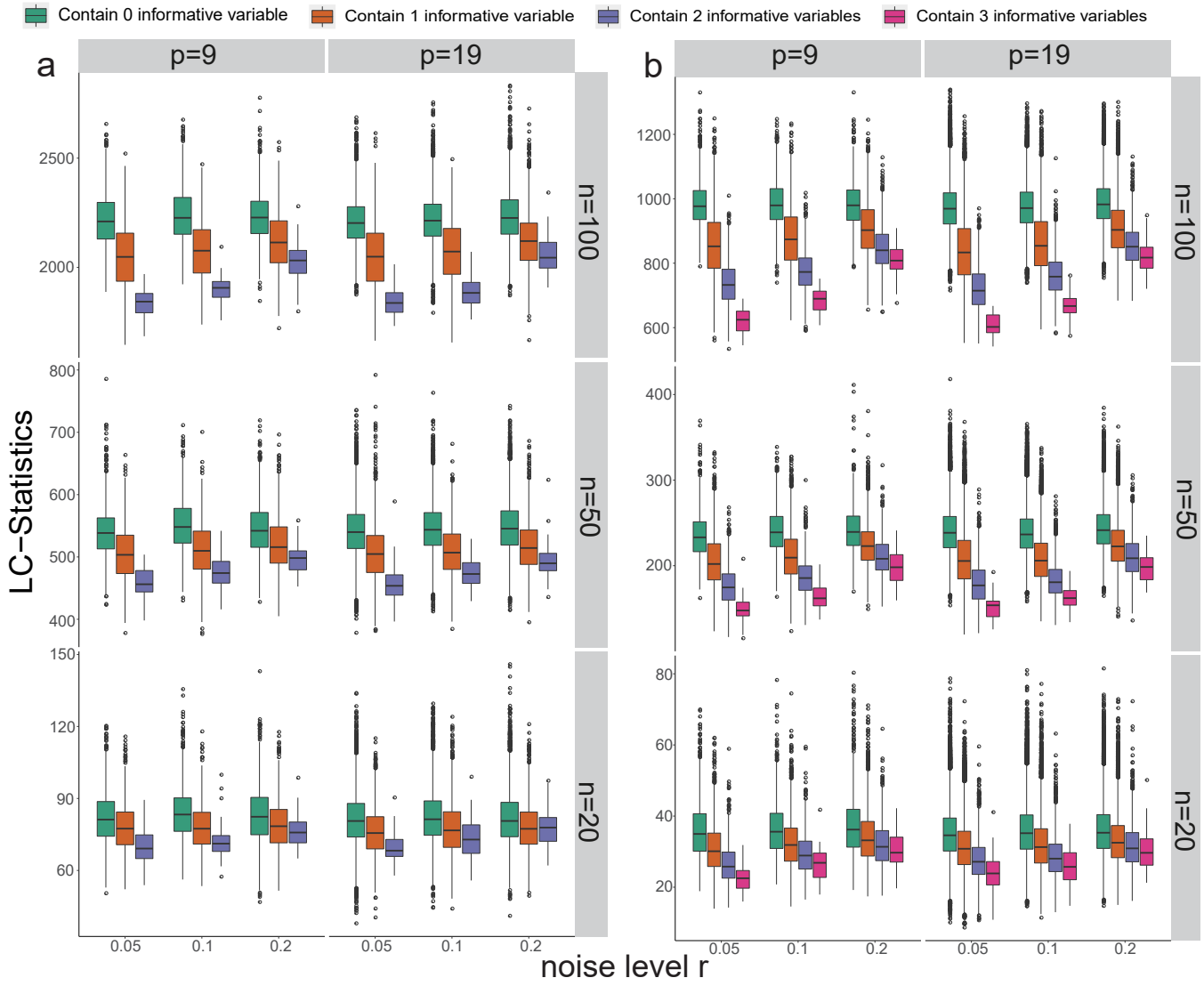

FIGURE 2.4. **Simulation results for LC optimisation for identifying combinations of Model 1.** The combinations are divided into 4 groups according to the informative variables it included. a) is Boxplots of LC statistics for different combination groups of Model 1. b) is Boxplots of LC statistics for different combination groups of Model 3 with $k = 3$. In a) and b) total numbers of informative variables are 2 and 3 respectively.

the response and it is quite stable when the number of variables is approximately as many as the number of observations.

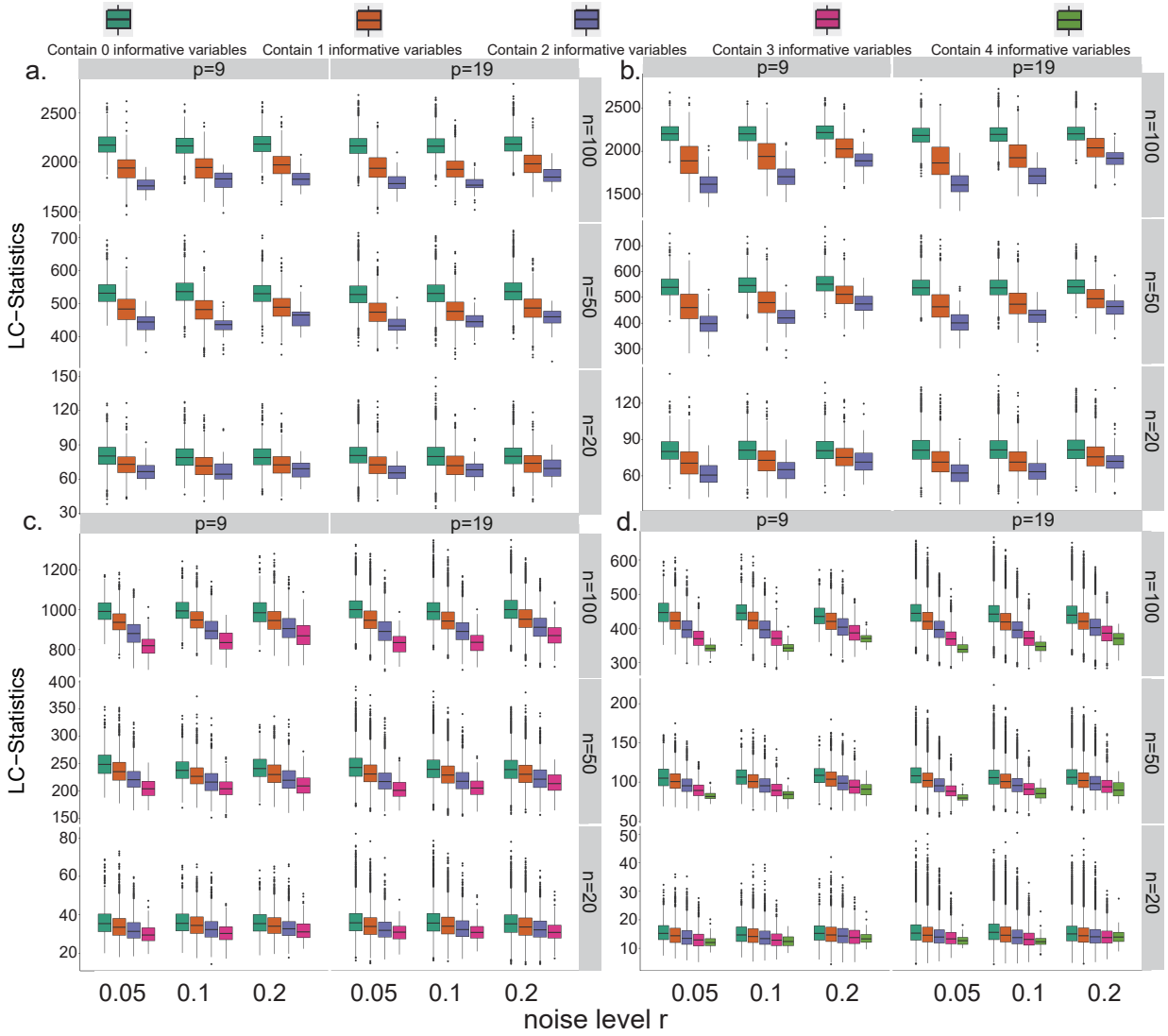

FIGURE 2.5. **Simulation results for LC optimization for identifying combinations of Model 2 and 3.** The combinations are divided into 5 groups according to the informative variables it included. Each model is evaluated under 3 different noise levels with different parameters r in the model. a) Boxplots of LC-Statistics for different combination groups of Model 2. b) Boxplots of LC-Statistics for different combination groups of Model 3 with $k = 2$. c) Boxplots of LC-Statistics for different combination groups of Model 4 with $k = 3$. d) Boxplots of LC-Statistics for different combination groups of Model 4 with $k = 4$. In a), b), c) and d) total numbers of informative variables are 2, 2, 3 and 4 respectively.

2.3.2 LC-Test: significance testing of the relationship between the combination of variables and given gene expression

The LC-Test tests if the selected variables have a non-random relationship with the response. In addition to finding the optimal nutrient components with the smallest LC statistics for a response, in practice, we may want to investigate the relationship between a specific response with certain nutrition components of interest. However, in some of the cases, the relationship of the nutrition variables and the genes is not significant, such as in Solon-Biet et al. [145], where the NGF of *Fgf21* with carbohydrate and fat is not significant.

We apply our LC-Test method to simulated data to evaluate if the relationship between combinations of variables and a particular gene expression is significant. In this simulation, the data are generated using three models (see section 2.3.1). Three tests are considered: our proposed LC-Test and two existing F-tests; first, the F-test in the linear regression model (F-test1) which tests under the null hypothesis whether all the regression parameters are equal to zero; and second, the F-test in the linear regression model which additionally considers the interaction term (F-test2) as in Cotter et al. [31]. For each model, we perform candidate tests on the informative variables (null scenario) and on randomly selected variables (alternative scenario). We simulate 200 data sets and then calculate the true positive rates (TPRs) and false negative rates (FPRs) at the 5% level and we sample again over sample sizes of 20, 50 and 100. The results of the top 20 nutrient combinations are shown in Table 2.1.

From Table 2.1 , it can be seen that in all cases, LC-Test always outperforms the two F-tests in terms of the TPRs while having relative small FPRs. Since the simulated response vector is generated using a non-linear function with respect to the covariates, which violates the assumption of both linear regression models with and without interaction terms, F-test1 and F-test2 result in relatively small TPRs in most cases except in Model 3 when $n = 100$. In the alternative scenario, F-test1 and F-test2 always show slightly smaller FPRs than LC-test, which means that they show similar performance in excluding non-informative variables. Comparing the results across different sample sizes, TPRs and FPRs somewhat deteriorate when n decreases for all the tests, demonstrating that a larger sample size is preferred.

Results in these simulations show that our proposed LC-Test has the ability to identify the combinations of informative covariates, for different models considered.

TABLE 2.1. Simulation result for LC-Test for identifying informative covariates. Results for TPR(FPR) at 5% level.

		Model 1	Model 2	Model 3
$n = 100$	LC-Test	0.97 (0.08)	0.99 (0.15)	1.00 (0.22)
	F-test1	0.19 (0.04)	0.19 (0.06)	0.69 (0.05)
	F-test2	0.17 (0.07)	0.22 (0.03)	0.83 (0.15)
$n = 50$	LC-Test	0.68 (0.07)	0.87 (0.14)	0.98 (0.18)
	F-test1	0.06 (0.06)	0.20 (0.07)	0.32 (0.03)
	F-test2	0.08 (0.03)	0.16 (0.04)	0.45 (0.09)
$n = 20$	LC-Test	0.53 (0.06)	0.69 (0.08)	0.84 (0.22)
	F-test1	0.01 (0.03)	0.08 (0.04)	0.15 (0.09)
	F-test2	0.01 (0.03)	0.09 (0.05)	0.34 (0.14)

2.3.3 Application of LC-N2G on mouse nutrition study reveals nonlinear relationships of dietary components and hub genes

We apply our proposed LC-N2G on a recent mouse nutrition study (see Section 2.2.5). We first use the LC-N2G approach to examine a specific gene *Fgf21* to validate the findings presented by Solon-Biet and colleagues [145].

The results are summarised in Table 2.2 and Figure 2.6(a). In Table 2.2 we see four sets of variable combinations, where in each either one variable is Protein (protein that is fed to each mouse) or Protein eaten (protein that is eaten by each mouse) and another is Carbohydrate or Carbohydrate eaten, rank highly, i.e. with small LC statistics and small corresponding p-values. This indicates that gene *Fgf21* shows a non-linear relationship with protein and carbohydrate. Figure 2.6(a) shows the gene expression values for *Fgf21* with "Protein eaten" and "Carbohydrate eaten", this visualisation result confirms prior knowledge that *Fgf21* response is largest in the low protein, high carbohydrate region.

To further investigate the relationship of nutrition and liver gene expression in this dataset, we perform LC-N2G for selected genes of interest to find potential relationships with nutrition variables. This analysis involves five nutrition variables, resulting in $\binom{5}{2} = 10$ pairs of

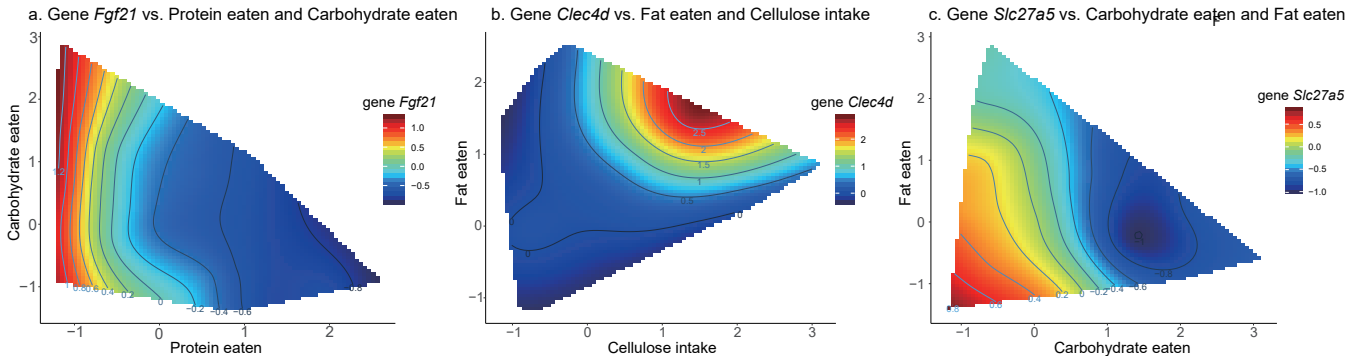

FIGURE 2.6. **Simulation results for LC optimisation for identifying combinations.** a)-c) NGF of *Fgf21*, *Slc27a5* and *Clec4d* with the informative combination identified by LC-N2G. *Fgf21* is investigated in [145]. *Slc27a5*, *Clec4d* are hub genes by WGCNA.

TABLE 2.2. The top 20 combinations of nutrition variables selected by LC-Opt for mouse nutrition study. The p-values are the permutation p-values from the LC-Test.

Combination			
Variable 1	Variable 2	LC statistic	p-value
Protein	Carbohydrate	85.02	0
Cellulose	Carbohydrate eaten	86.30	0
Carbohydrate	SFA	86.59	0
Carbohydrate	Protein eaten	88.58	0
Carbohydrate eaten	SFA	90.37	0
Carbohydrate	Dry weight food eaten	90.48	0
Dry weight food eaten	Carbohydrate eaten	91.20	0
Cellulose	Protein eaten	92.30	0
Carbohydrate	Cellulose	93.40	0
Protein	Carbohydrate eaten	96.38	0
Protein eaten	Carbohydrate eaten	96.55	0
Cellulose intake	Carbohydrate eaten	98.66	0
Carbohydrate	Energy intake	100.78	0
Carbohydrate	Fat	101.28	0
Protein	Cellulose	102.73	0
Protein eaten	Energy intake	103.63	0.005
Carbohydrate	Cellulose intake	103.69	0
Carbohydrate eaten	Energy intake	104.93	0
Dry weight food eaten	Protein eaten	106.61	0
Fat	Carbohydrate eaten	107.07	0.01

combinations and four genes (Ggcx, Slc27a5, Clec4d and Adgrg2). These four hub genes

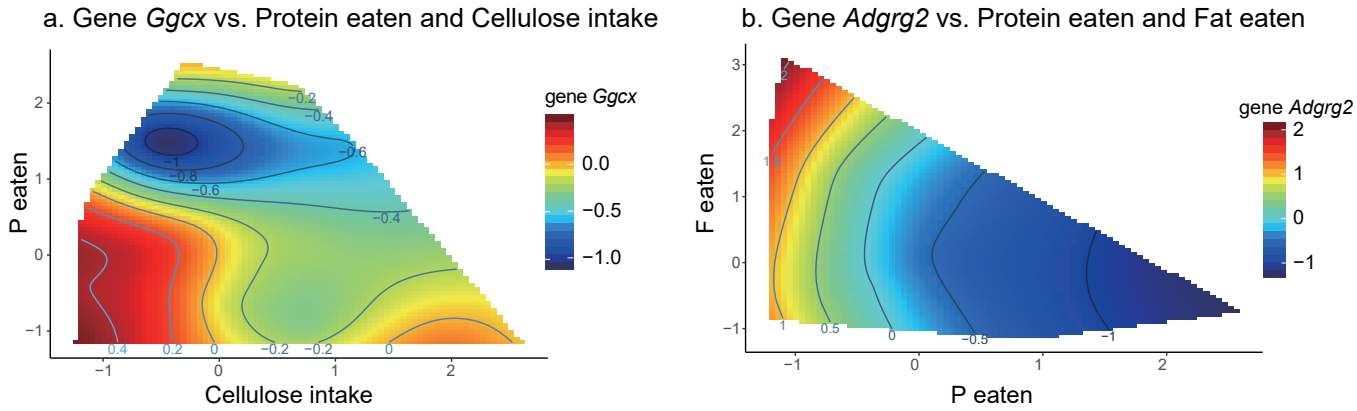

FIGURE 2.7. **a), b) NGF of *Ggcs* and *Adgrg* with the informative combination identified by LC-N2G.**

represent four distinct expression profiles selected after applying LC-test and the WGCNA [193] to the gene expression data (see Section 2.2 for more details).

Figure 2.6 (b) and (c) identifies the two key nutrition variables (Fat, Carbohydrate eaten) and (Fat, Cellulose intake) that are most associated with genes *Slc27a5* and *Clec4d*, respectively. In particular, "Fat" is identified to relate to both genes, which is consistent with current studies [11, 26, 48, 168] that recognised that fat level affects the expression level of these genes. It is interesting to note that the Spearman correlation of "fat eaten" with *Clec4d* and *Slc27a5* is -0.44 and 0.12, respectively. This indicates that LC-N2G has an ability to identify nutrition variables that may not have a marginal effect with a considered gene. Further results for genes *Ggcx* and *Adgrg2* are shown in Figure 2.7.

2.4 Discussion

In this chapter, We present a novel statistical framework, LC-N2G, that facilitates the examination of the relationship between a large number of nutrition variables and gene expression data. This involves developing approaches to estimate the stability of values in the two dimensional surface using a novel local consistency metric. By applying LC-N2G to simulated data, we demonstrate that our method can accurately recover the correct combination of nutrition variables related to a gene of interest. Furthermore, application to a real dataset not only

confirms the finding that *Fgf21* response is largest in the low protein, high carbohydrate region but we also find some potential non-linear relationship between nutrition variables with gene expressions.

We point out that the same form of $\boldsymbol{Q}(\boldsymbol{R})$ in Equation (2.1) is widely used in machine learning, especially in metric learning, such as in Neighbourhood Component Analysis (NCA) [52]. However, we do not constrain $\boldsymbol{Q}(\boldsymbol{R})$ to be normalised to a probability distribution, where the sum of each row equals 1, due to the effect of each sample not being the same when the response is involved. Another interesting fact is that the formula of the LC-Stat is similar to the objective function in spectral clustering [7], which aims at finding the best cut for a given graph structure. Spectral clustering does not make strong assumptions on the statistics of the data and shows good performance when data show some kind of sparsity [163]. Therefore, our method also enjoys these properties. One difference between our method and spectral clustering is that we substitute the square term with the absolute value; this alleviates the influence of possible outliers in the response. The main difference is that in spectral clustering, $\boldsymbol{R}$ is pre-specified from a certain structure in data, such as a network structure, and then the function optimised over $\boldsymbol{Z}$, while in our method, the response $\boldsymbol{Z}$ is observed from data, usually a specific gene expression level, and we try to find a good metric $\boldsymbol{R}$, under which LC is small enough.

In LC-Opt, the matrix $\boldsymbol{R}$ is a diagonal matrix consisting of 1's and 0's on the diagonal only. We use this set due to the fact LC-Opt aims to find a combination of nutrition variables with the smallest LC statistic. In a general sense, the matrix $\boldsymbol{R}$ can be considered as a semi-definite matrix. This makes the LC-Opt flexible enough to find several latent variables, linear combinations of nutrition variables as in principal component analysis (PCA) or partial least squares (PLS) [174], with small LC statistic.

We also noticed that the proposed LC statistic shares some similarities with correlation-based methods such as canonical correlation analysis (CCA) [55] and partial least squares (PLS) [174]. These two methods maximise the correlations between linear combinations of nutrition features and gene expression. However, in the nutriomics context, such correlation-based methods are better suited to capture linear relationships and their focus is on global patterns

[79]. Our LC-Stat, on the other hand, deals with the non-linear relationship between nutrition and gene expression. In addition, the proposed LC-Test is also closely related to the Mantel test[105], which tests the correlation between two distance matrices and is widely used in ecology. However, the Mantel test depends on the choice of distance measurement and does not incorporate the variable selection procedure. By introducing the diagonal matrix $\boldsymbol{R}$, our LC-N2G could be viewed as an extension of Mantel's test in the nutriomics context.

In practice, covariates always have different scales. Most of the macronutrition variables such as protein, fat and carbohydrate are measured in mg, while some micronutrition covariates such as vitamin are measured in International Unit (IU). Different scales impose non-uniform weights on the $LC(\boldsymbol{R})$ statistic and the optimisation result will tend to select the variable with largest range, as it makes the similarity matrix smallest. Thus, we standardised nutrition variables via $z-$score, which is centred by the mean and scaling by the standard deviation of a variable, to make their scale uniform.

The proposed LC-N2G framework is based on nutrigenomics. However, it is not restricted to exploring the relationship between nutrition and gene expression. Our method has potential to be used in other circumstances to identify a combination of variables that have a non-linear relationship with a certain response. One possible application for LC-N2G is in multi-omics, where the data interact in a complex way. The framework of LC-N2G can be easily extended by replacing nutrition data with one of the omics data such as metabolomics data and using another omics data or some phenotype as the response, which provides a method for understanding the interaction across multi-platform data.

We have developed a shiny webpage `http://shiny.maths.usyd.edu.au/LC-N2G/` that performs LC-N2G and provides an approach to further investigate the association between nutrients and gene expression for the mouse nutrition study.

In summary, we present LC-N2G, a novel statistical framework for ranking and identifying relationships between nutrition and gene expression information. LC-N2G finds combinations with small LC-Stat and tests these combinations using a permutation test to distinguish effects that are different to those from purely random combinations. We applied LC-N2G to both

simulated and real datasets and showed that this framework can accurately select combinations of nutrition variables that are related to a gene of interest. LC-N2G is implemented as in R and all the functions used in this chapter can be downloaded freely from https://github.com/SydneyBioX/LCN2G our GitHub repository.

CHAPTER 3

Identification of omics features clusters in experimental nutriomics data

---

Easier and more comprehensive access to large omics data such as genomics and proteomics data has not only revolutionised biological research [4] but also nutrition research. The emerging field of nutriomics refers to the joint study of omics and nutrition data which facilities a better understanding of the interactions between nutrients and the host, which unlocks novel insights on personalised nutrition [112].

Under the nutritional geometry framework (NGF) [141], we propose a method (Chapter 2) to discover the non-linear relationship between nutrition intake and selected omics features . A subsequent research question, as is pointed out in [141], is to understand how these omics features respond to experimental conditions. One widely used approach is to group these p different omics profiles into clusters and illustrate how these clusters of omics profiles depend on nutrition and other experimental factors [21, 120].

In this chapter, we present **eNODAL**, an **E**xperiment-guided **N**utri**O**mics **DA**ta c**L**ustering framework that clusters omics features based on their response to nutrition intake as well as other experimental factors. eNODAL uses a two-stage clustering approach to group the high-dimensional proteomics features into several interpretable subclusters with similar responses to the experimental factors as well as their interactions. eNODAL first performs a non-linear ANOVA-like test to categorise proteomics features into interpretable groups based on whether nutrition, drug treatment and their interaction show significant effects (Section 3.3.1). Then, in Section 3.3.2, an ensemble clustering further divides each group into subclusters based on their abundance profiles. We illustrate eNODAL by applying the workflow on a mouse nutrition study [87]. Applying eNODAL on a mouse proteomics study, we found that among the $2,951$ proteins that were significantly affected by nutrition and drug intake,

about half of these proteins $(1,350)$ are shaped by nutrition intake (see Section 3.3.2). These proteins are enriched in pathways related to splicesome, consistent with previous research [87]. Furthermore, in Section 3.3.3 and Section 3.3.4, eNODAL demonstrates its ability to identify clusters of proteins that are affected by interactions between nutrition information and drug intake. We observe that in a particular subcluster ("NxT_C4") where proteins show a clear response to the ratio of protein and carbohydrate intake in the presence of the drug metformin but not when rapamycin or resveratrol are present. This clusters is enriched with biological meaningful pathways such as AMPK.

3.1 Background

3.1.1 Domain background - nutrition

Understanding how nutrition interacts with health is important for precision nutrition and public health [156]. Throughout much of the 20th century, nutrition science was heavily focused on diseases that were caused by deficiency of specific nutrients [169]. Moving into the 21st century, many populations are plagued by non-communicable diseases that are not seemingly caused by deficiency per se, for example, metabolic syndrome, type-2 diabetes, obesity and cardiovascular diseases [47]. In providing solutions to these diseases, the 'single-nutrient-at-a-time' paradigm has largely failed [113].

One issue when applying the historical approach to these non-deficiency diseases is that both nutrition and the biological processes underlying the pathology are multidimensional. Nutrients which can have both individual and synergistic effects, typically co-vary with the foods that we eat [152]. Even at the macro level, the three primary energy sources: protein, carbohydrate, and fat, interact to affect metabolism, and cognitive functions [138]. What further complicates the challenge is that diets also possibly interact with lifestyle factors and medical interventions such as drug treatments [38]. Understanding the physiological effects of nutrients is also a multivariate task, requiring the assessment of the relationships between nutrient intake and a myriad of 'intermediary' phenotypes, such as those produced

from large-scale 'omics' experiments. These experiments produce datasets that shed light on the role of nutrition and drug in shaping health outcomes. However, it is statistically and computationally challenging to analyse the datasets produced from these experiments. One of the key difficulties is the high dimensionality of omics data, where the number of omics features tends to go much larger than the sample size. This makes it hard to understand how experimental factors including nutrition intake, drug treatment as well as their interaction, shape the omics response.

3.1.2 Data science challenges in pattern detection

A common strategy to address this issue is to group the high dimensional omics features into highly correlated clusters and then analyse the relationship between these clusters and experimental factors. Several methods serve for this procedure such as weighted correlation network analysis (WGCNA) [82] and ClustOfVar [21], which use an unsupervised way to cluster omics features based on the correlation structure or their abundance value. Such methods have been widely used in analyzing genomics and proteomics data [120]. However, these unsupervised clustering methods do not take the experimental information into account. In nutriomics study, the resulting clusters may be confounded with the information from the experimental design. We illustrate such a concept by a toy example shown in Figure 3.1: (a) and (b) show the abundance of proteins A and B. They have different patterns in the presence of drug1 and drug2. However, their Spearman correlation is still high. This makes these proteins easily clustered together when we only consider their marginal correlation. Furthermore, the unsupervised clustering methods do not provide any biological interpretation of the resulting clusters, which make it hard to understand how experimental factors affect the responses to these feature clusters [122].

3.1.3 Dataset description

Here we explore data from a study on the interactive effects of dietary macronutrients and drugs that target the aging mechanisms (DTAM) in mice [87]. In a nutshell, male

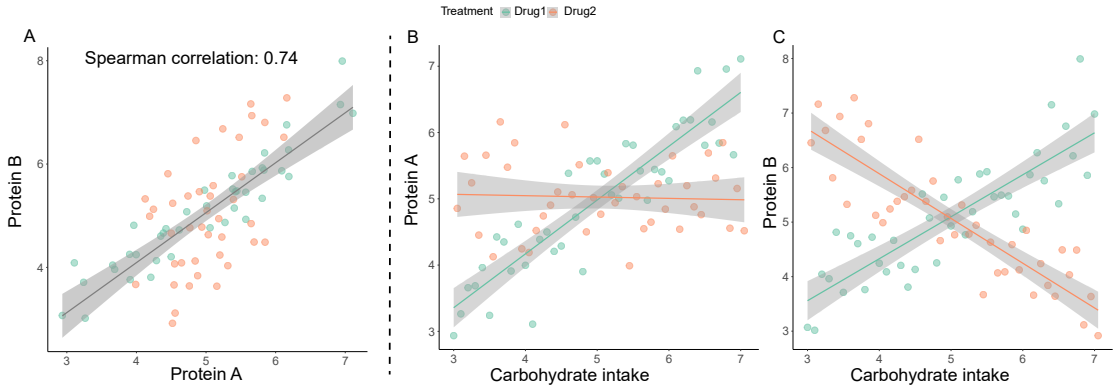

FIGURE 3.1. **A toy example illustrates that proteins with different response to nutrition and drug intake could be highly correlated** a) Protein A and B are highly correlated b) Protein A showed a positive correlation with Carbohydrate intake in the presence of drug 1. c) Protein B showed a different correlation with Carbohydrate intake with drug 1 and 2.

C57BL/6J mice were kept on one of ten different diets. These diets had five different ratios of macronutrients (i.e., $\%$ energy from protein, carbohydrate, and fat), and were replicated at two energy densities (8 kJ/g and14.8kJ/g). Layered over this nutritional design, animals were also on a control (no-drug) treatment or one of three DTAM: metformin, rapamycin, or resveratrol. In total, the fully factorial experimental design consists of 40 different treatments. Key metabolic traits as well as food intake were measured, and abundance of the liver proteome was quantified. The aim is to understand how drug-diet interactions affect the proteome, and their downstream impact of metabolic health.

The dataset has many of the challenges common to nutritional-omics datasets. First, there are a large number of outcomes; about $5,000$ proteins were quantified. Second, the effects of multiple nutritional dimensions are captured; the experiment utilised 10 diets that covaried in protein, carbohydrate, and fat content and overall energy density. Third, interactions between nutritional and non-nutritional factors are explored; each diet was replicated across three different drug treatments and one control. Finally, given the constraints in performing dietary experiments the sample size precludes the application of machine-learning approaches.

3.2 Overview of eNODAL

In this chapter, we propose a novel statistical workflow for an **E**xperiment-guided **N**utri**O**mics **DA**ta c**L**ustering framework, which we coin eNODAL. eNODAL first uses a non-linear ANOVA-like model to distinguish whether the protein shows significant effects on nutrition, drug intake, or their interaction. Then a consensus clustering method is performed to further cluster proteins with similar response profiles. We utilise proteomics data from a recent designed experiment in mice [87]. Our approach allows us to cluster the proteomics features as well as to link them to the phenotypes. Through the analysis of the dataset, we group the proteomics features into 29 interpretable subclusters with different responses to nutrition intake and drug treatment. The pathways enrichment analysis identified biologically meaningful pathways closely related to the interaction effect.

3.2.1 Experiment-guided nutriomics data clustering (eNODAL) method

eNODAL hierarchically groups high-dimensional omics features guided by the nutritional and drug experimental design (see Figure 3.2). It first categorises the omics data into interpretable groups based on whether the group of omics features shows significant effects of nutrition, drug treatment, or their interaction using an ANOVA-like hypothesis testing procedure[194]. For the current mouse data, we obtain five interpretable groups with the following respective main and interaction effects:

- "NxT" group: nutrient and drug main effect as well as their interaction
- "N+T" group: nutrient and drug main effect
- "N" group: nutrient main effect only
- "T" group: drug main effect only
- "non-sig" group: intercept only group.

eNODAL then further divides these interpretable groups into subclusters to reflect distinct patterns of omics features. The clustering is based on a consensus clustering strategy [77] integrating a variety of clustering methods and distance measurements (see Methods 3.2.2).

Besides the widely used correlation distance[77, 82], two weighted correlation distances are proposed to account for the similarity in terms of the roles of nutrition intake, drug treatment, and their interactions (see Methods 3.2.4). Such weighted correlation distance dynamically focuses on the correlations between variables as well as their contribution to experimental factors (see Figure 3.3).

To better understand the properties of these subclusters, eNODAL then annotates them in two aspects: experimental responses, and pathway enrichment. We first annotate these subclusters based on three sets of interpretable features described in Section 3.2.2.3. Such annotation reflects how the proteomics features respond to nutrition, drug treatment, and their interactions. We then perform enrichment analysis for each subcluster based on KEGG pathways [68] to annotate these subclusters from a biological view. This gives insights into the biological function of the subclusters.

In a final step, eNODAL integrates these subclusters, proteomics features, and metabolic phenotypes. We use a multiset hypothesis test procedure to determine whether the relationships between the subclusters and the phenotypes are significant [116]. This procedure further sheds light on how proteins are affected by nutrition and how it shapes metabolic phenotypes.

3.2.2 Two-stage clustering

The eNODAL framework uses a two-stage clustering method to group the high-dimensional omics features into subclusters. We will describe the details of each stage in the following sections.

3.2.2.1 Nonlinear ANOVA-like test

The development of the first stage clustering method is inspired by a nonparametric ANOVA (NANOVA) method which was first proposed to classify genes into different groups based on their factor effect [194]. We extend this method to categorise the proteomics features based on their response to a group of continuous variables (nutrition features) and a four-level categorical factor (drug treatment). We further consider the non-linear relationships among

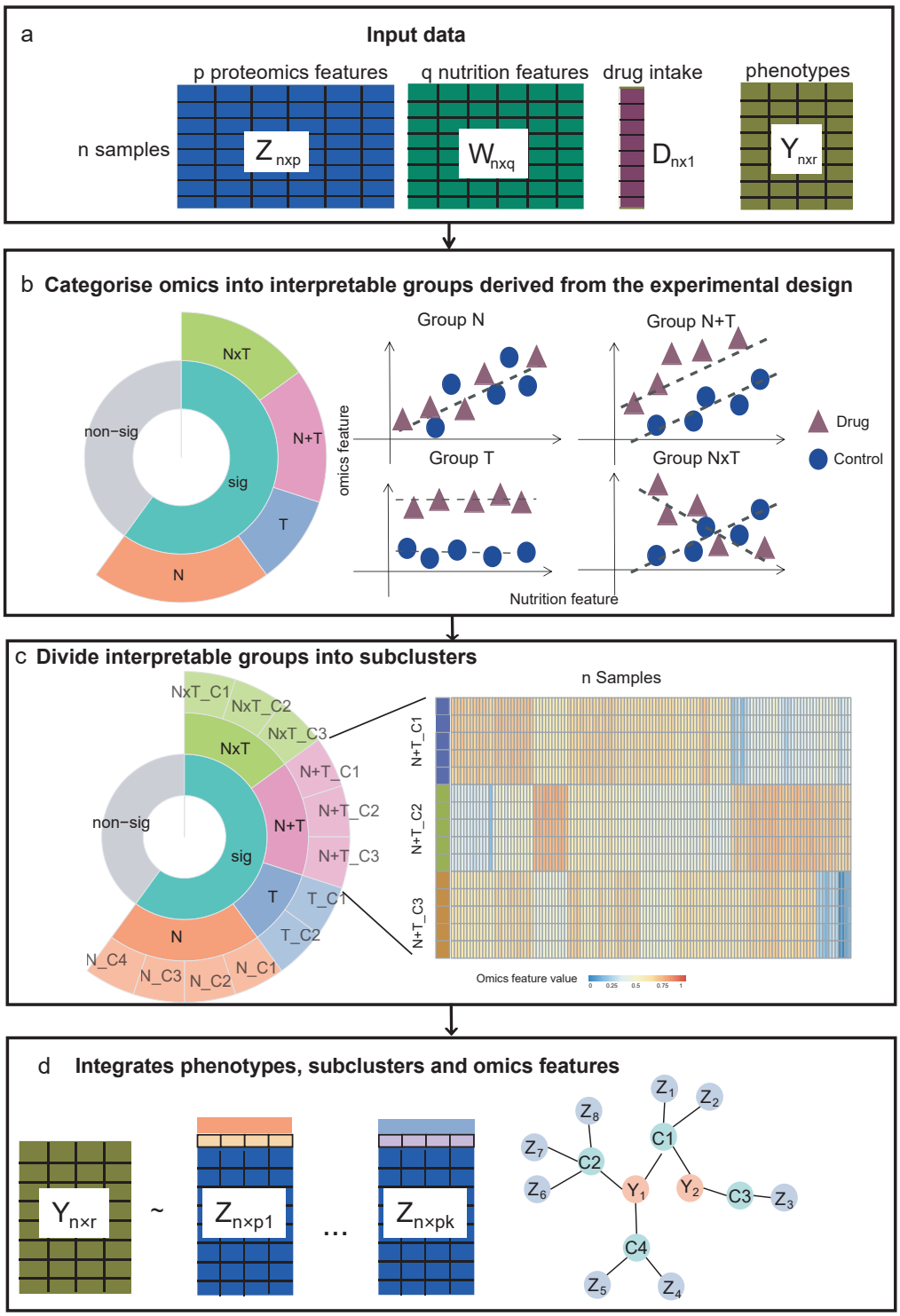

FIGURE 3.2. **An overview of eNODAL** a) Input of eNODAL including nutrition data, drug intake, omics data and metabolic phenotypes; b) Categorise omics features into interpretable groups derived from the experimental design by ANOVA-like test; c) Divide interpretable groups into subclusters via ensemble clustering; d) Integrate phenotypes, subclusters and omics features via multiset testing.

nutrition, treatment, and proteomics features. We define the five nested models (M1, M2, . . ., M5) as follows:

$$\text{M1:} \quad z_{ijk} = \mu + f_j(w_i) + \alpha_{jk} + g_{jk}(w_i) + \epsilon_{ijk}, \tag{3.1}$$

$$\text{M2:} \quad z_{ijk} = \mu + f_j(w_i) + \alpha_{jk} + \epsilon_{ijk}, \tag{3.2}$$

$$\text{M3:} \quad z_{ijk} = \mu + \alpha_{jk} + \epsilon_{ijk}, \tag{3.3}$$

$$\text{M4:} \quad z_{ijk} = \mu + f_j(w_i) + \epsilon_{ijk}, \tag{3.4}$$

$$\text{M5:} \quad z_{ijk} = \mu + \epsilon_{ijk}, \tag{3.5}$$

where z_{ijk} is the j^{th} proteomics feature for the i^{th} sample which received the k^{th} DTAM ($k = 1$ represents control group); μ is the overall effect; α_{jk} is the k^{th} treatment effect on j^{th} protein (in the mouse nutrition study, we have four different treatments corresponding to different drug intake); w_i is the nutrition features of i^{th} sample and f_j and g_{jk} are smooth functions representing the relationships between nutrition and proteomics features. For M1, different smooth functions f_j and g_{jk} are used to account for the main effects of treatment and nutrition as well as their interaction. For M2, f_j models the nonlinear relationship contributed by nutrition. Proteins of M2 are affected by both nutrition and drug, but their effects are independent. Proteins in M3 and M4 are only affected by nutrition and treatment effects in the model; Proteins in M5 are not influenced by any of the experimental conditions. As is in the usual ANOVA model, constraints $\sum_{k=1}^{4} \alpha_{jk} = 0$ and $\sum_{k=1}^{4} g_{jk}(w_i) = 0 \; \forall j$, are imposed for identifiability [194].

Next, we categorise all proteins into five interpretable groups. Each group corresponds to one of the above five nested models, that is M1, M2, . . ., M5, based on non-linear ANOVA-like testing as described below. We denote the set of all proteomics features as S.

The non-linear ANOVA-like testing proceeds as follows:

(1) First, we identify proteins whose abundance are affected by either nutrition or treatment. This is achieved by the LC-test described in Chapter 2, which tests whether the effect of nutrition and treatment is significantly different from randomly permuted protein abundance.

Proteomics features that show a significant response are assigned to the cluster "sig", denoted as C_0, and otherwise are assigned to the cluster "non-sig".

(2) For the proteins in cluster "sig", we use a nested ANOVA test as in [176, 178] for the generalised additive model (GAM) to test whether the interaction effect in M1 is significant. That is, we test for each proteomics feature, $H_0 : g_{j1} = g_{j2} = g_{j3} = g_{j4} = 0$ versus H_1 : at least one g_{jk} is not equal to zero. The set of proteins with a significant interaction effect is denoted as $C_{int} \subset C_0$

(3) For the proteins in set $C_0 \setminus C_{int}$, we fit M2 and test whether coefficients α_{jk} (i.e. for each j, $H_0 : \alpha_{jk} = 0, \quad \forall k$ vs. H_1 : at least one $\alpha_{jk} \neq 0$) and f_j (i.e. for each j, $H_0 : f_j = 0$ vs. $H_1 : f_j \neq 0$) is significant. Such a test also can be done via nested ANOVA tests in GAM. Proteins with $\alpha_{jk} \neq 0$ and $f_j = 0$ are classified as cluster "T", denoted as C_T, and those with $\alpha_{jk} = 0$ and $f_j \neq 0$ are classified as group "N", denoted as C_N.

(4) All proteins in $C_0 \setminus (C_{int} \cup C_N \cup C_T)$, form the "N+T" group.

After fitting the models and calculating the p-values, we use a hierarchical p-value adjustment [110] to correct the p-value. Then the Bonferroni method is used to control the false discovery rate. Through this procedure, we classify the proteins into five interpretable groups, i.e. "NxT", "N+T", "N", "T" , and "non-sig".

3.2.2.2 Consensus clustering

Based on the categorised five interpretable groups, we further divide the groups into sub-clusters using unsupervised clustering methods. We use a consensus clustering method [77] with different types of distance measurements described in Section 3.2.4 and varieties of clustering methods including affinity propagation [14, 43], Louvain clustering based on k-nearest neighbor graph [126] and a dynamic tree cut method for hierarchical clustering [83]. These methods use a data-driven way to find the number of clusters and adapt well to the complexity of individual datasets.

Then a consensus matrix [148] is created based on each individual clustering result. A binary similarity matrix is constructed from the corresponding clustering labels: if two features belong to the same cluster, their similarity is 1; otherwise, their similarity is 0. Finally, the resulting consensus matrix is clustered using the Louvain algorithm to get the resulting subclusters for each interpretable group.

3.2.2.3 Subcluster annotation

After the two-stage clustering, eNODAL provides two ways to annotate these subclusters from different perspectives: one angle is the relationship between nutrition intake and drug treatment, which is based on three sets of interpretable features described in the following parts; another perspective is based on the biological pathway information, which is done by a KEGG pathway enrichment analysis. The detail of the annotation of each subcluster is presented below.

Calculate interpretable features for each protein: Let z_{jk} denote the j^{th} proteomics measurement in the k^{th} drug treatment group ($k = 1, \ldots, 4$, where $k = 1$ represents the control group), the corresponding l^{th} nutrition intake is denoted w_{lk}, and $z_j = (z_{j1}, z_{j2}, z_{j3}, z_{j4})$, $w_l = (w_{l1}, w_{l2}, w_{l3}, w_{l4})$. In the mouse nutrition study, we focus on four nutrition intake features ($l = 1, \ldots, 4$), i.e. raw food intake in grams, and protein intake, carbohydrate intake, and fat intake in kJ. Three sets of interpretable features for the j^{th} proteomics measurement are described in the following.

Set 1: We first calculate the Fisher's z-test statistic, $\mathcal{Z}_{jl}$ $(l = 1, 2, 3, 4)$ from the correlation coefficients of proteomics feature z_j and nutrition feature w_l:

$$\mathcal{Z}_{jl} = \frac{1}{2} \ln \left(\frac{1 + \text{cor}(z_j, w_l)}{1 - \text{cor}(z_j, w_l)} \right),$$

where $\text{cor}(z_j, w_l)$ is the sample correlation coefficient between protein z_j and nutrition w_l. Each interpretable feature in Set 1 is calculated by $\sqrt{n-3}\mathcal{Z}_{jl}$ where n is the number of observations.

Set 2: We calculate the pairwise t-statistic of differential abundance of z_{jk} between control ($k = 1$) and each treatment group ($k = 2, 3, 4$):

$$\mathcal{T}_{jk} = \frac{\bar{z}_{jk} - \bar{z}_{jk}}{\sqrt{s_{jk}^2/n_k + s_{j1}^2/n_1}},$$

where $\bar{z}_{jk}$ and $\bar{z}_{j1}$ are the sample mean of protein abundance in drug group k and the control group, respectively; s_{jk}, s_{j1} and n_k, n_1 are the corresponding sample standard deviation and sample size respectively. Set 2 is composed of $\mathcal{T}_{jk}$, $(k = 2, 3, 4)$.

Set 3: Fisher's z-test statistic of differential correlations of z_{jk} and w_{lk} between the control ($k = 1$) and treatment groups ($k = 2, 3, 4$) is calculated:

$$\tilde{\mathcal{Z}}_{jlk} = \frac{\mathcal{Z}_{jlk} - \mathcal{Z}_{jl1}}{\sqrt{1/(n_k - 3) + 1/(n_1 - 3)}},$$

where $\mathcal{Z}_{jlk} = \frac{1}{2} \ln \left(\frac{1+\text{cor}(z_{jk}, w_{lk})}{1-\text{cor}(z_{jk}, w_{lk})} \right)$ is the z-statistic similar to Set 1 but only uses a subset of the data in each drug group. Set 3 is collection of $\tilde{\mathcal{Z}}_{jlk}$, $(l = 1, 2, 3, 4$ and $k = 2, 3, 4)$.

Set 1 shows the overall relationship between proteomics features and nutrition features. Set 2 describes how liver protein abundance marginally changes with respect to different drugs. Set 3 represents the change of relationship between nutrition and proteomics in the three-drug treatment groups, i.e. metformin, rapamycin, or resveratrol, which showed to have an interaction effect between nutrition and drug [46].

Annotate subclusters by interpretable features: The three created sets of interpretable features reflect different aspects of the relationship between proteomics features and experimental factors. We further annotate each subcluster based on these features. For the $\mathcal{J}^{th}$ subcluster, we take the annotation of its interaction effect, as an example: we first transform the related

interpretable features $\tilde{\mathcal{Z}}_{jlk}, (j \in \mathcal{J})$ to $\mathcal{F}_{jlk}$ as follows,

$$\mathcal{F}_{jlk} = \begin{cases} 1, & \tilde{\mathcal{Z}}_{jlk} > \Phi(0.95) \\ 0, & -\Phi(0.95) \leq \tilde{\mathcal{Z}}_{jlk} \leq \Phi(0.95) \\ -1, & \tilde{\mathcal{Z}}_{jlk} < -\Phi(0.95) \end{cases} \tag{3.6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\Phi(0.95) \approx 1.65$. Then we calculate the proportion of $\mathcal{F}_{jlk} = 1$ or $\mathcal{F}_{jlk} = -1$ for proteins within subcluster $\mathcal{J}$, i.e. $P_{Jlk} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} I(\mathcal{F}_{jlk} = 1)$ or $N_{Jlk} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} I(F_{jlk} = -1)$, where $I(\cdot)$ is the indicator function. If $P_{Jlk} > 0.7$, it indicates that at least 70% proteins in subcluster $\mathcal{J}$ show significantly increased correlation with nutrition variable l in drug group k compared with the correlation coefficients in the control group. Then we annotate cluster $\mathcal{J}$ with "increased correlation with variable l in drug group k". Similar annotation procedure works for $N_{Jkl} > 0.7$.

Annotate subclusters by pathway enrichment analysis: On the other hand, we also annotate each subcluster based on the enrichment of pathways in this cluster. This is done by enrichment analysis with the R package clusterProfiler [181] and the top enriched KEGG pathway [68] is also used to describe each subcluster.

3.2.3 Contribution of nutrition intake, drug treatment and their interaction

To better describe the interpretable groups “N+T” and “NxT”, we measure the contribution of groups of variables based on the difference of the residual sum of squares (RSS) between the full model and the reduced model, i.e. setting part of the parameters to zero. For group “NxT”, we fit the following GAM in Equation (3.1) - (3.4). Then we calculate the difference

of RSS between the full model and the reduced models:

$$\Delta\text{RSS}_{int} = \text{RSS}(\text{M2}) - \text{RSS}(\text{M1}), \tag{3.7}$$

$$\Delta\text{RSS}_{N} = \text{RSS}(\text{M3}) - \text{RSS}(\text{M2}), \tag{3.8}$$

$$\Delta\text{RSS}_{T} = \text{RSS}(\text{M4}) - \text{RSS}(\text{M2}), \tag{3.9}$$

where $\text{RSS}(\text{M})$ is the RSS of model M. The contribution of the interaction term V_{int} is defined as the normalised ΔRSS, i.e.

$$V_{int} = \frac{\Delta\text{RSS}_{int}}{\Delta\text{RSS}_{int} + \Delta\text{RSS}_{N} + \Delta\text{RSS}_{T}}.$$

A Similar calculation can be applied for the contribution of nutrition V_N and drug intake V_T. We also apply it to the "N+T" group and get the contribution of nutrition features and drug intake for each proteomics feature.

3.2.4 Distance for consensus clustering

We use four different types of distance for our consensus clustering. The first two are the widely used Pearson and Spearman correlation distance [77], which we denote as D_{pcc} and D_{spc}, respectively. We also use two types of weighted correlation distance to account for their similarity in terms of contribution of nutrition intake, drug treatment, and their interactions. We use D_{pcc} as an illustration and the same calculation can be applied to D_{spc}. First we calculate the Aitchison distance [54] of (V_N, V_T, V_{int}) as done in Section 3.2.3 between pairs of proteomics features. We denote this distance matrix as D_V. Then we normalise [77] D_V and D_{pcc} by $\tilde{D}_V = \frac{D_V}{max(D_V)}$ and $\tilde{D}_{pcc} = \frac{D_{pcc}}{max(D_{pcc})}$. The weighted Pearson distance is calculated by $D_{wpcc} = (1 - \tilde{D}_V)\tilde{D}_{pcc} + \tilde{D}_V^2$.

D_{wpcc} dynamically alternates the focus between D_V and D_{pcc}. As is illustrated in Figure 3.3, two pairs of points with different contributions of nutrition (N), drug treatment (T), and their interaction (NxT) are shown. We observe that when the (j_1, j_2)th element in $\tilde{D}_V$ is close to 0 (two green points), then the roles of the experimental factors are similar for protein j_1 and protein j_2; additionally, $D_{wpcc}(j_1, j_2)$ is close to $\tilde{D}_{pcc}(j_1, j_2)$, as $D_{wpcc}(j_1, j_2)$ focuses more

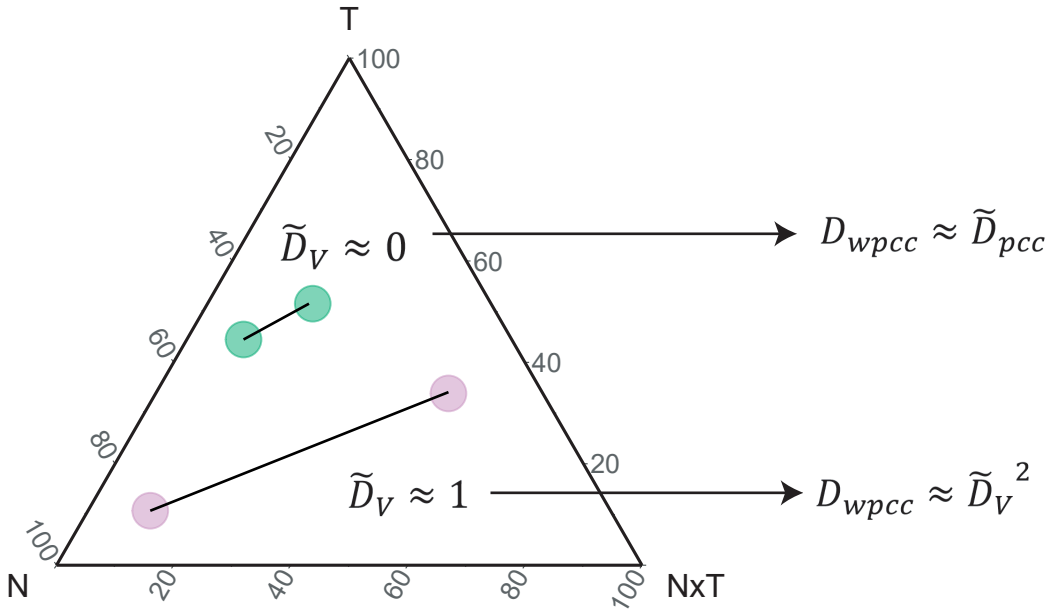

FIGURE 3.3. **Illustration of weighted Pearson distance** The axes in the triangle show the contribution of nutrition (N), drug treatment (T) and their interactions (NxT).

on the correlation structure. When $\tilde{D}_V$ is close to 1 (two pink points), which means that the contribution of some experimental factors shows the discrepancy between two proteins, then D_{wpcc} tends to focus more on $\tilde{D}_V$.

3.2.5 Creating a network among proteins, subclusters and phenotypes

We first calculate the Spearman correlations between the j^{th} proteomics feature and the m^{th} metabolic phenotype, denoted as ρ_{jm}. The p-value for testing against $H_0 : \rho_{jm} = 0$ is calculated. If the p-value is smaller than 0.01, the corresponding Spearman correlation is set to zero. Then we use a gene set enrichment analysis (GSEA) like the multiset test [116] method to determine the significance of the correlation between a subcluster and selected metabolic phenotype. If the p-value is smaller than 0.01, we put an edge (as in Figure 3.17) to emphasise the link between the corresponding subcluster and phenotype. Proteomics features

and subclusters are linked by proteins showing high correlation (that is correlation >0.7) with the first principal component of proteins in the subclusters. The resulting network is drawn using the R-package ggnetwork [159].

3.2.6 NGF visualisation

NGF visualisation of each proteomics feature varies across the interpretable groups. For the "N" group, we use all samples to fit a GAM model based on M4 in Equation (3.4); the fitted value is used to visualise the response surface. For the "N+T" group, the NGFs are based on Equation (3.2) by subtracting the treatment effect in the model. For the "NxT" group, the fitted value of the response surface of the NGFs is based on Equation (3.1) within each drug treatment. All of the significance of the visualisation is tested based on the LC-Test described in Chapter 2.

3.3 Results

3.3.1 Categorising omics data into interpretable groups derived from experiments

The proteomics features are first grouped by eNODAL based on their response to nutrition and drug intake (see Figure 3.4). A total of $2,951$ proteins out of $4,987$ proteins show significant responses to nutrition and drug intake. Among these proteins with significant responses, the "N", "N+T" and "NxT" groups are the majority groups with $1,350$, 830, and 717 proteins, respectively, whereas the "T" group only has 54 proteins. The unbalanced number in each interpretable group implies nutrition plays an important role in shaping proteomics profiles, where a large number of proteins only significantly responded to the nutrition intake (group "N") while only a small number of proteins are affected solely by drug treatment (group "T"). Rather, drugs affect protein abundance either additively or through a more complex way, that is by an interaction with diet (group "N+T" and "NxT"). Pathway analysis (see Figure 3.5) shows that RNA splicing pathways are enriched in the group "N" (rank 1, $p < 0.01$) and

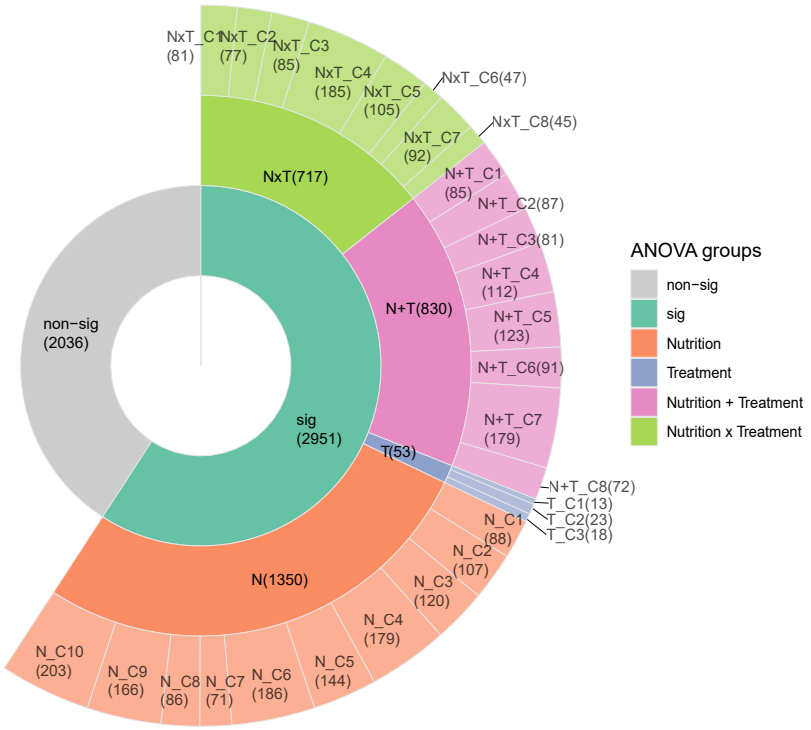

FIGURE 3.4. **Two-stage clustering result of proteomics features**. The $4,987$ proteins are categorised into four interpretable groups based on an ANOVA-like test. Then it further clustered into 29 subclusters within each group. The numbers in each subcluster are shown within the round brackets.

"N+T" (rank 2, p<0.01), a finding consistent with our previous results [87]. For group "NxT", the top enriched pathways are thermogenesis (p<0.01) and carbon metabolism (p<0.01). Several studies showed that thermogenesis is closely related to diet [171, 172] and drug treatment [2, 70, 158]. Further, there is evidence suggesting that the interaction between drug and diet impacts thermogenesis [133]. This implies that eNODAL can group proteomics features based on their response to experimental factors.

3.3.2 Dividing interpretable groups into subclusters reveals different patterns of the proteomics features

The five interpretable groups are further divided into subclusters by eNODAL. For the "N" group, we obtain nine subclusters, and Figure 3.6 shows that there are some clear differences in the abundance of proteins within the nine subclusters. Annotation of these subclusters

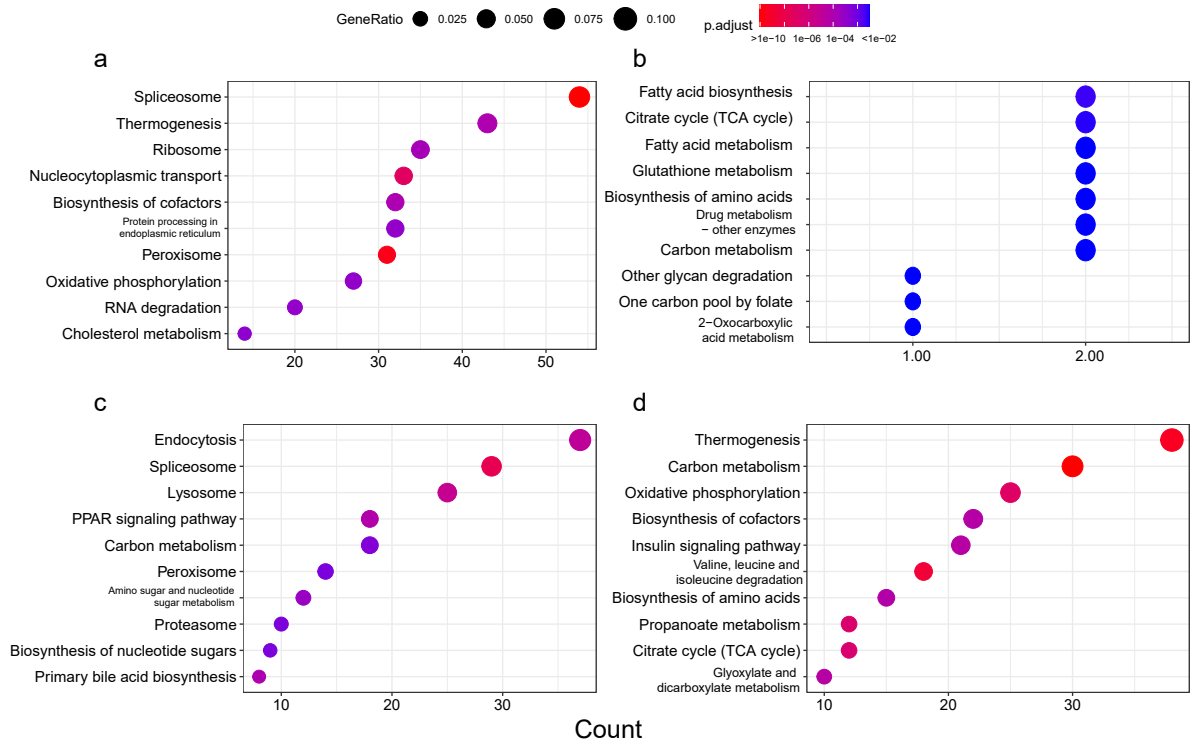

FIGURE 3.5. **KEGG pathway enrichment analysis of groups by ANOVA-like tests**. a)-d) KEGG pathway enrichment of the "N", "T", "N+T", "NxT" group respectively: y-axis represents the top 10 pathways with smallest adjusted pvalue, x-axis represents the number of proteins in the corresponding pathway.

suggests that correlations with nutrition intake are different among these clusters (see Figure 3.7). Taking subcluster 5 in the "N" group ("N_C5", 144 proteins include) as an example, it turns out that the majority of proteins in this subcluster have negative correlations with raw food intake in grams but positive correlations with carbohydrate and fat intake in kJ. Pathway analysis indicates that the Peroxisome pathway is enriched in this subcluster, and the Peroxisome pathway is known to be related to lipid metabolism [84] also consistent with our previous results [87]. To further visualise the effects of nutrient intake on within-subcluster protein abundance, we apply the surfaces-based approach from the NGF to the first principle component (PC1) of abundance within each cluster (see Figure 3.7). The subcluster "N_C5" for example contains proteins with a higher abundance of elevated carbohydrate or protein intake, while the opposing pattern is seen in the subcluster "N_C6". Similar results can be

found within the much smaller "T" group, which is further clustered into three subclusters with different responses to drug treatment (see Figure 3.8).

The "N+T" and "NxT" groups both contain groups of proteins whose abundance are affected in a more complex manner. Both the "N+T" and the "NxT" group contain eight subclusters (see Figure 3.9 and 3.12). In the "N+T" group, the effects of nutrition intake and drug treatment are additive. We use NGFs and boxplots to present the associations between protein abundance and nutrition intake as well as drug treatment as shown in Figure 3.13. For the "NxT" group, the interaction effect between nutrition and drug contributes to the abundance of proteins in these subclusters. In other words, the association between diet and abundance is modified by the drug treatment. Therefore, the association between nutrition intake and within cluster protein abundance (based on PC1 for the cluster) are visualised by the NGF surfaces applied to each drug group separately (see Figure 3.10).

The diverse NGFs in the "N+T" group and the "NxT" group further show the complex interplay between diet and drug in these clusters and the effect of the drug needs to be evaluated with respect to the nutrition context. For example, within subcluster 4 in the "NxT" group ("NxT_C4", see Figure 3.10) we see that in the absence of metformin, increasing the ratio of dietary protein to carbohydrate elevates the abundance of proteins in this subcluster. However, in the control and other drug treatment groups, the abundance of "NxT_C4" proteins is affected less by the ratio of macronutrients. Annotation of "NxT_C4" suggests that the majority of the proteins in this subcluster show an increased correlation with protein intake in the metformin and resveratrol group. We also notice that the AMPK, insulin, and glucagon signalling pathways are enriched in this subcluster (see Figure 3.11). Previous studies showed that intake of metformin is closely related to the activation of AMPK [167] and several studies also reported the interaction between diet and metformin [45, 106].

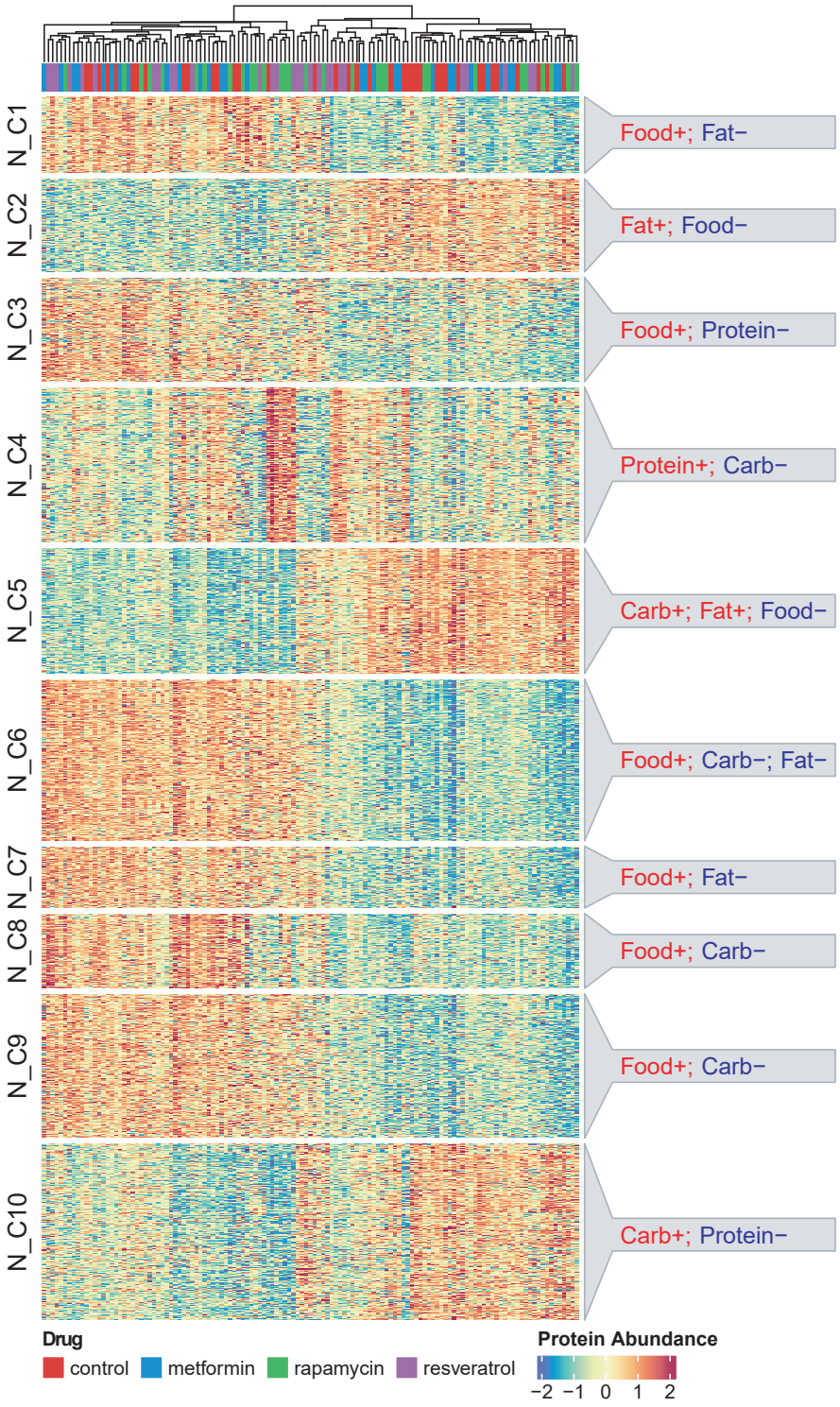

FIGURE 3.6. **Subclusters in the "N" group and their annotations**. Left panel shows the heatmap of protein abundance of proteins in the "N" group, split by subclusters. The right panel shows the annotation of each subcluster.

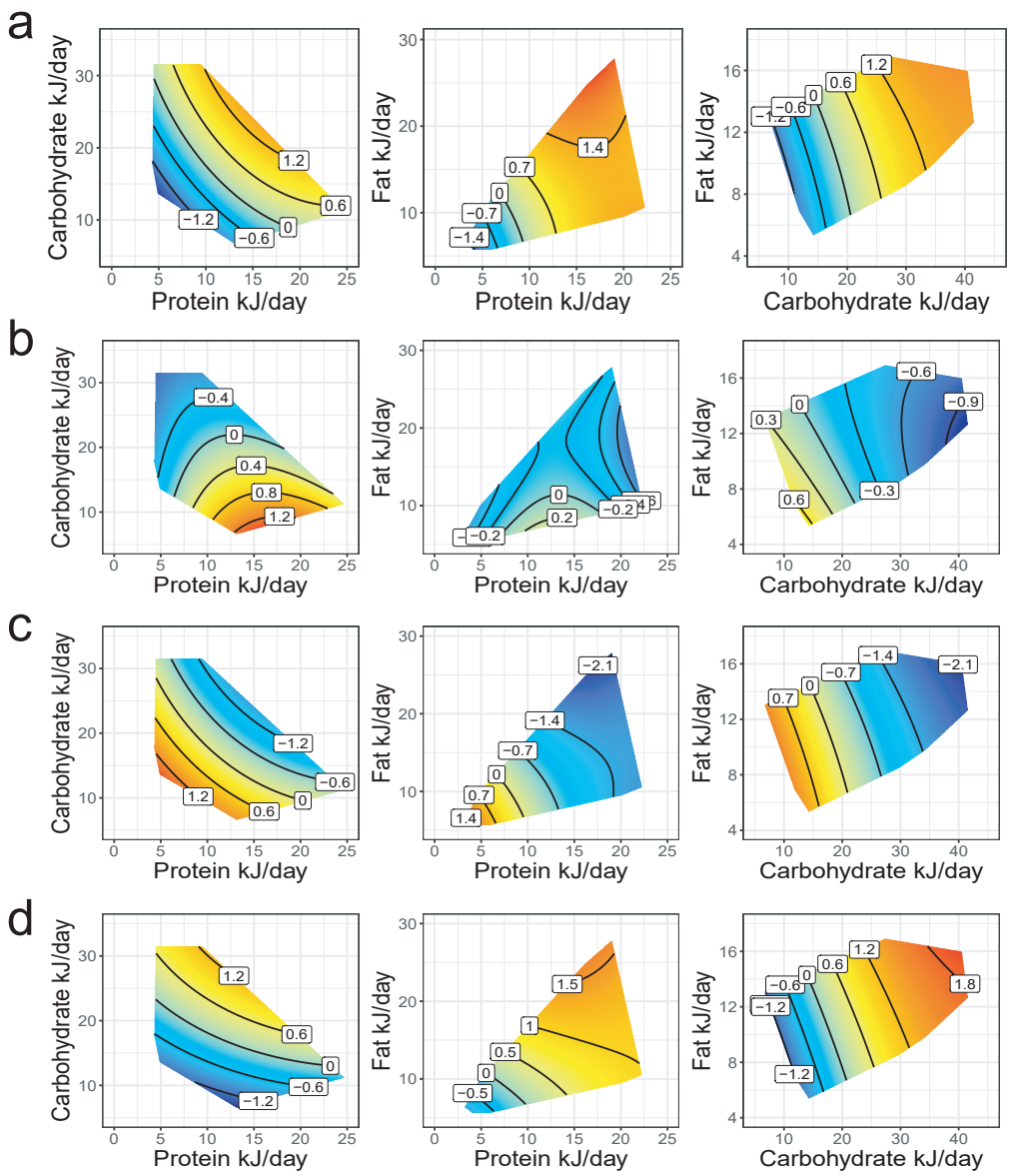

FIGURE 3.7. **NGFs of the first PC of four subclusters in the "N" group**. a) NGF of PC1 of subcluster "N_C2", b) NGF of PC1 of subcluster "N_C4", c) NGF of PC1 of subcluster "N_C6", d) NGF of PC1 of subcluster "N_C10".

3.3.3 eNODAL identifies subclusters of proteins with similar response of nutrition intake and drug treatment

We evaluate the clustering result between eNODAL and the result from WGCNA [82], a widely used method for categorising high-dimensional features. In this mouse nutriomics

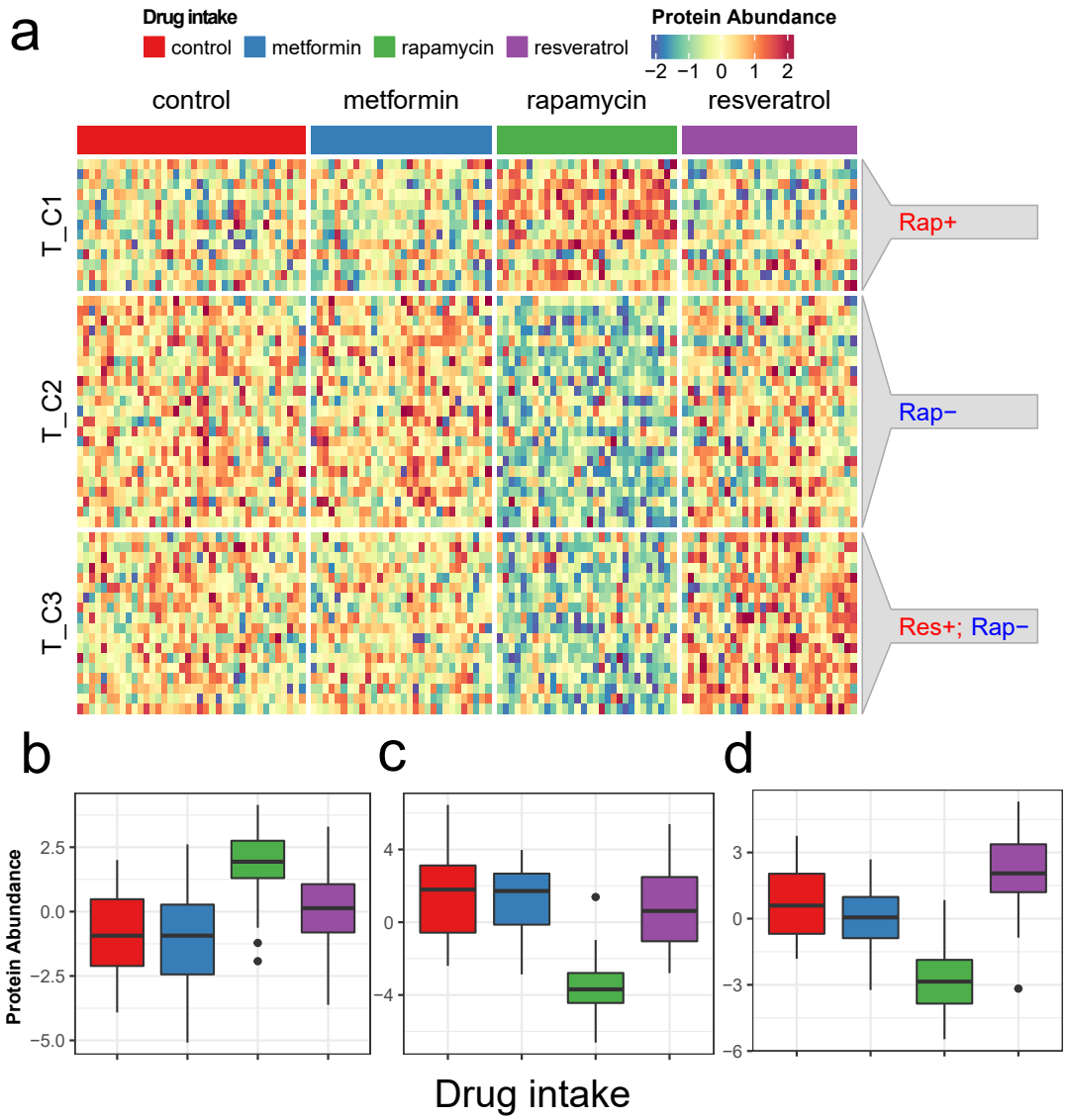

FIGURE 3.8. **Subclusters in the "T" group and their annotations**. a) Protein abundance of three subclusters in "T" group, b) boxplot of PC1 of subcluster "T_C1", c) boxplot of PC1 of subcluster "T_C2", d) boxplot of PC1 of subcluster "T_C3".

dataset, WGCNA obtains 5 clusters of proteomics features, containing 1671, 970, 140, 97, and 73 proteins respectively. The adjusted rand index (ARI), a cluster comparison statistic, is quite low (≈ 0.08) between clustering results from WGCNA and eNODAL. We use the subcluster annotation procedure (see Section 3.2.2.3) to annotate WGCNA clusters. It turns out most of the subclusters are driven by a marginal correlation between nutrition intake and proteomics abundance. As is shown in the previous result, macronutrients have a much larger

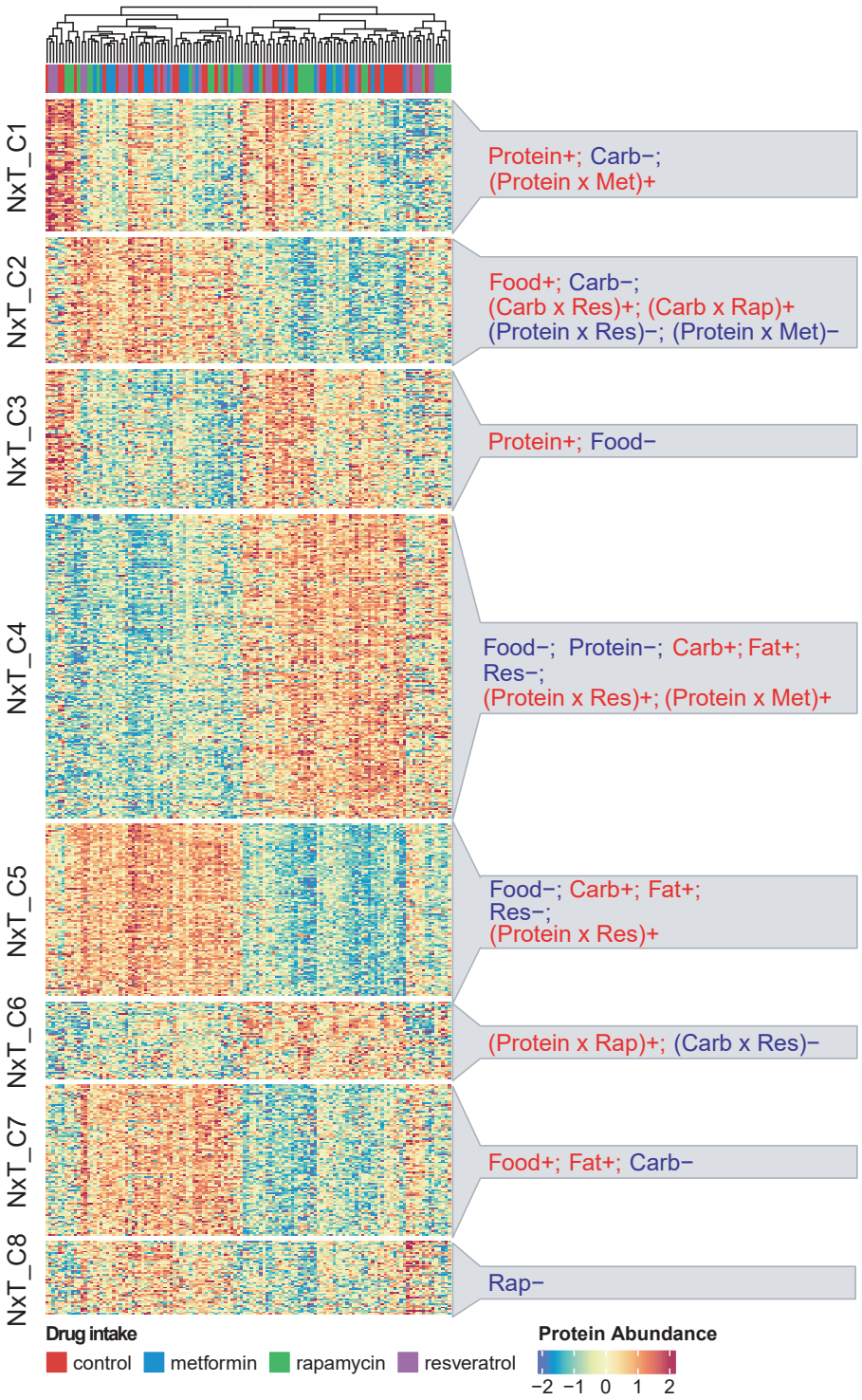

FIGURE 3.9. **Subclusters in "NxT" group and its annotation**. The left panel shows the heatmap of protein abundance of proteins in the "NxT" group, split by subclusters. The right panel shows the annotation of each subcluster.

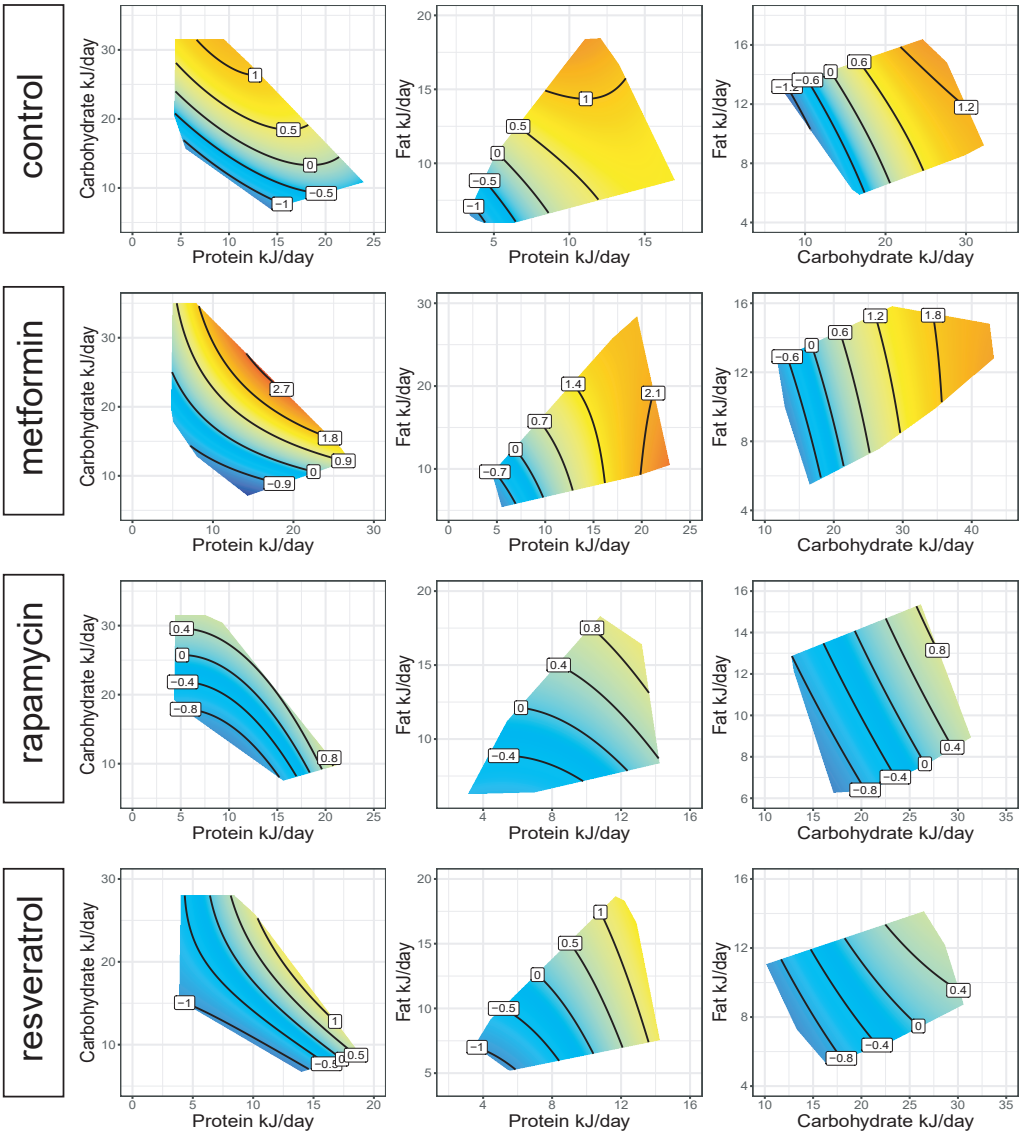

FIGURE 3.10. **NGFs of PC1 in subcluster "NxT_C4"**. NGFs of PC1 in subcluster "NxT_C4", four rows corresponding to NGF within the control, metformin, rapamycin and resveratrol group respectively.

effect on proteins than treatment [87]. eNODAL, however, further reveals relationships among protein abundance, nutrition intake, and drug treatment in detail compared with WGCNA. Taking the WGCNA cluster 2 ("W_C2") as an example, the majority of its proteins show a negative correlation with raw food intake and a positive correlation with fat intake; while, in the eNODAL result, the majority of these proteins are distributed in subclusters "N_C2",

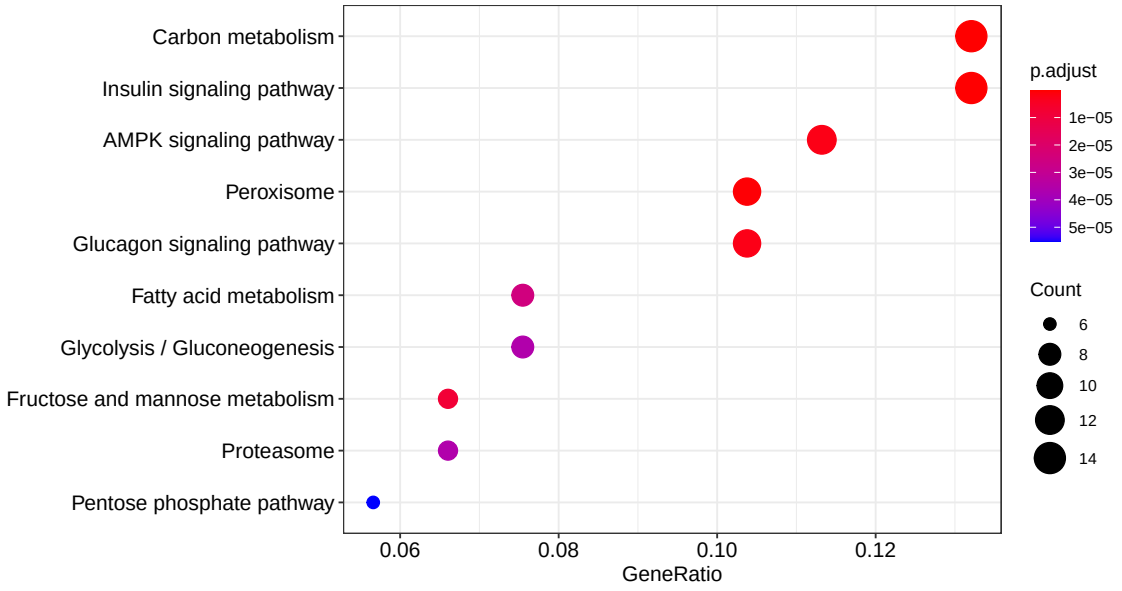

FIGURE 3.11. **KEGG pathway enrichment analysis of subcluster "NxT_C4"**. Biological signalling pathways such as the AMPK, insulin and glucagon signalling pathway are enriched in this subcluster.

"N_C5", "N_C10", "N+T_C4, "N+T_C5", "N+T_C8", "NxT_C4" and "NxT_C6" (see Figure 3.14). We noticed that proteins in subcluster "N+T_C5" not only share a similar marginal correlation with nutrition intake to "W_C2", but they also exhibit lower abundance in the resveratrol group. Similarly, subcluster "NxT_C4" further reveals the interaction between nutrient intake and drug treatment.

We notice that the insulin signalling pathway is enriched in "W_C2" ($p < 0.01$, rank 3, Figure 3.15). At the same time, this pathway is also enriched in eNODAL subclusters "N+T_C8"($p < 0.01$, rank 1, Figure 3.16) and "NxT_C4"(p < 0.01, rank 2, Figure 3.11). Except for the relationship with nutrient intake, proteins in subcluster "N+T_C8" show lower abundance in the rapamycin group and resveratrol group in "NxT_C4". Several studies pointed out that this pathway is affected by high carbohydrate or fat diet intake [94, 97] as well as rapamycin [91] and resveratrol [6, 132]. This result indicates that subclusters by eNODAL can integrate the information from nutrition intake, drug treatment, and protein abundance and provide a biologically meaningful interpretation of these subclusters.

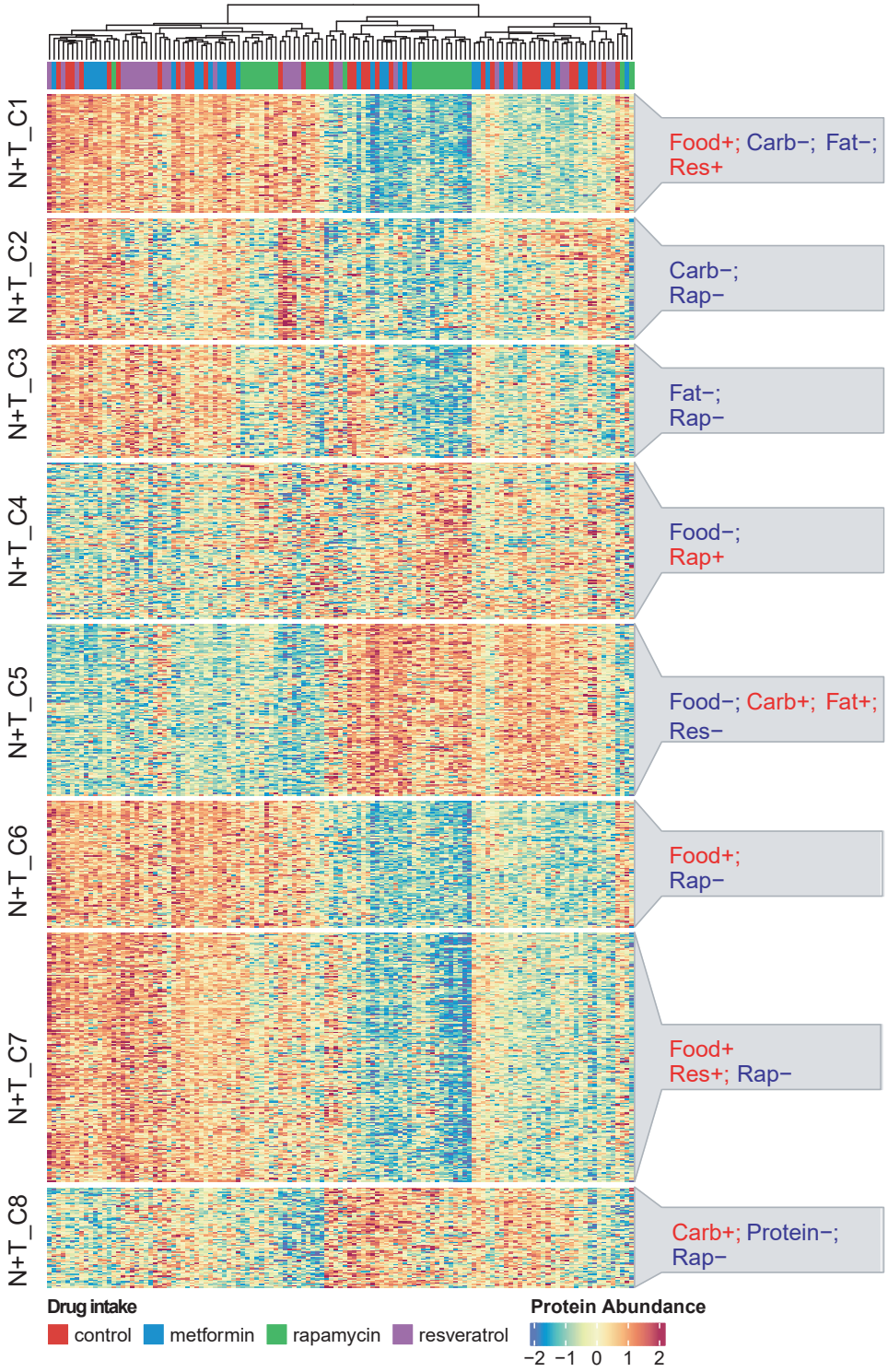

FIGURE 3.12. **Subclusters in "N+T" group and its annotation**. The left panel shows the heatmap of protein abundance of proteins in the "N+T" group, split by subclusters. The right panel shows the annotation of each subcluster.

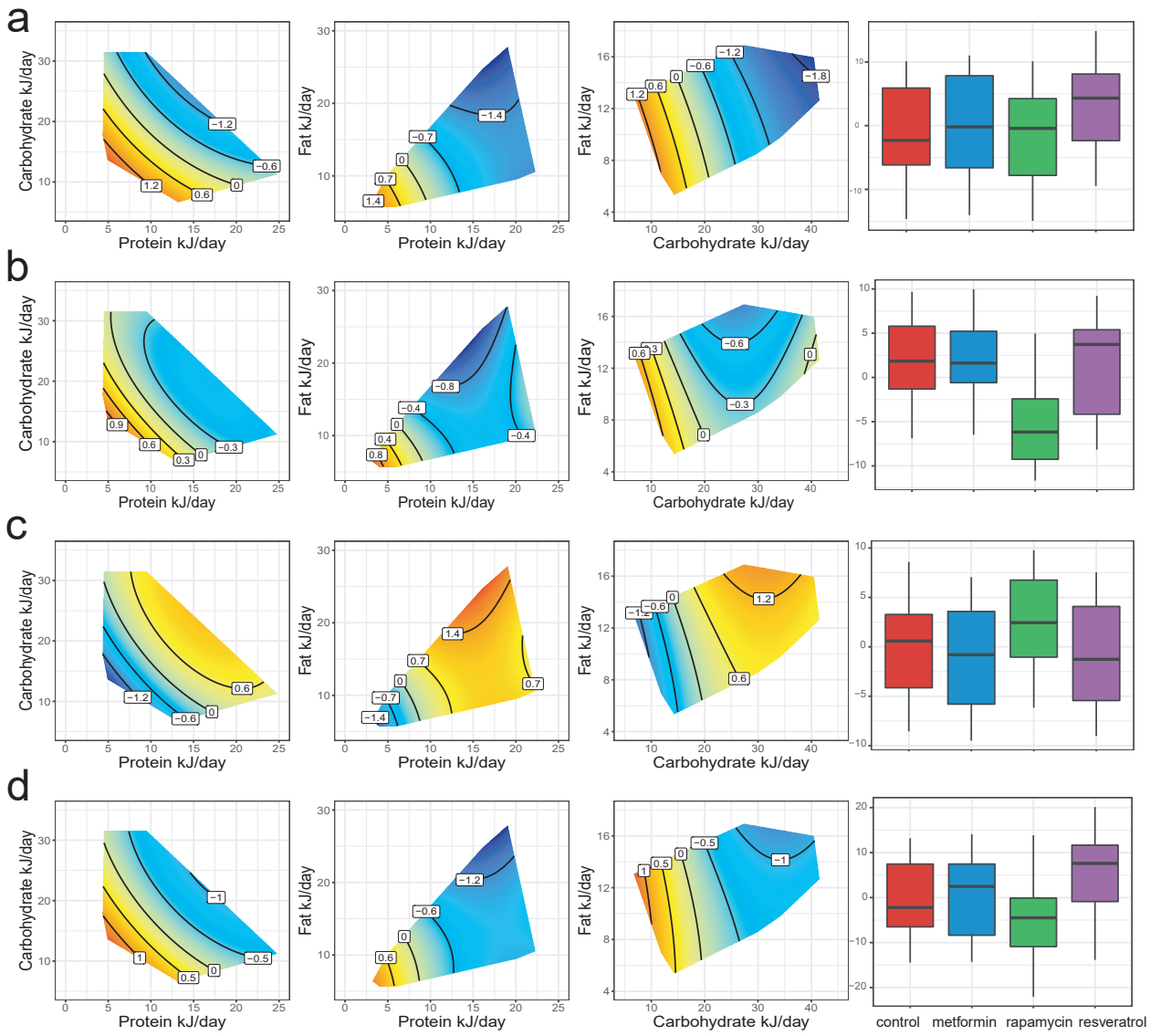

FIGURE 3.13. **NGFs and boxplots of the first PC of four subclusters in "N+T" group**. a) NGF and boxplot of PC1 of subcluster "N+T_C1", b) NGF and boxplot of PC1 of subcluster "N+T_C3", c) NGF and boxplot of PC1 of subcluster "N+T_C4", d) NGF and boxplot of PC1 of subcluster "N+T_C7".

3.3.4 eNODAL reveals interplay between dietary components, drug intake, liver proteins and metabolic phenotypes

In this mouse nutriomics study, eNODAL provides a natural way to combine proteins, subclusters, and diet-related metabolic phenotypes revealing how diet-drug-proteins contribute

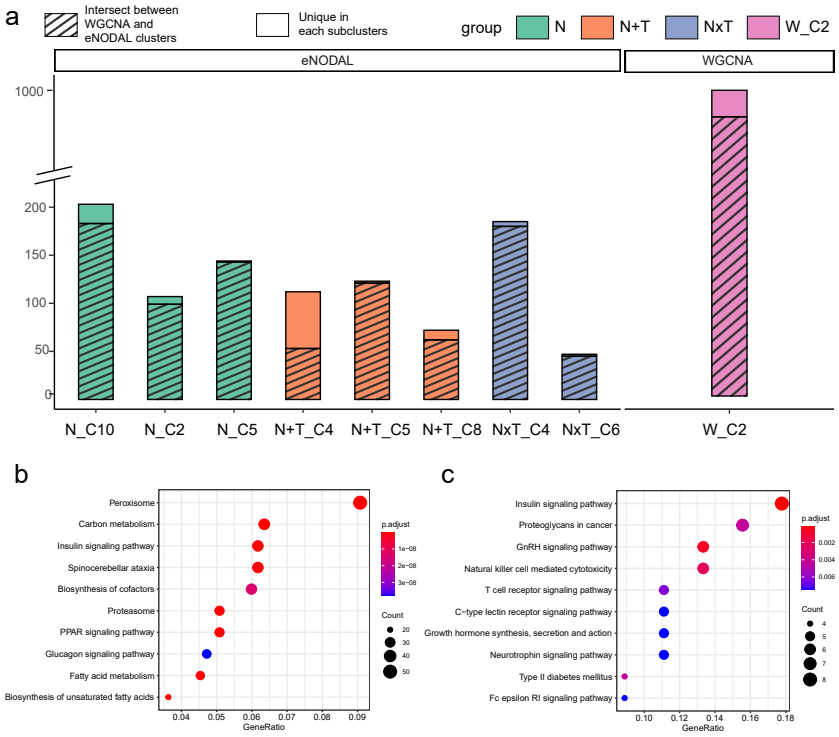

FIGURE 3.14. **Relationship between WGCNA cluster "W_C2" and related eNODAL subclusters**. Left panel shows bar plot of eight majority subclusters related to "W_C2", shaded bars represent that the proteins in each subcluster intersect with "W_C2". The right panel shows a bar plot of proteins in "W_C2"; the shaded bars represent that the proteins in "W_C2" intersect with all eight subclusters in eNODAL.

to the health of mice. We create a network to link them and reveal their associations (see Section 3.2.5) as shown in Figure 3.17: we note that most of these metabolic phenotypes include body weight, fast insulin, and the amount of retroperitoneal fat pad correlates with subclusters that respond to the amount of raw food intake, such as in the "N_C2", "N+T_C5" and "NxT_C4" subclusters, consistent with our previous finding [87]. On the other hand, we notice the incremental area under the curve for insulin (iAUC) associated with subclusters affected by the intake of rapamycin ("T_C1" and "N+T_C4"). Similar results have been observed in other studies[170, 187].

Figure 3.17 also provides a way to analyse the interplay between proteins and metabolic phenotypes: we notice that protein Pex11, a hub protein in subcluster "NxT_C4", is positively correlated with many of the metabolic phenotypes. Previous studies have also shown that this

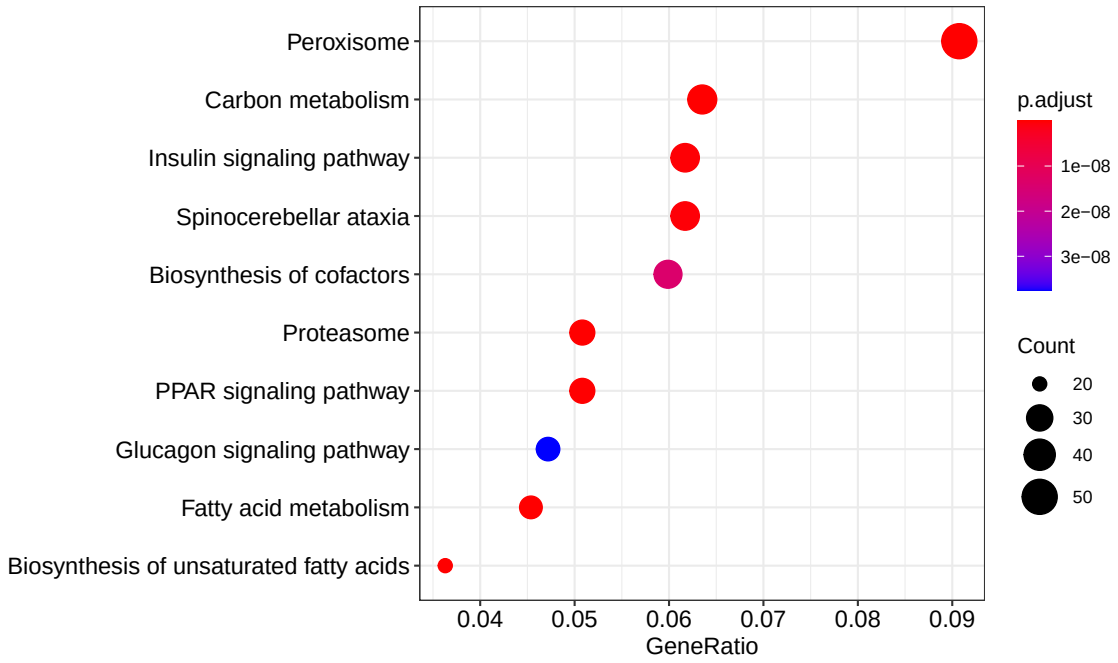

FIGURE 3.15. **KEGG pathway enrichment analysis of subcluster "W_C2"** Biological signaling pathways such as insulin signalling pathway are enriched in this subcluster

protein relates to phenotypes such as body weight and insulin level [22]. NGFs of this protein (see Figure 3.18) also indicate that its abundance is elevated in the metformin group with an increasing ratio of dietary protein to carbohydrate. Similar relationships were also observed in other studies [93, 139].

3.4 Discussion

This chapter provides a method to integrate the nutriomics experimental information with high-dimensional proteomics features by a two-stage clustering procedure called eNODAL. eNODAL provides a framework to analyse the effect of nutrition and drug in a two-stage manner: the first stage categorises the features into interpretable groups based on whether the proteomics features showed significant responses to nutrition and drug intake. These interpretable groups determine whether the change in experimental conditions affects protein abundance. The second stage further clusters these features within each interpretable group

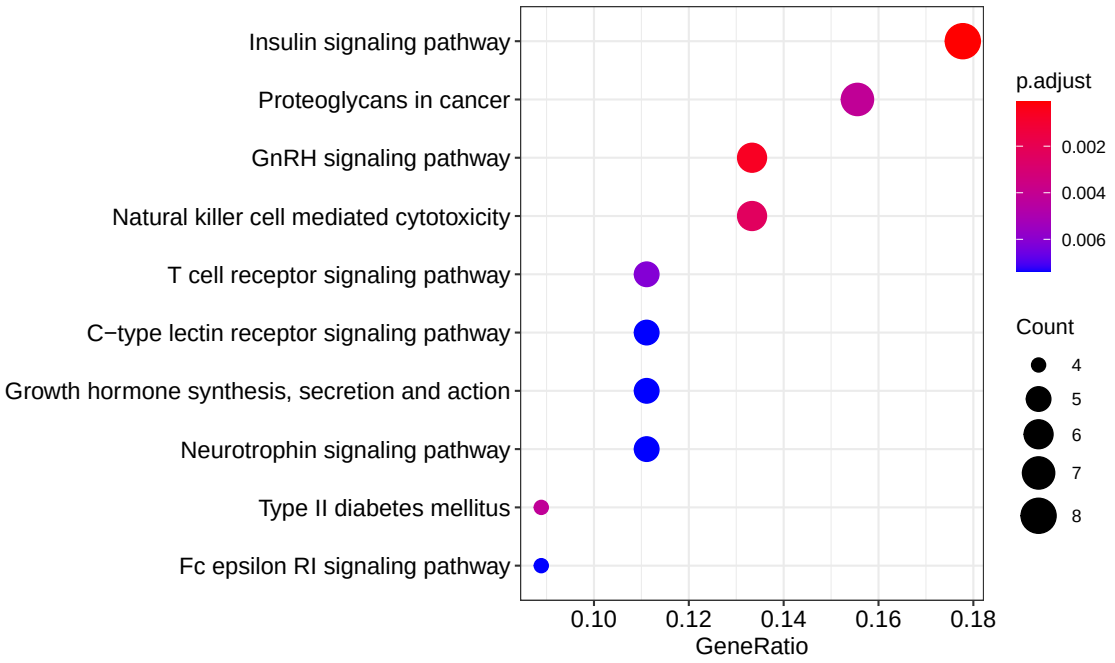

FIGURE 3.16. **KEGG pathway enrichment analysis of subcluster "N+T_C8"** Biological signaling pathways such as insulin signalling pathway are enriched in this subcluster

into subclusters with similar abundance profiles. The interpretation of these subclusters is then determined by the relationship with experimental conditions and enrichment of biological pathways. When applying eNODAL to a mouse nutrition study, we obtained 29 subclusters of proteomics features representing different biological pathways. These subclusters could potentially provide insights into how nutrition and drug shape the proteomics profile and further affect our health.

The first component of eNODAL methodology aims to incorporate experimental design information (e.g., factorial design structure) in the overall classification of omics features. This component is not limited to the ANOVA-like test approach used in this Chapter. In our development, we use GAMs to model the association between experimental factors and the abundance of each proteomics feature motivated by the flexibility of GAM. GAM is a nonparametric model able to capture the nonlinear relationships between covariates and response variables. Due to the complex interaction among nutrients, modeling such nonlinearity is appropriate in nutrition context, (see Chapter 2 and Simpson et al. [141]). While

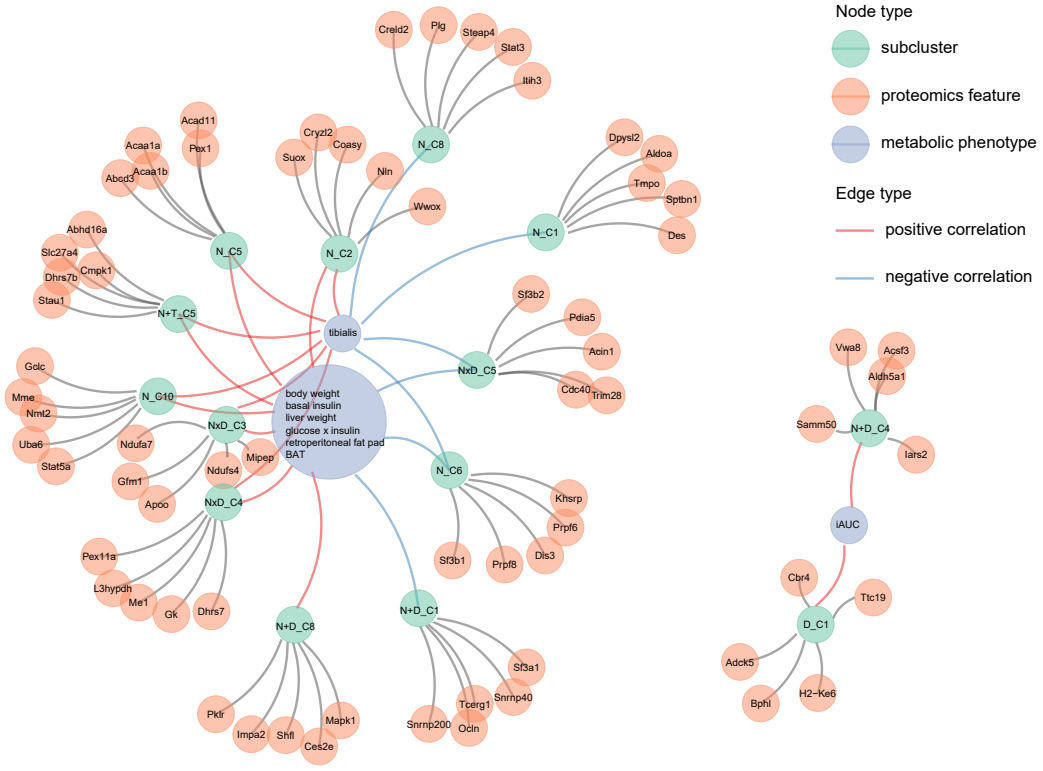

FIGURE 3.17. **Network among proteomics features, subclusters and metabolic phenotypes**

in the setting where we anticipated the majority of features have linear relationships between covariates and response variables in the data, the use of linear models in the first stage is able to reduce computing complexity.

The second component of the eNODEL workflow is the using data-driven clustering approach to generate refined subclusters. We pointed out that these clustering methods are also able to be user-defined to accommodate data structure or prior biological knowledge. Our current strategy is to use an ensemble of clustering algorithms with varying distance measures, with the goal of ensuring the number of estimated clusters is stable and robust to differences between distance measures. However, if the number of subclusters is known through external biological knowledge, such information can be incorporated with alternative unsupervised clustering such as kmeans or hierarchical clustering approach.

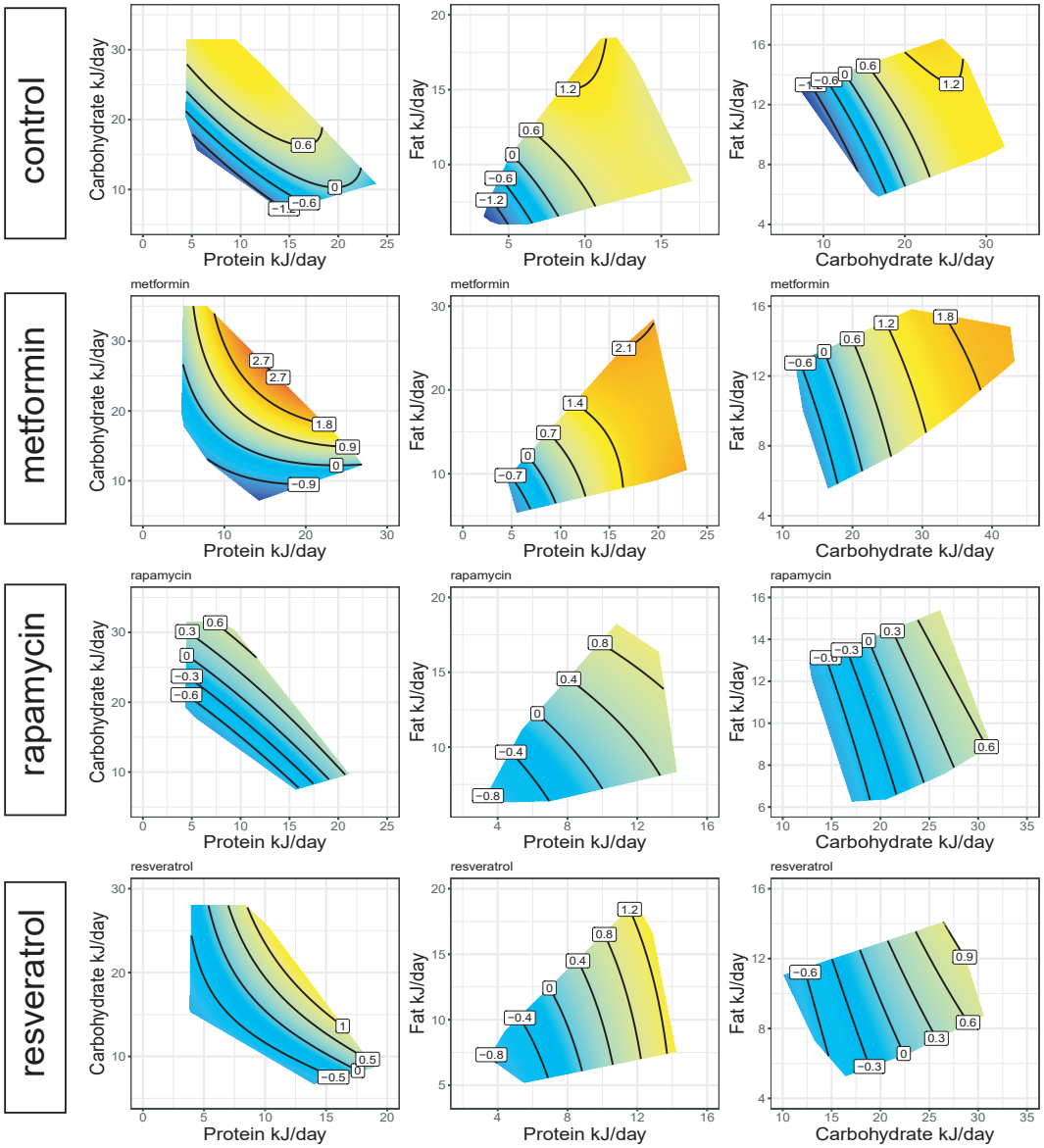

FIGURE 3.18. **NGFs of protein Pex11**

The proposed eNODAL method is based on an analysis of the effect of nutrition and drugs on the proteome. However, our approach is not restricted to the examination of the three data modalities described in this Chapter. A natural extension of eNODAL is to handle experiments generating multi-omics data. Our method can be easily adapted by extending both stages of eNODAL to handle multi-omics data. In the first component, we classify each omics feature according to whether it responds significantly to each experimental factor or

interaction effect. Various multi-omics clustering algorithms, as described in a recent review [127], can be used to modify the subclustering in the second component. For example, we construct a similarity network for each omics feature inside each interpretable group before integrating and clustering multi-omics features using the network fusion [165].

In summary, we present eNODAL, a two-stage clustering method to cluster proteomics features into interpretable subclusters. This method first classifies the proteomics features based on an ANOVA-like test to determine whether experimental factors as well as their interactions have a significant effect on proteomics features in several interpretable groups. Then an unsupervised clustering method was used to further cluster the proteomics features within each interpretable group into subclusters with similar profiles. The application of eNODAL on the mouse nutrition study demonstrated its ability to identify biologically meaningful subclusters, such as subcluster "NxT_C4" containing proteins enriched in AMPK signaling pathways, which is not found (i.e. significantly enriched) by popular method WGCNA. eNODAL is implemented as an R package and can be downloaded freely from https://github.com/SydneyBioX/eNODAL.

CHAPTER 4

Identification of diet subcohorts in observational nutriomics data

---

The gut microbiota of each individual is unique and maintains a symbiotic relationship for the benefit of humans, as it has been observed that these microbes are key contributors to various vital metabolic requirements [154]. Diet is one of the key factors that influence the composition and balance of the gut microbiome [142]. A deep understanding of the interplay between diet, the microbiome, and health state could enable the design of personalised intervention strategies and improve the health and wellbeing of individuals [107, 186].

In Chapter 3, we provided a method to cluster the omics features for experimental nutriomics data analysis. However, in nutriomics observational data, the population could be heterogenous, as different people with different diet intakes may experience very different relationships between omics features and their respective health outcomes. In the microbiome context, a similar concept is "enterotype", which in essence captures that long-term diet is closely linked to the composition of the gut microbiome [180]. One strategy to address this heterogeneity is to divide the population into subcohorts, i.e. into several groups, each having people who have similar diet intake. The analysis of the microbial profile occurs then within each subcohort. However, the classification of subjects based solely on diet is unlikely to reflect the microbiome-host health relationship or the taxonomic microbiome makeup [151].

In this chapter, we present a novel approach, the Nutrition-Ecotype Mixture of Experts (NEMoE) model, for establishing associations between gut microbiota and health state that accounts for diet-specific cohort variability using a regularised mixture of experts model

*Most of this work can also be seen in: **Xiangnan Xu**, Michal Lubomski, Andrew J. Holmes, Carolyn M. Sue, Ryan L. Davis, Samuel Muller, and Jean YH Yang. NEMoE: A nutrition aware regularized mixture of experts model addressing diet-cohort heterogeneity of gut microbiota in Parkinson's disease. *Under Review* (2021).

framework with an integrated parameter sharing strategy to ensure data-driven diet-cohort identification consistency across taxonomic levels. NEMoE was demonstrated through a series of simulation studies in Section 4.3.1, in which NEMoE showed robustness with regard to parameter selection and varying degrees of data heterogeneity. Further application to real-world microbiome data from a Parkinson's disease cohort revealed that NEMoE is capable of not only improving predictive performance for Parkinson's Disease (Section 4.3.2) but also of identifying diet-specific microbiome markers of disease (Section 4.3.4).

4.1 Background

The human body is home to complex microbial communities, collectively known as the microbiome, which is mostly made up of prokaryotes (bacteria) and archaea [90]. Considerable evidence has emerged indicating that the microbiome is an important contributor to an individual's health [180]. This has been illustrated by links between gut microbiome and numerous diseases, including irritable bowel syndrome [25], Crohn's disease [118], type 2 diabetes [124], cardiovascular disease [78] and Parkinson's disease (PD) [98]. The gut microbiome is known to change throughout our lives as a result of various environmental influences. Diet, being one of these factors, has the greatest known long-term interaction with the gut microbiome [189]. Thus, a deep understanding of the relationship between diet and the gut microbiome and the consequential impact on disease processes, holds promise for developing personalised dietary intervention strategies to modulate and maintain a healthy microbiome population [107, 186].

Diet has a direct impact on the microbial community in the gut, which governs the activity of the intestinal ecosystem and can have considerable implications for an individual's health [35, 129]. This is conceptualised in Figure 4.1 where, for illustration purposes, the macronutrient intake is separated into three perfectly distinct subcohorts with different associations between microbiome composition and PD. In practice, several studies have demonstrated that variations in nutrient intake, such as different ratios of protein, carbohydrate [59] or dietary fiber [32] intake, can influence the host-microbiome association. These discoveries are generally based

on elaborate experimental design using model organisms [59] or dietary interventions [33, 57, 191]. Recent observational studies suggest that long-term diets could be associated with the microbiome [10], and this can further affect overall health. In a similar context, our recent study of the gut microbiome in PD showed that when partitioning individuals based on carbohydrate intake, the predictive performance of the microbiota profile to indicate PD was increased [99, 100]. Together, these studies suggest that dietary differences can impact relationships between microbiome composition and host health/disease status.

To uncover complex heterogeneous relationship structures between diet, microbiome, and host health, it is important to identify homogeneous subcohorts or latent structures in data that can be explained by a set of features. This is similar to the concept of “ecotype”, which is commonly used to refer to a variant that has observable phenotypical differences in a local environment [95]. Hence, using a data-driven approach, it is able to divide a population into multiple subcohorts with distinct microbiological signatures for health that can be best described by nutrient combinations, resulting in what we term "nutritional-ecotypes". These subcohorts can be thought of as diet-based latent classes where they capture the interaction between the constraints imposed by nutrient intake of individuals on the community dynamics of their microbiomes [136, 153].

Methods to discover such diet-based latent classes could be hypothesis-driven based on prior knowledge [63, 137] or guided by an unsupervised statistical learning method, such as clustering [61], followed by latent class analysis [60]. Although these methods identify nutrient-classes with an altered overall nutritional profile, one limitation is that the defined cohorts may not reflect the heterogeneous microbiome-host health relationship: the drivers of “diet x microbiome” outcomes, “diet x host” outcomes, and “host x microbiome outcomes” are overlapping, but not perfectly congruent. Consequently, classification models built within a subcohort defined just by diet (or microbiome) will not necessarily improve the prediction of the health/disease state [151].

Similar concepts of identifying cohort heterogeneity to improve prediction performance have been developed in other omics settings and for other diseases [119, 150]. However, simple adaptations of methodologies developed for other omics platforms remain challenging as

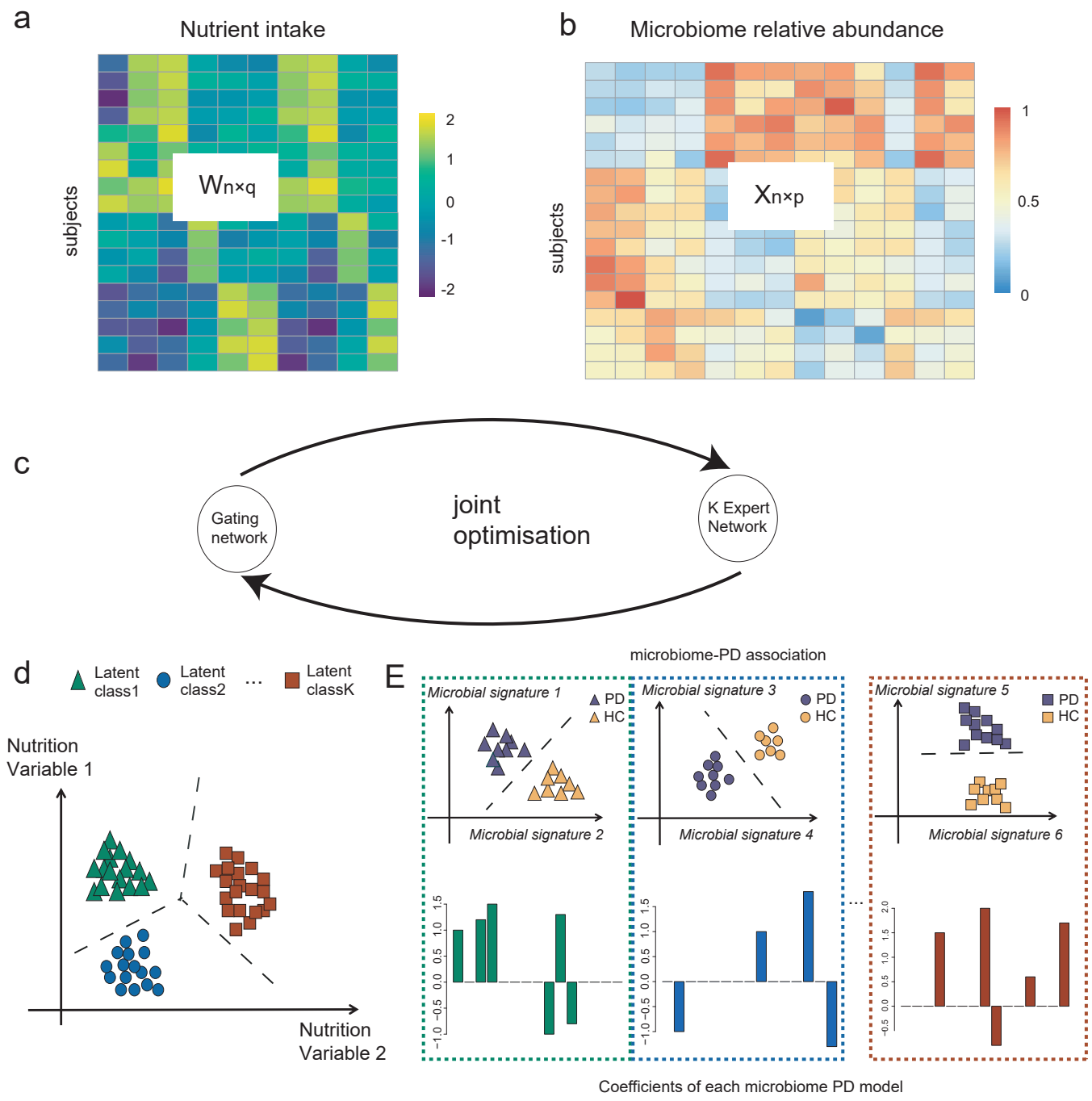

FIGURE 4.1. **Illustration of NEMoE.** a),b) The input matrices of NEMoE: n samples with q nutrient features and p microbial features. c) A conceptual workflow of NEMoE, where the joint optimisation is achieved by the EM algorithm to maximise the regularised likelihood function. d) A toy example showing a nutritional-ecotype in the microbiome PD study. The nutrient intake is clustered into K latent classes. e) In each latent class, the microbial signatures of PD are different, which is reflected by the coefficients in the experts network.

these do not account for the hierarchical taxonomic structure observed in the study of the diet-microbiome-host interaction. That is, each individual should be in the same diet-specific cohort across all taxonomic levels to keep the hierarchical fidelity of the microbial community, i.e. a consistent nutrition class across Phylum, Class, Family, Genus, etc.

To this end, we propose a novel Nutritional-Ecotype Mixture of Experts (NEMoE) approach for uncovering associations between the gut microbiome profile and the health state of an individual that takes into account diet-specific cohort heterogeneity (Figure 4.1 and Figure 4.2). This is achieved by using a regularised mixture of experts model to simultaneously optimise the separations between nutritional-ecotypes, the classification performance of microbiota, and the health state. The mixture of expert models has been widely used in integrating different types of data. Lê Cao, Meugnier and McLachlan [85] have used it for combining clinical data and genomics data. However, this work does not use sparse regularisation and lacks interpretability i.e. it is unable to obtain unique markers in each experts network. NEMoE also integrates a model parameter sharing strategy to account for the taxonomic information contained in microbiome data, ensuring coherent nutritional classification is maintained across all taxonomic levels. We show through empirical computational simulation research that NEMoE is robust to parameter changes. We also apply NEMoE to real microbiome data from a PD cohort and show that the model outperforms existing approaches in predictive performance and is able to uncover candidate diet-specific microbiome markers of complex diseases.

4.2 Methods

4.2.1 Overview of NEMoE

NEMoE identifies nutritional-ecotypes that represent differential dietary intake as well as the relationship between microbiome structure and host health (Figure 4.1 and 4.2). This approach has two distinct components: first, a gating network aimed at estimating latent classes shaped by nutritional intake, and second, an experts network aimed at modelling the relationship

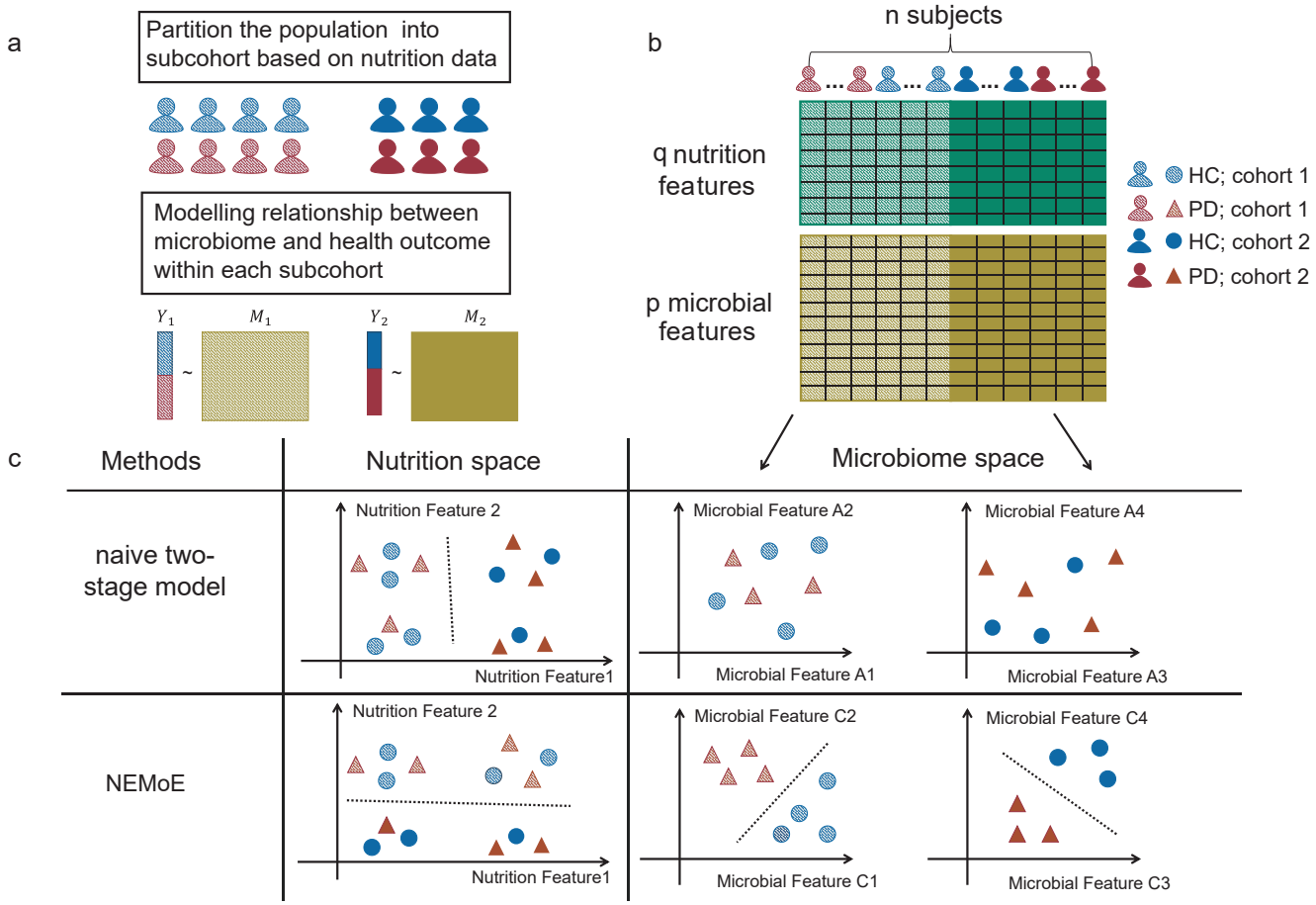

FIGURE 4.2. **Illustration of NEMoE and two-stage model**. a) A common workflow of a two-stage model: NEMoE first clusters the cohort based on the nutrition intake, then builds a model between microbiome data and health outcomes within each cohort. b) Illustration of a two-stage model with two latent classes. c) Illustration of two methods NEMoE and naive two-stage model in both nutrition space and microbiome space. Naive two-stage model identified two latent classes that showed best separation in the nutrition space but do not count for the relationship between microbiome and health outcome; two latent classes identified by NEMoE showed a differential relationship between microbial features and health outcome.

between the microbiota composition and the health state within each latent class [85, 190]. The input of the gating network is a nutrition matrix, with each variable being the nutrient intake of the individual and the corresponding microbiome measurements are used as input of the experts network. Similar to the non-regularised mixture of experts (MoE) models, fitting NEMoE involves estimating the parameters via maximum likelihood estimation to optimise

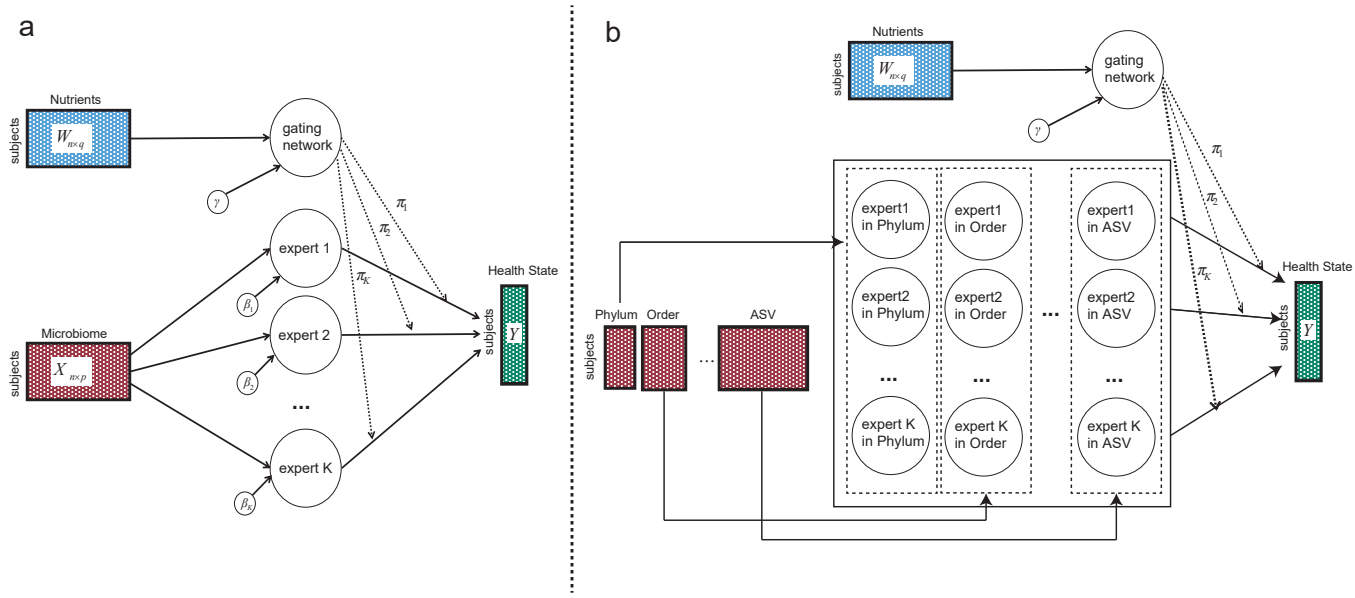

FIGURE 4.3. **Graphical model representation of NEMoE.** a) A graphical model representation of the mixture of experts model. b) A graphical model representation of NEMoE extends the MoE to multi-level with a shared gating network.

the separations amongst nutritional-ecotypes, microbiome classification performance, and the health state simultaneously (see Figure 4.3). The optimisation procedure is usually achieved by an expectation maximisation (EM) algorithm. However, the MoE model does not extend to a large number of feature variables (p) and a small sample size (n) framework, which often occurs in diet and microbiome data where there are many more features than observations. Instead, NEMoE adopts a regularisation component to the MoE (RMoE [19]) by adding elastic net penalties [195] on both the gating function and the experts network (details in the Methods section). Next, NEMoE employs a parameter sharing strategy that involves a shared gating network for the microbiome relative abundance matrices across taxonomic levels, to ensure coherent latent classes across all taxonomic levels. Compared with a latent class using purely nutritional intake, our nutritional-ecotype has two advantages: (i) it takes the relationship between microbiome and health outcome into account and is beneficial for identifying diet-specific microbial signatures (Figure 4.2). (ii) It incorporates the taxonomic structure in the latent class and keeps the hierarchical fidelity of the microbial community.

4.2.2 Mathematical formulation of NEMoE

The development of NEMoE was inspired by a mixture of experts approach to model heterogeneous data as shown in Figure 4.4 (a). In machine learning, the concept of "gate" [104] can be thought of as a decision-making component given some input. Our approach consists of two key components, a "gating network" sets up to determine which nutritional-class the sample belongs to and a "k-experts network" of size k to build classifiers for each nutritional-class. NEMoE uses a regularised MoE (RMoE) model, which adds elastic-net penalties to both the gating network and the experts network. Regularisation is needed here because a non-regularized MoE does not extend to a large p small n framework [44] where the number of features (p) is much larger than the number of samples (n). This data characteristic often occurs in diet and microbiome data where there are many more microbial features (p) than individual samples (n). NEMoE further incorporates the taxonomic information into RMoE by jointly optimising RMoE models from all taxonomic levels with the added constraint that all RMoE share the same gating network (Figure 4.4 b).

For a transformed microbiome data at taxonomic level l, we use the matrix $X_{n\times p_l^{(l)}}$to denote the relative abundance of n samples of p_l taxa. The corresponding diet information, measured as a nutrients intake matrix, is denoted as $W_{n\times q}$, where the q columns are the nutrient metrics for the same n samples. Let Y_n denote the binary response of the health outcome, with $Y = 1$ and $Y = 0$ representing individuals with and without disease, respectively. NEMoE models the heterogeneous relationship between the microbiome and the health outcome by a mixture distribution, i.e.

$$\mathrm{P}_l(Y = 1|X^{(l)}, W) = \sum_{k=1}^{K} \pi_k \frac{\exp(X^{(l)}\beta_k^{(l)})}{1 + \exp(X^{(l)}\beta_k^{(l)})}, \tag{4.1}$$

where $\pi_k = \frac{\exp(W\gamma_k)}{\sum_{i=1}^{K}\exp(W\gamma_i)}$ is the nutrition class mixing weight of shared components determined by nutrients intake, and where $\beta_k^{(l)}$ and γ_k are the corresponding effect size for the gating network and the experts network, respectively, and K denotes the predetermined number of nutrition classes. NEMoE estimates the regularised sum of all levels of the log-likelihood (LL) function in Equation (4.1), where the regularisation term consists of elastic net penalties

for both the gating network and the experts network:

$$\text{rLL} = \sum_{l=1}^{L}\sum_{k=1}^{K}\{\sum_{i=1}^{n}\log[P(Y_i|X_i^{(l)}, W_i)] - \phi(\lambda_{1k}^{(l)}, \alpha_{1k}^{(l)}, \beta_k^{(l)}\} - \phi(\lambda_2, \alpha_2, \gamma), \tag{4.2}$$

where $\phi(\lambda, \alpha, \beta) = \lambda[\alpha + \frac{1}{2}(1-\alpha)||\beta||_2^2]$ is the elastic net penalty function and $\lambda_{1k}^{(l)}$, $\alpha_{1k}^{(l)}$, λ_2, α_2 are the corresponding parameters for penalties in the experts network and in the gating function.

The regularised log-likelihood (rLL) can be maximised through a proximal Newton Expectation Maximisation algorithm [44]. Details of the optimization procedure can be found in the reference manual of the NEMoE package https://sydneybiox.github.io/NEMoE.

4.2.3 NEMoE algorithm

The implementation of NEMoE involves the EM algorithm, a special case of the Minorise-Maximisation (MM) algorithm and widely used in estimating parameters for finite mixture models. By alternatively inferring the latent variables given the parameters (E steps), and then optimising the parameters given the "filled in" data, the EM algorithm iteratively finds an appropriate local maximiser of the log-likelihood function.

In this section, we introduce the implementation of NEMoE. We first describe the EM algorithm parameters estimation for RMoE, and then introduce the EM algorithm for NEMoE. To achieve robust parameter estimation, we use three different strategies for initialisation. We also implement different variants of the EM algorithm to suit different scales of the problem. Finally, the selection of the tuning parameter in NEMoE is also introduced.

4.2.3.1 EM algorithm of RMoE

In this section, we first derive the optimisation for RMoE, i.e. set $L = 1$ in (4.2). We denote $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ and $\boldsymbol{\theta}^{(t)} = \left\{\boldsymbol{\gamma}^{(t)}, \boldsymbol{\beta}^{(t)}\right\}$ as the parameters in the t^{th} iteration. For the regularised log-likelihood function in Equation (4.2), the EM algorithm runs as follows:

E-step: Compute the conditional expectation of the complete data log-likelihood function given the observed data D and current parameter, the corresponding expected complete data log-likelihood is as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta^{(t)}}) &= \mathbb{E}[PL(\theta)|D; \theta^{(t)}] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t)}[\log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}) + p_i(\beta_k; \boldsymbol{x_i}, y_i)] \\ &- \sum_{k=1}^{K} \lambda_{1k}\left[\alpha_{1k}|\beta_k| + \frac{1}{2}(1-\alpha_{1k})||\beta_k||_2^2\right] - \lambda_2\left[\alpha_2|\boldsymbol{\gamma}| + \frac{1}{2}(1-\alpha_2)||\gamma||_2^2\right], \end{aligned} \tag{4.3}$$

where $p_i(\beta_k; \boldsymbol{x_i}, y_i) = y_i \log\left[\frac{\exp(\boldsymbol{x_i}^T\beta_k)}{1+\exp(\boldsymbol{x_i}^T\beta_k)}\right] + (1-y_i)\log\left[\frac{1}{1+\exp(\boldsymbol{x_i}^T\beta_k)}\right]$ is the log-likelihood function of the logistic distribution and

$$r_{ik}^{(t)} = \frac{\pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}^{(t)})p_i(\beta_k^{(t)}; \boldsymbol{x_i}, y_i)}{\sum_{k=1}^{K}\pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}^{(t)})p_i(\beta_k^{(t)}; \boldsymbol{x_i}, y_i)} \tag{4.4}$$

is the responsibility that latent class k takes for sample i.

M-step: In the M step, we maximise the expected complete data log likelihood function in Equation (4.3). This maximisation can be achieved by maximising $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as follows:

$$\beta_k^{(t+1)} = \text{argmax}_{\beta_k \in \mathbb{R}^p}\sum_{i=1}^{n} r_{ik}^{(t)} p_i(\beta_k; \boldsymbol{x_i}, y_i) - \lambda_{1k}\left[\alpha_{1k}|\beta_k| + \frac{1}{2}(1-\alpha_{1k})||\beta_k||_2^2\right], \tag{4.5}$$

$$\boldsymbol{\gamma}^{(t+1)} = \text{argmax}_{\gamma \in \mathbb{R}^q}\sum_{k=1}^{K}\sum_{i=1}^{n} r_{ik}^{(t)} \log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}) - \lambda_2\left[\alpha_2|\boldsymbol{\gamma}| + \frac{1}{2}(1-\alpha_2)||\gamma||_2^2\right]. \tag{4.6}$$

From Equation (4.5), it can be seen that the updating of β_k is a weighted sparse logistic regression problem with $\boldsymbol{x_i}$ as covariates, y_i as response and $r_{ik}^{(t)}$ as weight for i^{th} sample. Equation (4.6) shows that the updating of $\boldsymbol{\gamma}$ is also a sparse multinomial regression problem with $\pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma})$, defined in Equation (4.1), as covariates, $r_{ik}^{(t)}$ as response.

To solve the optimisation (4.5) and (4.6), a proximal-Newton iteration approach can be used, which repeatedly approximates the regularised log-likelihood by a quadratic function (see section 4.2.3.3).

4.2.3.2 MM algorithm of NEMoE

For the regularised log-likelihood function in Equation (4.2), we use the MM algorithm to estimate the parameters.

Let $\Theta = \left\{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(L)}, \boldsymbol{\gamma}\right\}$ be aggregated parameters in Equation (4.2). The estimated parameters and log likelihood function at t^{th} iteration is $\Theta^{(t)}$ and $\mathrm{rLL}_m(\Theta^{(t)})$. We have

$$\begin{aligned}\mathrm{rLL}_m(\Theta) - \mathrm{rLL}_m(\Theta^{(t)}) \geq & \sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(l),(t)}[\log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}) + p_i(\beta_k^{(l)}; \boldsymbol{x_i}, y_i)] \\ & - \sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(l),(t)}[\log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}^{(t)}) + p_i(\beta_k^{(l),(t)}; \boldsymbol{x_i}, y_i)],\end{aligned} \tag{4.7}$$

where

$$r_{ik}^{(l),(t)} = \frac{\pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}^{(t)}) p_i(\beta_k^{(l),(t)}; \boldsymbol{x_i}^{(l)}, y_i)}{\sum_{k=1}^{K} \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}^{(t)}) p_i(\beta_k^{(l),(t)}; \boldsymbol{x_i}^{(l)}, y_i)}. \tag{4.8}$$

Similar to Equation (4.3), we can define the multi-level Q function as

$$\begin{aligned}Q(\Theta, \Theta^{(t)}) = & \sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(l),(t)}[\log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}) + p_i(\beta_k^{(l)}; \boldsymbol{x_i}^{(l)}, y_i)] \\ & - \sum_{l=1}^{L}\sum_{k=1}^{K} \lambda_{1k}^{(l)} \left[\alpha_{1k}^{(l)}|\beta_k^{(l)}| + \frac{1}{2}(1-\alpha_{1k}^{(l)})||\beta_k^{(l)}||_2^2\right] \\ & - \lambda_2 \left[\alpha_2|\boldsymbol{\gamma}| + \frac{1}{2}(1-\alpha_2)||\gamma||_2^2\right].\end{aligned} \tag{4.9}$$

With this definition of the Q function, we have

$$\mathrm{rLL}_m(\Theta) - \mathrm{rLL}_m(\Theta^{(t)}) \geq Q(\Theta, \Theta^{(t)}) - Q(\Theta^{(t)}, \Theta^{(t)}). \tag{4.10}$$

Thus $Q(\Theta, \Theta^{(t)}) - Q(\Theta^{(t)}, \Theta^{(t)}) + \mathrm{rLL}_m(\Theta^{(t)})$ is a minoriser [62] of $\mathrm{rLL}_m(\Theta)$ at $\Theta^{(t)}$. Similar to the EM algorithm, a local maximum can be computed by maximising Equation (4.9) iteratively. Similar to equations (4.5) and (4.6), we have

$$\hat{\beta}_k^{(l),(t+1)} = \mathrm{argmax}_{\beta_k^{(l)} \in \mathbb{R}^{pl}} \sum_{i=1}^{n} r_{ik}^{(l),(t)} p_i(\beta_k^{(l)}; \boldsymbol{x_i}^{(l)}, y_i) - \lambda_{1k}^{(l)} \left[\alpha_{1k}^{(l)} |\beta_k^{(l)}| + \frac{1}{2}(1 - \alpha_{1k}^{(l)}) ||\beta_k^{(l)}||_2^2 \right], \tag{4.11}$$

$$\hat{\boldsymbol{\gamma}}^{(t+1)} = \mathrm{argmax}_{\gamma \in \mathbb{R}^q} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{i=1}^{n} r_{ik}^{(l),(t)} \log \pi_{ik}(\boldsymbol{w_i}, \boldsymbol{\gamma}) - \lambda_2 \left[\alpha_2 |\boldsymbol{\gamma}| + \frac{1}{2}(1 - \alpha_2) ||\gamma||_2^2 \right]. \tag{4.12}$$

Equations (4.11) and (4.12) also can be solved by a proximal Newton method. Taking together, we have Algorithm 1 to estimate the parameters in NEMoE.

Algorithm 1 NEMoE Algorithm

Input: L-levels microbiome matrix (phylum to ASV) $X_{n \times p_l}^{(l)}$, diet matrix $W_{n \times q}$;
regularisation parameters $\lambda_{1k}^{(l)}$, λ_2, $\alpha_{1k}^{(l)}$, α_2,
initial values of parameters $\beta_k^{(0)}$, $\gamma^{(0)}$, $k = 1, \ldots, K$.

Output: MLE of $\hat{\beta_k^{(l)}}$, $\hat{\gamma}$

1: t = 1
2: **repeat**
3: At iteration t
4: **E step:** Update $r_{ik}^{(l),(t)}$ by (4.8), $k = 1, \ldots, K$, $l = 1, \ldots, L$.
5: **M step:** Update $\beta_k^{(l),(t+1)}$ and $\boldsymbol{\gamma}^{(t+1)}$ by solving (4.11) and (4.12), $k = 1, \ldots, K$.
6: t = t+1
7: **until** the stopping criterion is satisfied.

4.2.3.3 Proximal Newton iteration for EM inner loop optimisation

One of the most efficient algorithms to solve problem (4.5), (4.6), (4.11), (4.12) is the proximal Newton method. It first replaces the cross entropy type of term (the first term) in Equation (4.5) and (4.6) with its quadratic approximation. Then a weighted elastic net problem is optimised with coordinate descent methods.

We take the optimization of (4.5) as an example and note that a similar procedure can be implemented for the other parameter estimations or, alternatively, one can directly use sparse multinomial regression as for example implemented in the R package glmnet.

Let $J(\beta_k) = \sum_{i=1}^{n} r_{ik}^{(t)} p_i(\beta_k; \boldsymbol{x_i}, y_i)$, we have

$$\nabla J(\beta_k) = \sum_{i=1}^{n} r_{ik}^{(t)} \left[y_i - p_i(\beta_k; \boldsymbol{x_i}, y_i) \right] \boldsymbol{x_i}, \tag{4.13}$$

$$\nabla^2 J(\beta_k) = \sum_{i=1}^{n} r_{ik}^{(t)} \left[1 - p_i(\beta_k; \boldsymbol{x_i}, y_i) \right] p_i(\beta_k; \boldsymbol{x_i}, y_i) \boldsymbol{x_i} \boldsymbol{x_i}^T. \tag{4.14}$$

Thus, given the current estimator $\tilde{\beta}_k$, $J(\beta_k)$ can be approximated by

$$\begin{aligned} I(\beta_k) &\approx J(\tilde{\beta}_k) + (\beta_k - \tilde{\beta}_k)^T \nabla J(\beta_k) + \frac{1}{2} (\beta_k - \tilde{\beta}_k)^T \nabla^2 J(\beta_k) (\beta_k - \tilde{\beta}_k) \\ &= -\frac{1}{2} \sum_{i}^{n} r_{ik}^{(t)} s_i (c_i - \boldsymbol{x_i} \beta_k)^2 + C(\tilde{\beta}_k), \end{aligned} \tag{4.15}$$

where $C(\tilde{\beta}_0, \tilde{\beta}_k)$ is a constant unrelated to β_k and

$$c_i = \boldsymbol{x_i} \tilde{\beta}_k + \frac{y_i - p_i(\tilde{\beta}_k; \boldsymbol{x_i}, y_i)}{p_i(\tilde{\beta}_k; \boldsymbol{x_i}, y_i) \left[1 - p_i(\tilde{\beta}_k; \boldsymbol{x_i}, y_i) \right]}, \tag{4.16}$$

$$s_i = p_i(\tilde{\beta}_k; \boldsymbol{x_i}, y_i) \left[1 - p_i(\tilde{\beta}_k; \boldsymbol{x_i}, y_i) \right]. \tag{4.17}$$

Denote the objective function in Equation (4.5) as $I(\beta_k)$ and we have

$$\begin{aligned} I(\beta_k) &= J(\beta_k) - \left[\alpha_{1k} |\beta_k| + \frac{1}{2} (1 - \alpha_{1k}) ||\beta_k||_2^2 \right] \\ &\approx -\frac{1}{2} \sum_{i}^{n} r_{ik}^{(t)} s_i (c_i - \boldsymbol{x_i} \beta_k)^2 - \left[\alpha_{1k} |\beta_k| + \frac{1}{2} (1 - \alpha_{1k}) ||\beta_k||_2^2 \right] + C(\tilde{\beta}_k). \end{aligned} \tag{4.18}$$

Equation (4.18) is a weighted elastic net problem and can be efficiently solved by the coordinate descent algorithm. In our implementation, we use functions in the R package glmnet to solve this optimisation problem.

For other optimisations such as Equation (4.6), a similar procedure can be made by just dropping the $r_{ik}^{(t)}$ term in Equation (4.18) and replacing y_i in Equation (4.16) with the current $r_{ik}^{(t)}$.

4.2.3.4 Initial values of the EM algorithm

Maximisation of (4.2) is a non-convex problem and heavily depends on the choice of initial values. Random initialisation will give an unstable estimation of parameters in NEMoE. To alleviate the effect of initialisation and give a more stable estimation, we adopt the following ways to initialise Algorithm 1.

Clustering initialisation: This initialisation first uses a clustering method such as kmeans with the input of the gating network, $W_{n\times q}$, which clusters the data into K classes. Then the corresponding classes are used as the initial value of r_{ik}. Finally, the initial parameters of $\beta_k^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ can be computed by solving (4.5) and (4.6), $k = 1, \ldots, K$.

Elnet initialisation: Here we first fit an elastic net model using all of the samples. Then all the experts networks of NEMoE are initialised with the same parameters as in the elastic net. The parameter r_i or $r_i^{(t)}$ are initialised with $\frac{1}{K}, \ldots \frac{1}{K}$ for each of the expert networks.

Small EM initialisation: Small EM initialisation uses a small number of iterations (10 iterations in our implementation) of the EM or MM algorithm with different initialisations. In our method, we implement both randomised, clustering and elnet initialisation with a small EM procedure and compare their rLL after 10 iterations. Parameters in the EM with the largest rLL are chosen in the following optimisation.

4.2.3.5 Selection of tuning parameters

In Equation (4.2), appropriate parameters λ_{1k}, λ_2 and K need to be determined to get good estimation of parameters. Some studies such as [19, 74, 146] use the Bayesian information criterion (BIC) for parameter selection. The BIC is given by

$$BIC(\hat{\beta}, \hat{\gamma}, K) = -L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + \frac{d(\lambda_{1k}, \lambda_2, K)}{2} \log n, \tag{4.19}$$

where $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ are the parameters estimated by Algorithm 1 and $d(\lambda_{1k}, \lambda_2, K)$ is the number of free parameters involved in the estimated model.

It was shown that for finite mixture models, under certain regularity conditions, the BIC consistently estimates the number of mixture components [72], and numerical experiments show that the BIC works well at a practical level [42, 146].

However, as is pointed out in Biernacki, Celeux and Govaert [13] and Baudry et al. [12], the BIC tends to overestimate the number of clusters, in the case of MoE, with the number of parameters involved in the degrees of freedom, not only the number of clusters but also the variables in the model tend to be overestimated. Also, the BIC used in Chamroukhi and Huynh [19], Khalili and Chen [74] and Khalili [73] substitute the degrees of freedom with the number of variables involved in the model, which cannot be directly applied in the case where W and X are not identical.

To overcome this overestimation issue, Biernacki, Celeux and Govaert [13] proposed an integrated classification (ICL) criterion. Instead of approximating the integrated observed likelihood function, the ICL uses an integrated completed likelihood approximation and is given by

$$ICL(\lambda_{1k}, \lambda_2, K) = -L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + \mathrm{Ent}(\hat{r}_{ik}) + \frac{d(\lambda_{1k}, \lambda_2, K)}{2} \log n \tag{4.20}$$

where $\mathrm{Ent}(\hat{r}_{ik}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{r}_{ik} \log \hat{r}_{ik}$ is the entropy of the fuzzy classification matrix and $\hat{r}_{ik}$ is the estimated latent class defined in (4.4).

In our implementation, we implement all of the criteria above including BIC, ICL and 5-fold cross validation to allow users to choose the parameters flexibly. In our simulations, we find that ICL tends to select the least number of latent classes while cross validation selects the most. To get better biological meaning, we use cross validation in our real data analysis.

4.2.4 Simulation study

Our simulation is inspired and integrated from multiple simulation studies. These include the generation of the nutrition data based on a multivariate Gaussian distribution [23] and a sparse multinomial regression model; the generation of the microbiome data using a zero-inflated latent Dirichlet allocation model (zinLDA) [36] and constructing the health outcome using sparse logistic regression. The details of the simulation are described as follows.

The simulation of the nutritional data consists of two main components, the values of the nutritional measurements (W) and the underlying latent class which we refer to as nutritional-ecotype (L).

Constructing the effect size between W and L: For a given number of nutritional features q and given the total number of latent classes K, we first simulate the effect size between W and L, denoted by the matrix $\gamma_{q\times K}$, where each element γ_{jk}, $j = 1, \ldots, q$ and $k = 1, \ldots, K$ represents the effect size of the j^{th} nutrition feature on the k^{th} latent class. We randomly select five nutrition features to have non-zero effect size and its value is either c_g or $-c_g$ with equal probability, i.e.

$$\gamma_{jk} = \begin{cases} 2c_g\left(B_{jk} - \frac{1}{2}\right), & j \in A_k, \\ 0, & j \notin A_k, \end{cases} \tag{4.21}$$

where $B_{jk} \sim$ *Bernoulli* $\left(\frac{1}{2}\right)$ and A_k denotes a set of nutritional variables with non-zero effect size randomly drawn from the q nutritional variables, $c_g > 0$ is a constant that controls the overall strength of the effect size and B_{jk} is a Bernoulli random variable with probability parameter 1/2. This is designed such that the effect of a nutritional feature can randomly have a positive or negative value with equal probability. Selection of five nutrition features allows only a fraction of nutrition features to contribute to the ecotype.

Simulating nutritional values W and nutritional-ecotype L: The nutritional values are simulated with a q-variate Gaussian mixture distribution with K components. Given the i^{th} sample, we first generate its nutritional-ecotype L_i by randomly drawing an index from the set $\{1, \ldots, K\}$ indicating which latent class it belongs to. For $L_i = k$, its corresponding

nutrition data W_i is generated from a normal distribution

$$W_i|L_i = k \sim N\left(\mu_k, \Sigma_k\right), \tag{4.22}$$

where μ_k and Σ_k are the mean and covariance matrix of the k mixture component, respectively. In our simulation, we set Σ_k to be a matrix with diagonal entries equal to 1 and off-diagonal entries equal to ρ, we further set μ_k to be proportional to the effect size of the k^{th} latent class i.e.

$$\mu_k = \eta\gamma_k. \tag{4.23}$$

The use of Gaussian mixture distributions for nutrition data is an extension of the simulation method by Chen and Li [23]. The levels of heterogeneity between dietary patterns were controlled by different levels of η. A larger η represents a greater gap between the means of the components and thus, the larger that gap the easier it is to distinguish the nutritional ecotype k (Figure 4.6 a). In our simulation, we define "none", "weak", "mild" and "strong" separation by setting the maximum absolute value η to be 0, 0.1, 0.3, and 0.5, respectively.

Simulating microbiome data X: The generation of the microbiome data follows a zero-inflated latent Dirichlet allocation model (zinLDA) [36]. This model simulates several typical characteristics of the microbiome data including over-dispersion, zero-inflation and high dimensionality. The distribution of the microbiome data X is given by,

$$X \sim \textit{zinLDA}\left(\pi_m, a_m, b_m, \alpha_m, K_m, N_m\right), \tag{4.24}$$

where N_m is the total number of reads, K_m is the number of microbiome subcommunities, α_m is the parameter in the Dirichlet distribution, π_m relates to the generation of the subcommunity, a_m and b_m are the parameters in the zero-inflated generalized Dirichlet distribution related to the generation of counts within each subcommunity. Our simulation uses the default parameters as in zinLDA [36]: N_m was drawn from a discrete uniform distribution with a lower bound of 5,000 and an upper bound of 25,000, $\alpha_m = 10$, $K_m = 5$, $\pi_m = 0.4$, $a_m = 0.05$ and $b_m = 10$ respectively. The microbial counts at ASV level were generated using the corresponding distribution of p variables.

For each simulation, we generate n samples of microbiome data at 5 taxonomic levels, where level 1 to level 5 correspond to Phylum, Order, Family, Genus and ASV, respectively and $(p_1, \ldots, p_5) = (10, 20, 50, 80, p)$ corresponds to the number of microbiome variables for each level respectively. The hierarchical structure of the taxonomic table is generated by randomly grouping microbial features from the higher neighbouring level $(l+1)$ into the p_l group when going from the $(l+1)$ level to the l level. We denote $\text{ASV}^{(l)}(j)$ as the set of taxa at ASV level that is mapped to taxa j at level l. The simulated microbiome data of the l taxonomic table was thus hierarchically aggregated from the taxa from the $(l+1)$ taxonomic table and the corresponding count matrix is denoted as $X^{(l)}_{n \times p_l}$.

Simulating health outcome Y: The relationship between microbiome and health outcome was simulated based on a mixture of sparse logistic regression which extends the model by Dong et al. [37]. We simulate the binary health outcome in three steps.

Step 1: Simulate the microbial signatures of each latent class at ASV level. The microbial signatures were selected from candidates' taxa that satisfy two conditions: prevalence is larger than 50% (counts are non-zero in at least 50% of samples) and the variance of its relative abundance is larger than 10^{-6}. Then we randomly select 5 taxa from the candidate taxa for the k^{th} latent class and denote the corresponding sets of taxa as $A_k^{(\text{ASV})}, k = 1, \ldots, K$.

Step 2: Simulate the effect size between microbiome $X^{(l)}$ and health outcome Y at the l^{th} taxonomic level and latent class k. We first generate the effect size at ASV level, which we denote as $\beta_k^{(ASV)}$. The i^{th} element of $\beta_k^{(ASV)}$ was generated as follows:

$$\beta_{ik}^{(\text{ASV})} = \begin{cases} \frac{2c_e}{\sigma_{ik}} \left(B_{ik} - \frac{1}{2}\right), & i \in A_k^{(\text{ASV})} \\ 0, i \notin A_k^{(\text{ASV})}, \end{cases} \tag{4.25}$$

where $B_{ik} \sim$ *Bernoulli* $\left(\frac{1}{2}\right)$, σ_{ik} is the standard deviation of taxa i in nutrition class k and $c_e > 0$ is a constant that controls the strength of the effective size, and B_{jk} is a Bernoulli random variable with probability parameter 1/2. Then the effect size of the j^{th} taxa at level l, that is $\beta_{jk}^{(l)}$ is generated by aggregating its corresponding effect

size at ASV level,

$$\beta_{jk}^{(l)} = \sum_{i \in A_k^{(\text{ASV})}} \beta_{ik}^{(\text{ASV})}. \tag{4.26}$$

Note that for taxonomic level l, the corresponding effect sizes were generated as the sum of the level $l+1$ coefficients similar to as in Wang and Zhao [166]. Here, the strength of effect size is inversely proportional to σ_{ik} and ensures that the contribution of the microbiome signature is not mainly affected by its relative abundance.

Step 3: Simulate the health outcome Y. We simulate the probability that "the health outcome Y equals 1" as an average of all 5 levels mixture sparse logistic regression with the mixing weight $\pi_k = \frac{\exp(W\gamma_k)}{\sum_{i=1}^{K}\exp(\mathrm{W}\gamma_\mathrm{i})}$,

$$\mathrm{P}\left(Y=1|X^{(1)},\ldots,X^{(5)},W,\beta_k^{(1)},\ldots,\beta_k^{(5)},\gamma\right) = \frac{1}{5}\sum_{l=1}^{5}\sum_{k=1}^{5}\pi_k\frac{\exp\left(X^{(l)}\beta_k^{(l)}\right)}{\sum_{k=1}^{K}\exp\left(X^{(l)}\beta_k^{(l)}\right)}. \tag{4.27}$$

For given microbiome data $X_{n\times p_l}^{(l)}$, nutrition data $W_{n\times q}$ and the corresponding effect size γ and $\beta_k^{(l)}, (l=1,\ldots,5)$. We draw n Bernoulli pseudo-random samples with probability parameter given by equation (4.27) to obtain the binary health outcome vector $Y_{n\times 1}$.

Our simulation first generated independent data of $2n$ samples from the procedure described above, then the first n samples were used for training and another n samples were used to calculate the predicted accuracy. The details of parameter settings in each simulation are described in Table 4.1.

4.2.5 Real world datasets

We also evaluate NEMoE on a set of in-house and publicly available microbiome-PD datasets. We provide information about the process and details of the datasets below.

TABLE 4.1. Summary of Simulation settings.

Simulation description	n	p***	q	η	c_e	c_g	K	ρ
Evaluate the effect of n	$(100, 200, 500, 1000)$	50	30	0.1	2	2	2	0
Evaluate the effect of η	200	50	30	$(0, 0.1, 0.3, 0.5)$	2	2	2	0
Evaluate the effect of p	200	$(30, 50, 80, 100)$	30	0.1	2	2	2	0
Evaluate the effect of q	200	50	$(30, 50, 80, 100)$	0.1	2	2	2	0
Evaluate the effect of ρ	200	50	30	0.1	2	2	2	$(0, 0.1, 0.3, 0.5)$
Evaluate the effect of K *	$(100, 200, 500, 1000)$	50	30	$(0, 0.1, 0.3, 0.5)$	2	2	3	0
Evaluate the effect of multi-level data **	500	50	30	0.1	2	2	2	0

* For the evaluation of the effect of K, the underlying simulation data is generated based on $K = 3$, while the fitted NEMoE is based using K ranging from 2 to 4.
** For the evaluation of the multi-level, we compare the adjusted rand index between NEMoE using all 5 levels data (Phylum, Order, Family, Genus and ASV) with NEMoE using only one level data.
*** Except the evaluation of multi-level, all evaluations were performed based on single level data. For the multi-level data, the number of variables for Phylum, Order, Family, Genus and ASV levels are 30, 50, 80, 100 respectively.

PD-microbiome: A gut microbiome dataset containing 101 PD patients and 83 healthy controls. The stool samples were collected and 16S rRNA V3-V4 amplicon sequencing was performed on an Illumina MiSeq platform.

PD-diet: Dietary information was collected by a comprehensive Food Frequency Questionnaire and resulted in a table of nutrient intake with 23 macronutrients, presented earlier [117]. Details of the sample information and sequence processing can be found in Lubomski et al. [100] and Lubomski et al. [99].

Public validation (PV) studies: We curated a series of datasets from eight different publicly-available microbiome studies [1, 58, 65, 135, 164] to further validate results from NEMoE. All the datasets were processed using the dada2 pipeline [17] (v1.16) and microbiome taxa

were annotated using taxonomy reference "silva 138" [125, 188]. Samples with low sequence reads (<1000) were excluded from the analysis. More information on these datasets can be found in Table 4.2.

TABLE 4.2. Summary of eight publicly available Parkinson's Disease microbiome studies used for validation of the NEMoE model.

Study	Design	Country	Sample size	16S region	Accession number
Lubomski_0 Lubomski_6 Lubomski_12	Longitudinal	Australia	74 PD, 74 HC	V3-V4	PRJNA808166
Wallen_1	Cross sectional	USA	323 PD, 184 HC	V4	PRJNA601994
Wallen_2	Cross sectional	USA	197 PD, 130 HC	V4	PRJNA601994
Aho(baseline) Aho(follow up)	Longitudinal	Finland	64 PD, 64 HC	V3-V4	PRJEB27564
Weis	Cross sectional	Germany	34 PD, 25 HC	V4-V5	PRJEB30615
Pietrucci	Cross sectional	Finland	72 PD, 72 HC	V1-V3	PRJEB4927
Scheperjans	Cross sectional	Germany	34 PD, 25 HC	V4-V5	PRJEB30615
Jin	Cross sectional	China	72 PD, 68 HC	V3-V4	PRJEB588834

Studies Lubomski_0, Lubomski_6 and Lubomski_12 were part of the same longitudinal data set by Lubomski and colleagues[99–101] and they represent samples that were measured at 0, 6 and 12 months, respectively.
Studies Aho(baseline) and Aho(follow up) were part of the same longitudinal data set by Aho and colleagues [1]. The same subjects were measured twice, at baseline and then later at follow-up, which was on average 2.25 years apart.
Studies Wallen_1 and Wallen_2 were part of two large cohort studies set by Wallen and colleagues [164].

4.2.6 Data processing

We process the microbiome and diet dataset to exclude outliers samples as well as keep the core features. Details of processing are described as follows.

PD-microbiome data processing. We excluded 7 samples with extremely large energy intake (>20,000 kJ per day), one subject with low microbial read counts (Total counts < 10000) and

two samples with missing nutrition measurements, resulting in 175 samples (75 HC individuals and 100 PD individuals). Raw counts from microbiome data were first normalised by total sum scaling, i.e. the counts (totals) were normalised into a composition proportion. Then core microbial features were kept and further transformed: Features that had more than 30% zeros in the n samples and features which had sample variance smaller than 10^{-5} were filtered out at each taxonomic rank resulting in the core microbial features of 7 Phylum, 19 Order, 27 Family, 41 Genus and 101 ASVs and 3,152,746 total reads were kept from 6,024,011 reads; variance stability transformation, i.e. an arcsin square root transformation, was performed on taxa proportion [37, 102]; the arcsin transformed data were further standardised to have mean zero and unit variance (z-score).

PD-diet features construction. In addition to the nutrients intake values, we calculated the percentage of energy intake as protein ($\%$EP), percentage of energy intake as fat ($\%$EF), percentage of energy intake as carbohydrate ($\%$EC) and protein intake and carbohydrate intake ratio (P:C) as additional variables. These transformations of nutritional features are widely used in nutri-omics studies [128, 141]. All of the 27 nutritional features were z-scored.

4.2.7 Performance evaluation

The performance of NEMoE and other methods are presented below. We also introduce the naive two-stage method to compare their ability to identify diet-subcohort. Finally, differential abundance analysis was used to compare the microbiome composition between PD and HC groups.

Comparison methods: Table 4.3 contains a summary of all methods used in the comparison study. We included the most commonly used methods in microbiome analysis as well as a naive two-stage approach. All of the comparisons were performed on simulation datasets and on in-house data on the Genus level.

Naive two-stage approach: The approach first clustered the nutrition data using unsupervised learning methods such as k-means. Then, based on the clustering result, samples in each cluster were used to build a classification model of microbiome and health state. The choice

of classification models we used in our simulation includes sparse logistic regression, support vector machine and random forest. For the sparse logistic model, SVM and the random forest model, we use microbiome data as input.

Differential abundance: We compared differential relative abundance between PD and HC in all datasets. The comparison was based on a non-parametric bootstrapping procedure. We resampled the data with replacement and then calculated the difference of the average relative abundance between PD and HC. This procedure was repeated 10,000 times for each taxon and the 95% confidence interval of the differential relative abundance was calculated.

4.3 Results

4.3.1 NEMoE is able to accurately identify nutritional latent classes shared across different taxonomic levels

We evaluated the efficiency of NEMoE in determining nutritional-ecotypes based on microbiota across different taxonomic levels using both simulated and experimental data. In our simulation study, we created a four-level dataset of 500 samples with a shared latent structure, where each individual belonged to a nutritional-ecotype and the relationship between microbiota and health status differed between two simulated nutritional-ecotypes. The adjusted rand index (ARI), a cluster comparison statistic, was used to compare the estimated nutritional-ecotypes and the underlying simulated latent classes shown in Figure 4.4 (a). We discovered that by incorporating hierarchical taxonomy information in our NEMoE approach, the estimated nutritional class was cohesive and performed better (higher ARI = 0.80) than nutritional-ecotypes estimated from a single taxonomy level (ARI = 0.75). NEMoE achieved this by sharing information across taxonomic levels and the estimated latent class incorporated information from all levels.

Next, we applied NEMoE to our in-house data from a gut microbiome PD study [99, 100]. A scatter plot from the first two components of a principal component analysis (PCA) of scaled nutrient intake (details of nutritional features see Section 4.2) from all individuals are shown

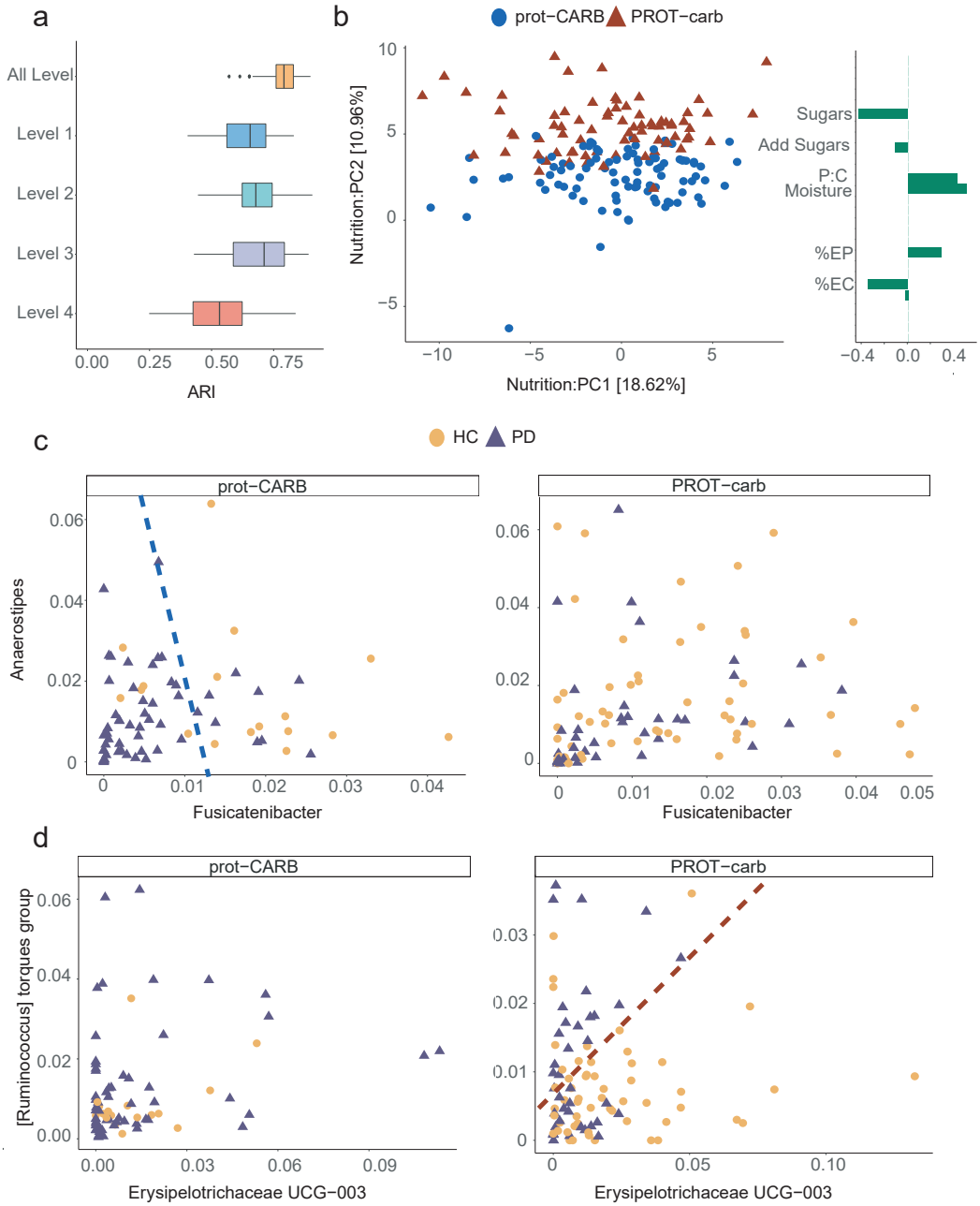

FIGURE 4.4. **Identification of nutritional-ecotype by NEMoE**. a) Boxplot comparing NEMoE and single level NEMoE in estimating shared latent classes. The ARI (x-axis) is calculated by comparing the estimated latent class and the true latent class from the data generating model. In all settings, NEMoE using multiple level information performs better. b) PCA plot of scaled nutrient intake for subjects colored by the two nutrition classes as estimated by NEMoE. Estimated coefficients of the gating network showed high coefficients for Sugar, Protein:Carbohydrate and Moisture. We denote the two nutrition classes as prot-CARB and PROT-carb with low protein-high carbohydrate intake and vice versa. c) Scatter plot of genera *Fusicatenibacter* and *Anaerostipes*. The Left panel shows that Parkinson's Disease and Healthy Controls in the prot-CARB subcohort roughly separate but there is no such separation in the PROT-carb right panel. d) Scatter plot of genera *Erysipelotrichaceae UCG-003* and *[Ruminococcus] torques group* showed a different relationship between Parkinson's Disease and Healthy Controls in two nutritional-ecotypes.

in Figure 4.4 (b), with the two nutritional-ecotypes best described as "high protein"-"low carbohydrate" (PROT-carb; shown in red) and "low protein"-"high carbohydrate" (prot-CARB; blue). The corresponding loadings show that these two ecotypes have very different ratios of protein to carbohydrate intake. Figure 4.4 (c) and (d) illustrate that the relationships between gut microbiota and PD status are different between these two nutrition-ecotypes, PROT-carb and prot-CARB. It is important to note that two identified subcohorts are significantly different to clusters identified by unsupervised clustering, such as subcohorts estimated by the k-means algorithm (ARI ~ 0, Figure 4.5).

We further established the generalisability of NEMoE by examining its impact when applied to data with different levels of heterogeneity. Here, we created synthetic datasets with four different degrees of separation (Figure 4.6 and Section 4.2.4), and demonstrated that NEMoE performs better than other existing approaches in detecting latent classes and this difference was more evident in challenging situations where the true separation between latent classes was small (Figure 4.7). This implies that NEMoE has the potential to perform well in many observational studies where nutrient intake patterns are mixed or difficult to separate, and hence the NEMoE approach can be applied broadly to human disease datasets with diverse dietary intakes.

4.3.2 NEMoE outperforms existing supervised methods in predicting Parkinson's disease state

We evaluated the predictive performance of NEMoE using both simulation and real data based on leave-one-out cross validation (LOOCV) to the area under the receiver operating characteristics curve (AUC) for the various models described in Table 4.3. In simulation studies we showed that under all comparison settings, NEMoE was able to achieve higher prediction accuracy (Figure 4.7), which implies NEMoE is robust to different parameter settings, such as n and p. Figure 4.6 (c) highlights that when NEMoE was applied to our in-house dataset from a gut microbiome PD study [100] with 2 latent classes (AUC = 0.78), it outperformed all other approaches, with the next best being random forest (AUC = 0.71). We also validated using 10-fold Monte Carlo cross validation. Compared to LOOCV, all of the

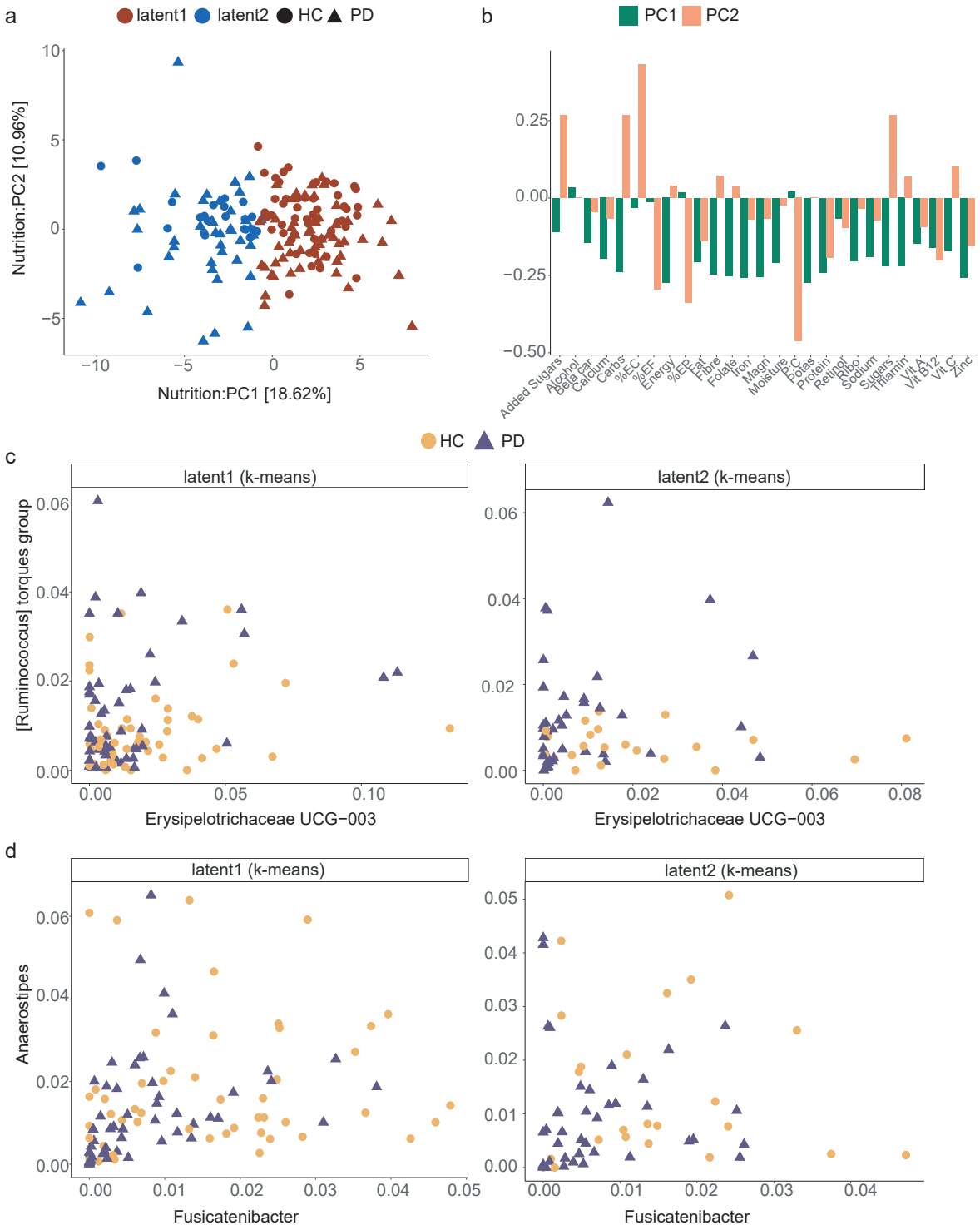

FIGURE 4.5. **Nutrition classes determined by k-means do not show an informative relationship between microbiome and PD.** a), PCA plot of scaled nutrient intake for subjects colored by two latent classes estimated by k-means. b), Loadings of the first two PCs. Loadings of the second PC shared some important variables with the coefficients of the gating network in NEMoE. c),d), Variables selected by NEMoE do not show a clear difference between PD and HC.

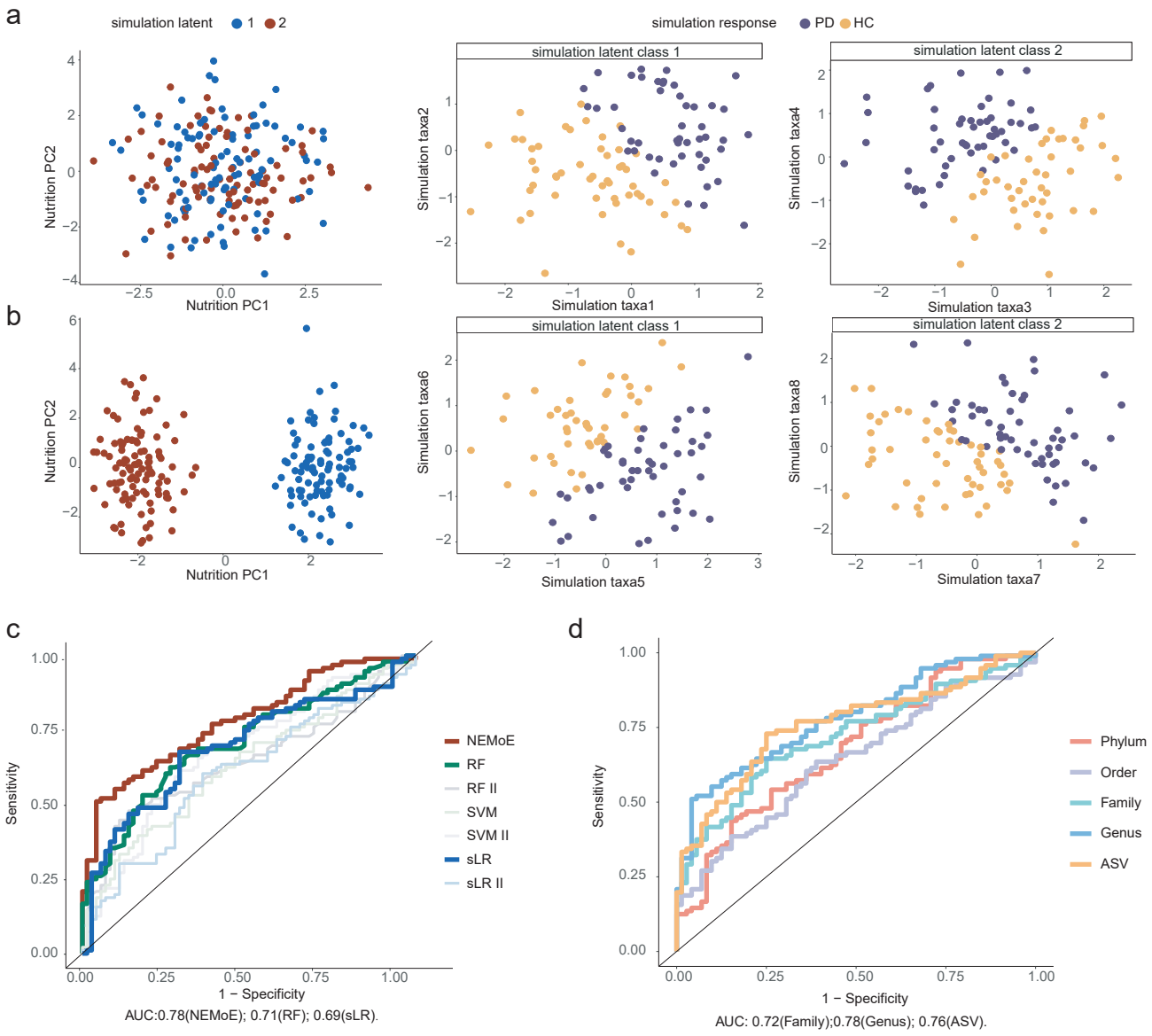

FIGURE 4.6. **Comparison of NEMoE on simulation dataset and real dataset.** a) An illustration of a non-separable case where nutrition intake does not show a difference between two nutritional-ecotypes, but each subcohort shows a different relationship between microbiome taxa and health state. b) An illustration of a separable case where nutrition intake is significantly different between two nutritional-ecotypes and relationships in each model are similar to the illustration in a). Simulation studies showed that NEMoE can identify both case a) and case b). c) Receiver Operating Characteristics curve of different methods (See Table 4.3) in predicting Parkinson's Disease using LOOCV. NEMoE showed the best LOOCV-AUC (AUC = 0.78). d) ROC plot of NEMoE at different taxonomic levels using LOOCV. Genus level showed the best predictive performance (AUC = 0.78).

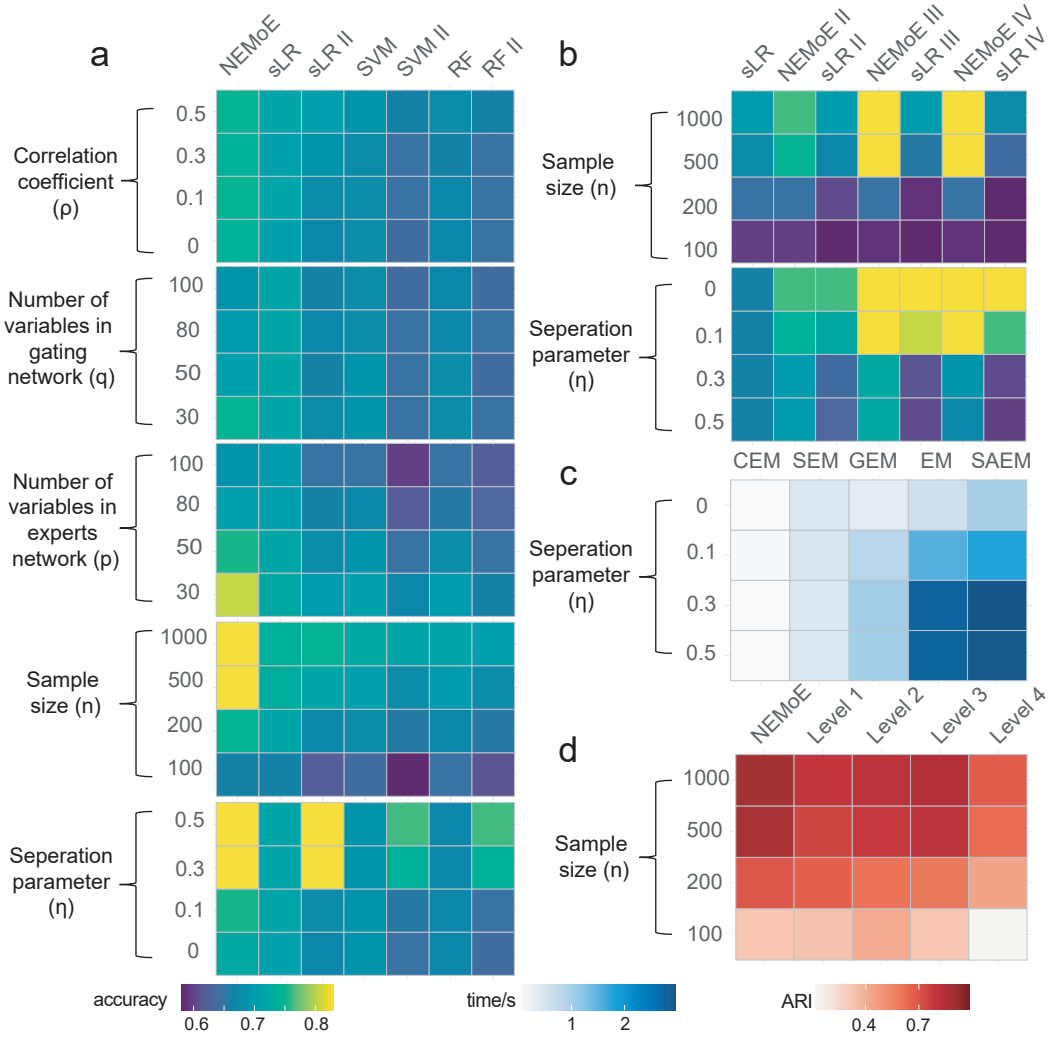

FIGURE 4.7. **Simulation results of NEMoE and other methods under different settings.** a) simulation results under different simulation parameters settings including the separation parameter η, sample size n, number of variables in the gating network q and in the experts network p and correlation ρ between variables. NEMoE achieves the best predictive performance compared with others in all settings with latent class structure. b) Simulation results of three latent classes with different n and η. NEMoE III and NEMoE IV perform well under most parameter settings. c) Time consumption of different EM-algorithms implemented in NEMoE. CEM was the fastest while achieving lower regularised LL, while EM and SAEM achieved higher regularised LL but required more time. d) Comparison between different taxa levels with NEMoE in the estimation of shared latent classes. Level K represents fitting RMoE with data in level $K(K = 1, 2, 3, 4)$. The ARI is calculated by comparing the estimated latent class and the generated true latent class.

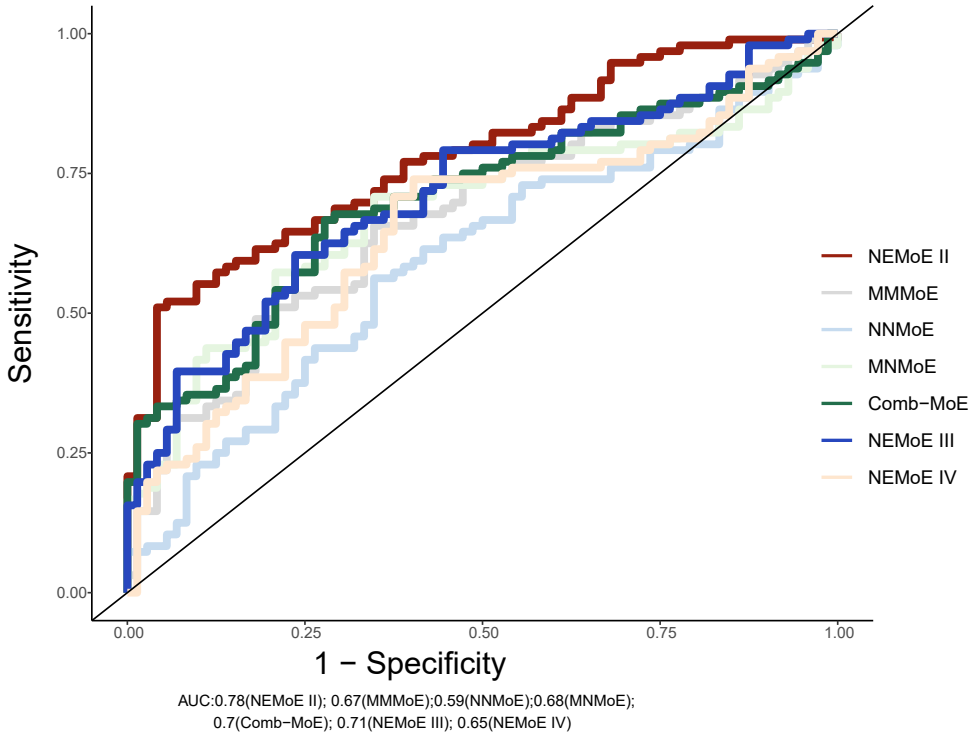

FIGURE 4.8. **Prediction performance of different types of input for NEMoE.** Using nutrient intake as the input of the gating network and microbiome as the input of the experts network showed better prediction performance than for the other cases. NEMoE with two nutrition classes showed the best prediction performance in our dataset.

methods showed a decreased AUC, however, NEMoE still outperform other methods (see Figure 4.9). Figure 4.8 further highlights that increasing the number of latent classes for this data did not improve the overall AUC.

NEMoE's ability to detect meaningful subcohorts via its joint optimisation approach is a key driver of this increase in accuracy. For example, when comparing to a naive two-stage model that uses unsupervised clustering to identify latent classes before fitting two independent models, the performance of NEMoE is considerably better, as indicated by the large difference in AUC (NEMoE = 0.78, sLR II= 0.6).

We further assessed NEMoE's capabilities on enterotype-separated subcohorts [29] within our PD dataset. Enterotype, a widely used concept in microbiome research, refers to the categorisation of an individual's microbiomes by the variance in composition [9, 180]. It is

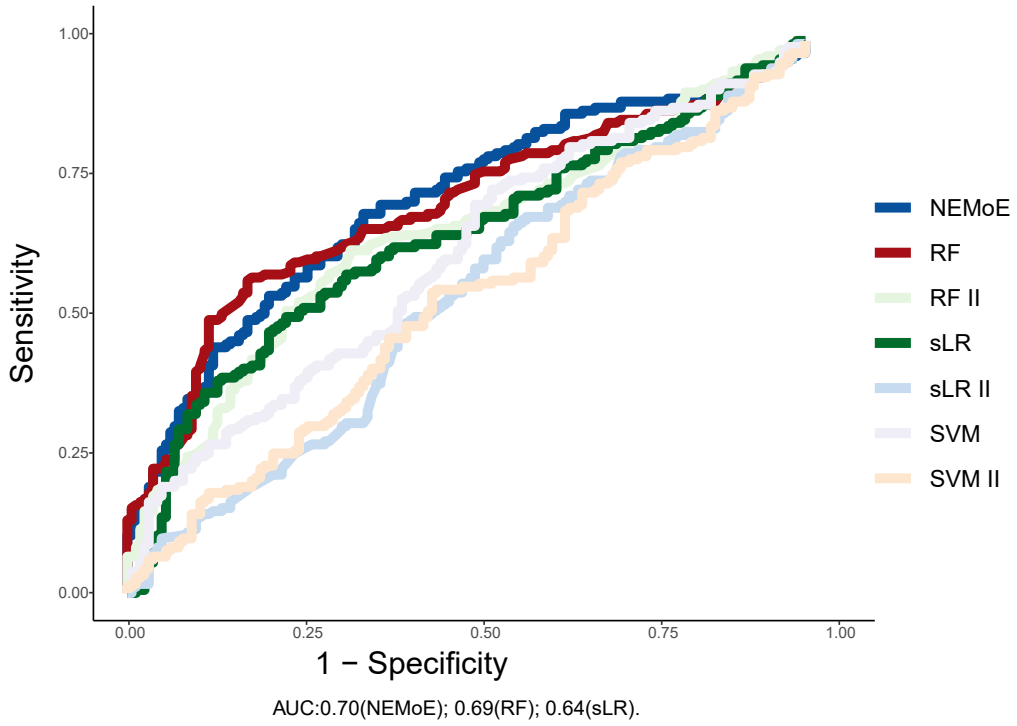

FIGURE 4.9. **Prediction performance the real dataset using 10-fold Monte Carlo cross validation.** 10-fold Monte Carlo cross validation showed similar result compared to LOOCV.

widely accepted that the enterotype captures stable compositional features of individuals and differences in community-type prevalence across populations with different long-term diets. In this study, we classify 87 samples as Enterotype B, 81 samples as Enterotype F and no samples as Enterotype P. The cluster memberships between the subcohorts determined by NEMoE and by enterotype had no more overlap than pure chance (ARI = 0). Furthermore, building a different classifier for each of the two enterotypes had a much lower (LOOCV-AUC = 0.65) predictive ability than NEMoE (LOOCV-AUC = 0.78). This suggests that NEMoE allows the model to focus more on each latent class and increases prediction performance by more precisely identifying subcohorts with differential microbiome-PD relationships.

TABLE 4.3. Summary of methods for comparison.

Method	Input data of identification subcohort	Input data of modelling within each subcohort	Model
sLR		Microbiome	Sparse logistic regression
SVM		Microbiome	Support Vector Machine
RF	Nutrition	Microbiome	Random forest
sLR K*	Nutrition	Microbiome	Two stage sLR with K latent class
SVM II	Nutrition	Microbiome	Two stage SVM
RF II	Nutrition	Microbiome	Two stage RF
NEMoE K**	Nutrition	Microbiome	NEMoE with K latent class
MMMoE***	Microbiome	Microbiome	RMoE
NNMoE	Nutrition	Nutrition	RMoE
MNMoE	Microbiome	Nutrition	RMoE
Comb-MoE	Microbiome+Nutrition	Microbiome+Nutrition	RMoE

* Two-stage sparse logistic regression fitted with two, three and four latent classes were denoted as sLR II, sLR III and sLR IV.
** NEMoE fitted with two, three four latent classes were denoted as NEMoE II, NEMoE III and NEMoE IV. When not explicitly including the number of latent classes, we refer to NEMoE II.
*** NEMoE is easy to extend to partition the population with different types of data. We also investigate the different types of data as input of the NEMoE model. Results showed using nutrition to split the population obtained best performance in our dataset

4.3.3 Identification of informative taxonomic levels and consensus candidate microbial PD signatures in multiple independent cohorts

In our in-house gut microbiome PD investigation, NEMoE provided a natural criterion to examine which of the taxonomic levels (Phylum, Order, Family, Genus, and ASV) was most informative with respect to different nutrient intakes. We achieved this by evaluating the predictive performance for PD at each taxonomic level to determine the most informative. Figure 4.6 shows that genus was most predictive compared to the other taxonomic levels, with an LOOCV-AUC of 0.78.

Next, our NEMoE model determined a separate set of PD microbial signatures for each nutritional-ecotype. The derived coefficients represent the level of association between microbiota and health/disease state in each nutritional-ecotype shown in Figure 4.10 (a) and (b). Results for all taxa are given in Supplementary Data 1. We can broadly group the microbiota taxa into five categories based on their coefficient estimates: (i) significant in both

classes with different directions; (ii) significant in both classes with the same direction; (iii) significant in prot-CARB only, (iv) significant in PROT-carb only and (v) not-significant in both classes. The first category "significant in both classes with different directions" represents consistent abundance changes in both nutritional-ecotypes shown in Figure 4.10 (b). It was noted that the genera *Fusicatenibacter* and *Blautia* showed consistent negative coefficients in both PROT-carb and prot-CARB nutritional-ecotypes. Such genera may be considered stable PD microbial signatures, with several studies showing their underrepresentation in PD [3, 49, 99, 101, 131, 164].

The underrepresentation of *Fusicatenibacter* and *Blautia* were further validated using data from eight independent PD microbiome studies (Table 4.2). We processed the publicly available datasets using the dada2 pipeline [17](v1.16) and taxonomy reference "silva 138" [125, 188]. The relative abundance changes of the genus *Fusicatenibacter* were examined across all datasets, as shown in Figure 4.10 (c). In all but one dataset [65], *Fusicatenibacter* had significantly lower relative abundance among PD individuals. Similar results were observed for *Blautia* (see Figure 4.11), verifying NEMoE's ability to identify consensus microbial signatures of PD in multiple independent cohorts.

4.3.4 Identification of the microbiome that are differentially represented in specific nutritional classes

We note that taxa categories (i)-(iii) represent differential abundance changes that are unique in the two nutritional-ecotypes prot-CARB and PROT-carb, which indicate some microbial signatures of PD are diet-specific (Figure 4.10 c). We discovered that the genus *Escherichia-Shigella* was significantly underrepresented in the prot-CARB nutritional-ecotype but not in the PROT-carb ecotype. This genus belongs to the family *Enterobacteriaceae* (including E. *coli*, *Shigella*, *Salmonella*, *Klebsiella*, etc.), which are facultative anaerobes and known for utilising soluble sugars as a carbon source. When an individual's diet has a higher intake of sugars (or simple starch) it can be expected that the relative abundance of these microbiota will likely increase. Recent studies found that *Escherichia-Shigella* is a pathogenic bacteria that

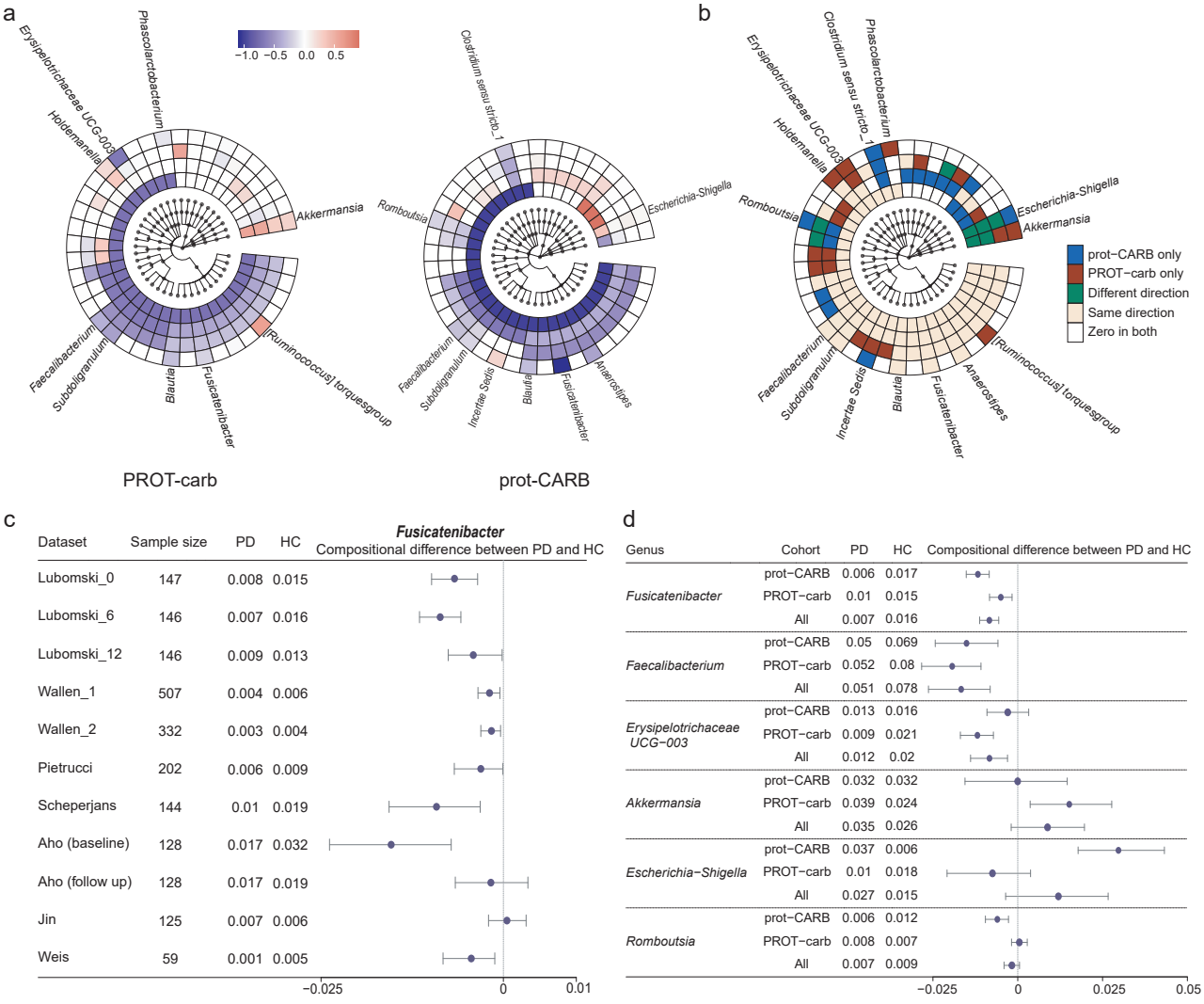

FIGURE 4.10. **Results of NEMoE on gut microbiome-PD study.** a) Coefficients of the experts network in NEMoE at different taxonomic levels. The two latent classes showed distinctly different microbiome patterns. b) Identification of diet specific microbial signatures of PD. The "Same direction" class showed consistent function in different dietary patterns. The "PROT-carb only" and "prot-CARB only" classes tended to be important only with specific dietary intake. The "Different direction" class changed their coefficients in different dietary patterns. c) Validation of differential relative abundance of genus *Fusicatenibacter* in 11 different datasets. With the exception of one dataset [65] all other datasets showed decreasing Fusicatenibacter in PD. d) Forest plot of 95% confidence intervals of selected taxa showed NEMoE is able to identify the species that are differentially represented in specific nutritional-ecotypes.

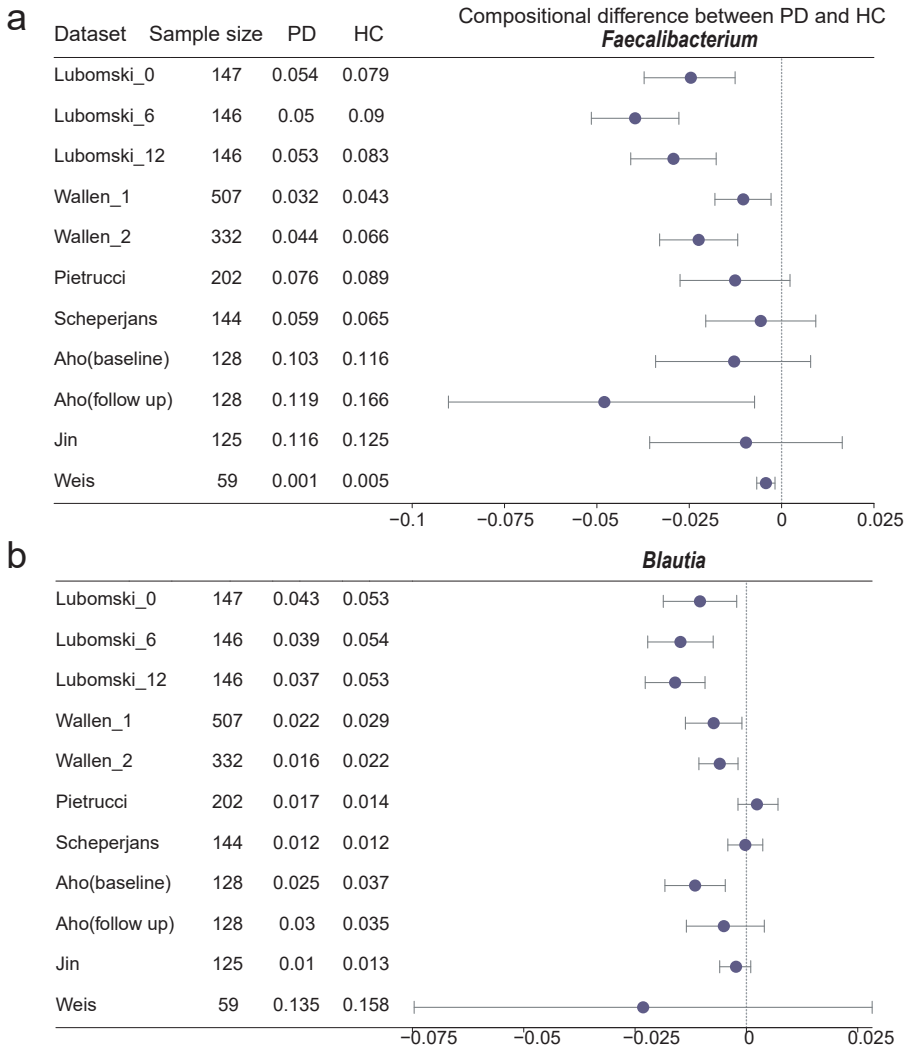

FIGURE 4.11. **External validation of consensus taxa Faecalibacterium and Blautia.** a) Validation of differential relative abundance of genus Faecalibacterium in eight different datasets. b) Forest plot showing the validation of differential relative abundance of genus Blautia in eight different datasets.

potentially reduces short-chain fatty acid production and produces endotoxins and neurotoxins [1, 69].

We also found a significant increase in the relative abundance of the genus Akkermansia, but only in the PROT-carb class (see Figure 4.10 d). These bacteria are known to impact immune response and constipation, with many studies reporting an overrepresentation in PD [16, 49, 101, 131]. *Akkermansia* breaks down mucins and turns them into short-chain fatty acids;

further, their relative abundance is thought to increase when "diet-specialise bacteria" decline as a direct impact of changes in microbially accessible carbohydrates (MAC). Generally, a low carbohydrate diet will lower MAC, thus lowering the number of diet-specialist microbes and allowing *Akkermansia* to become overrepresented, consistent with our discovery.

Most importantly, neither of these two genera (*Escherichia-Shigella*, *Akkermansia*) were discovered in our previous analysis using the ALDE model [41], where both classes were combined for microbiome biomarker identification (Escherichia-Shigella: p-value 0.14, Akkermansia: p-value 0.55) [99]. This highlights the relevance and importance of nutritional-ecotypes identification in microbiome marker discovery.

4.4 Discussion

The aim of this chapter is to investigate and unravel the complex interaction between diet, the microbiome and an individual's health. We achieve this by exploring the effects of dietary patterns (or composition) on the relationship between the microbiome and host health and by developing a method called NEMoE that detects such heterogeneity. Through a series of simulation studies, NEMoE shows strong prediction performance when the underlying data show heterogeneity explained by different nutrient intakes. Furthermore, we illustrate the practical performance of NEMoE on a gut microbiome PD study in which nutritional-ecotypes and microbial signatures of disease are found. We show that NEMoE outperforms the predictive accuracy of previous models (higher AUC) and identifies multiple known PD microbiome markers. Two different nutritional-ecotypes are also identified within our data with distinct protein-to-carbohydrate intake ratios and novel candidate signatures that were indicative of a diet-specific cohort.

While we focus on the discovery of microbial signatures of PD by splitting the population based on dietary profile, the architecture of NEMoE means its flexible algorithm can take different types of data for subcohort detection (data used for gating networks) or biomarker identification (data used for expert networks). Therefore, an alternate research question could be to identify nutrients as disease markers for diverse microbiome profiles, and the

NEMoE system can readily adapt to this new problem by changing the input of the gating network and experts network. Often, clinical knowledge or interest guides the decision on question formulation. However, if we consider both the dietary and microbiome profiles to be equivalent proxies for one's nutrition system, then performing NEMoE in two different ways allows us to empirically compare the effectiveness of nutritional signatures versus microbial signatures and provides us with insight into the natural heterogeneity in the microbiome and in nutritional intake.

NEMoE is designed to partition samples based on their associated nutrient intake and can be viewed as a data-driven strategy for subcohort or latent class identification. An alternative option is to investigate a knowledge-driven strategy to achieve the same goal and one example is the use of "enterotype". Similar to unsupervised learning, stratifying samples based on "enterotype" whilst providing an alternative way to stratify samples, does not explicitly take disease prediction performance into account. As a result, the aggregate predictive ability of the three separate enterotypes is lower than the nutritional-ecotypes division discovered by the NEMoE approach.

The proposed NEMoE method is based on diet-microbiome-host health interaction. However, it is not restricted to diet and microbiome data. Our method can be expanded to other multi-omics studies to identify subcohorts determined by the heterogeneity in relationships between covariates and response. One potential application is in the clinical heterogeneity of the relationship between multi-omics and host health. In such scenarios, the subcohorts are determined by their clinical index while the omics data are used to model the relationship between host health and information from a specific molecular platform.

In summary, we present NEMoE, a novel statistical method to model the heterogeneity of diet and the gut microbiome in disease. NEMoE identifies nutritional-ecotypes based on a maximum likelihood framework and uses an Expectation-Maximisation step to estimate the model parameters. Our proposed framework also enables identification and then accounts for multiple levels of structure in the feature set, a unique characteristic in microbiome data, where we are able to estimate a shared latent class for each individual at different taxonomic levels. The effectiveness of NEMoE is validated at three levels. First, we demonstrate through

a series of extensive simulation studies the model's ability to accurately identify latent classes and to increase microbiome predictability. Second, we validate the performance of NEMoE on a real disease dataset and show that this method outperforms existing two-stage methods. Finally, the downstream impact and practical importance of NEMoE is further demonstrated by the discovery of diet-specific PD microbiome markers, such as *Escherichia-Shigella* and *Akkermansia*, which are not identified by the ALDE model [41].

CHAPTER 5

Conclusion

Interrelationships among nutrients, omics features, microbiome and physiological outcomes are complex [8]. Diet could directly affect the health outcomes [66] or induce the changing of omics features, such as altering the expression and control of genes [15] or composition and function of the microbiome [142], further influencing outcomes. Conversely, diet-induced physiological changes in the host could impact the profile of omics features, for example, inflammatory bowel disease risk failure to sense beneficial commensal microbes [142]. In addition, several studies have shown evidence of causal relationships among these factors [8]. For instance, experiments have demonstrated that the gut microbiome is causally linked to differential effects of dietary fiber on host metabolic phenotypes [134], and animal studies have supported causal roles of host genotype and gut microbiome in the development of type 2 diabetes (T2D), insulin resistance, and obesity [80]. Uncovering such complex interactions in nutriomics data necessitate addressing many statistical and data science challenges including non-linearity, high-dimensionality, and heterogeneity. In this thesis, we developped novel statistical methods to deal with these challenges.

In Chapter 2, we focused on the discovery of the non-linear associations between nutrition and omics features. We proposed a novel method called LC-N2G to quantify the relationship between a combination of nutrients intake and a given omics feature. The LC-N2G method helps to select combinations of nutrition variables that relate to a gene of interest as well as test the significance of their associations in a non-linear setting. LC-N2G is available at `https://github.com/SydneyBioX/LCN2G` as an R package.

In Chapter 3, we further provided a method to group the high-dimensional omics features based on their response to nutrition intake as well as other experimental factors. We

presented a novel method, called eNODAL, to identify clusters of omics features in experimental nutriomics data. The eNODAL approach uses a two-stage clustering method to classify proteomics features into several interpretable subclusters and helps to illustrate how nutrition and other factors shape proteomics profiles. eNODAL is available at `https://github.com/SydneyBioX/eNODAL` as an R package.

In Chapter 4, we addressed the heterogeneity in observational studies in the context of diet-microbiome-host health interaction. A novel statistical method, NEMoE, has been proposed to divide the population into subcohorts that have a similar diet profile and model relationship between microbiome and host health within each subcohort. NEMoE also incorporates the characteristic of microbiome, that is the hierarchical taxonomic table, and provides insights into diet-specific microbial signatures for a disease. NEMoE is available at `https://sydneybiox.github.io/NEMoE` as an R package.

The three challenges we addressed in this thesis arise in nutriomics studies. However, the methods we developed are also applicable in a broader context in precision medicine. For example, LC-N2G in Chapter 2 is also able to non-linear associations between two different omics platforms such as genomics and proteomics; eNODAL provides a way to cluster high-dimensional omics features with response incorporate higher order interaction effect between drug treatment; and NEMoE, presented in Chapter 4, could also be applied when analysing the heterogeneous relationship based on other clinical variables such as age, sex and BMI.

As an emerging interdisciplinary field of study, nutriomics integrates techniques, knowledge, and methods from nutrition science, bioinformatics, statistics and machine learning. With the continual advancement of technology in all of these fields as well as the ongoing explosion of data, we expect several future extensions to be developed based on this thesis.

We briefly present three such possible future works from different discipline perspectives:

(1) From a nutritional perspective, in this thesis we focused on the nutrient intake to represent the nutrition information. However, several studies indicate that food quality and diversity are also essential to human health [161]. Integration of current

nutriomics methods into the food knowledge graph [56] might shed new light on a balanced diet.

(2) In this thesis, we focused on learning from the nutrition data and from one type of omics features. From a bioinformatics perspective, extending our developed methods to suit nutriomics studies with multi-omics data modality could reveal the relationships between nutrition and health more systematically and holistically. Based on this thesis, we could further consider the integration between nutrition and multi-omics to design a more comprehensive personalised nutrition strategy.

(3) More complex statistical models also hold promise to systematically elucidate the uncertainty in nutriomics studies. As we pointed out in Chapter 1, in this thesis, we did not account for the measurement error in the nutrition data matrix. Several statistical methods such as the NCI method [155] or PC-SIDE [18, 192], to just name two, could be incorporated to improve the model accuracy in nutriomics studies.

(4) Methods developed in this thesis are mainly based on observed correlations among nutrients, omics features, microbiome and physiological outcomes. However, the causal relationships between these factors are not yet fully understood [40]. Several studies have used Mendelian randomization [96] and mediation analysis [143] to investigate such causality. By incorporating such causal inference methods, it may be possible to gain a deeper understanding of the intricate interplay among these factors and shed light on the underlying mechanism that impacts our health.

A deeper understanding of the complex interplay among nutrients, omics features and health outcomes from nutriomics studies provides opportunities for precision medicine and personalised nutrition [20]. Towards this end, we developed methods and tools that contribute to the increasing demand of statistical methodology that serve nutrition science as well as related biology. It could be expected that based on these methods and tools, future work will give new insight into how nutrition shapes our health and further guides a healthy life.

Bibliography

[1] Velma TE Aho et al. 'Relationships of gut microbiota, short-chain fatty acids, inflammation, and the gut barrier in Parkinson's disease'. In: *Molecular Neurodegeneration* 16.1 (2021), pp. 1–14 (cit. on pp. 86, 87, 100).

[2] Goiuri Alberdi et al. 'Thermogenesis is involved in the body-fat lowering effects of resveratrol in rats'. In: *Food Chemistry* 141.2 (2013), pp. 1530–1535 (cit. on p. 50).

[3] Keshavarzian Ali et al. 'Colonic bacterial composition in Parkinson's disease'. In: *Movement Disorders* 30.10 (2015), pp. 1351–1360 (cit. on p. 98).

[4] Akram Alyass, Michelle Turcotte and David Meyre. 'From big data analysis to personalized medicine for all: challenges and opportunities'. In: *BMC Medical Genomics* 8.1 (2015), pp. 1–12 (cit. on p. 35).

[5] Thomas R Anderson et al. 'Geometric stoichiometry: unifying concepts of animal nutrition to understand how protein-rich diets can be "too much of a good thing"'. In: *Frontiers in Ecology and Evolution* 8 (2020), p. 196 (cit. on p. 15).

[6] Diana M André et al. 'High-fat diet-induced obesity impairs insulin signaling in lungs of allergen-challenged mice: Improvement by resveratrol'. In: *Scientific Reports* 7.1 (2017), pp. 1–12 (cit. on p. 58).

[7] Y N Andrew, I J Michael and W Yair. 'On spectral clustering: analysis and an algorithm'. In: *NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Ed. by T G Dietterich, S Becker and Z Ghahramani. MIT Press Cambridge, MA, USA. 2001, pp. 849–856 (cit. on p. 32).

[8] Anissa M Armet et al. 'Rethinking healthy eating in light of the gut microbiome'. In: *Cell Host & Microbe* 30.6 (2022), pp. 764–785 (cit. on p. 104).

[9] Manimozhiyan Arumugam et al. ‘Enterotypes of the human gut microbiome’. In: *Nature* 473.7346 (2011), pp. 174–180 (cit. on p. 95).

[10] Francesco Asnicar et al. ‘Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals’. In: *Nature Medicine* 27.2 (2021), pp. 321–332 (cit. on p. 69).

[11] B Ason et al. ‘Apobsirna-induced liver steatosis is resistant to clearance by the loss of fatty acid transport protein 5 (fatp5)’. In: *Lipids* 46 (2011), pp. 991–1003 (cit. on p. 31).

[12] Jean-Patrick Baudry et al. ‘Combining mixture components for clustering’. In: *Journal of Computational and Graphical Statistics* 19.2 (2010), pp. 332–353 (cit. on p. 81).

[13] Christophe Biernacki, Gilles Celeux and Gérard Govaert. ‘Assessing a mixture model for clustering with the integrated completed likelihood’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7 (2000), pp. 719–725 (cit. on p. 81).

[14] Ulrich Bodenhofer, Andreas Kothmeier and Sepp Hochreiter. ‘APCluster: an R package for affinity propagation clustering’. In: *Bioinformatics* 27.17 (2011), pp. 2463–2464 (cit. on p. 43).

[15] Annie Bouchard-Mercier et al. ‘Associations between dietary patterns and gene expression profiles of healthy men and women: a cross-sectional study’. In: *Nutrition Journal* 12.1 (2013), pp. 1–13 (cit. on p. 104).

[16] Clara Bullich et al. ‘Gut vibes in Parkinson’s disease: the microbiota-gut-brain Axis’. In: *Movement Disorders Clinical Practice* 6.8 (2019), pp. 639–651 (cit. on p. 100).

[17] Benjamin J Callahan et al. ‘DADA2: High-resolution sample inference from Illumina amplicon data’. In: *Nature Methods* 13.7 (2016), pp. 581–583 (cit. on pp. 86, 98).

[18] Alicia L Carriquiry. ‘Understanding and assessing nutrition’. In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 123–146 (cit. on pp. 4, 106).

[19] Faicel Chamroukhi and Bao-Tuyen Huynh. ‘Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models’. In: *Journal de la Société Française de Statistique* 160.1 (2019), pp. 57–85 (cit. on pp. 73, 80, 81).

[20] Nisha Chaudhary et al. ‘Personalized nutrition and-omics’. In: *Comprehensive Foodomics* (2021), pp. 495–507 (cit. on p. 106).

[21] Marie Chavent et al. ‘ClustOfVar: An R Package for the Clustering of Variables’. In: *Journal of Statistical Software* 50.13 (2012), pp. 1–16 (cit. on pp. 10, 35, 37).

[22] Congcong Chen et al. ‘Pex11a deficiency causes dyslipidaemia and obesity in mice’. In: *Journal of Cellular and Molecular Medicine* 23.3 (2019), pp. 2020–2031 (cit. on p. 62).

[23] Jun Chen and Hongzhe Li. ‘Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis’. In: *The Annals of Applied Statistics* 7 (2013), pp. 418–442 (cit. on pp. 82, 83).

[24] Jun Chen et al. ‘Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis’. In: *Biostatistics* 14.2 (2013), pp. 244–258 (cit. on p. 5).

[25] Judy H Cho and Clara Abraham. ‘Inflammatory bowel disease genetics: Nod2’. In: *Annual Review of Medicine* 58 (2007), pp. 401–416 (cit. on p. 68).

[26] M Clément et al. ‘Necrotic cell sensor clec4e promotes a pro-atherogenic macrophage phenotype through activation of the unfolded protein response.’ In: *Circulation* 134 (2016), pp. 1039–1051 (cit. on p. 31).

[27] Ricki J Colman et al. ‘Caloric restriction delays disease onset and mortality in rhesus monkeys’. In: *Science* 325.5937 (2009), pp. 201–204 (cit. on p. 15).

[28] International Human Genome Sequencing Consortium. ‘Finishing the euchromatic sequence of the human genome’. In: *Nature* 431 (2004), pp. 931–945 (cit. on p. 16).

[29] Paul I Costea et al. ‘Enterotypes in the landscape of gut microbial community composition’. In: *Nature Microbiology* 3.1 (2018), pp. 8–16 (cit. on p. 95).

[30] Sheena C Cotter et al. ‘Macronutrient balance mediates trade-offs between immune function and life history traits’. In: *Functional Ecology* 25.1 (2011), pp. 186–198 (cit. on p. 17).

[31] Sheena C Cotter et al. ‘Diet modulates the relationship between immune gene expression and functional immune responses’. In: *Insect Biochemistry and Molecular Biology* 109 (2019), pp. 128–141 (cit. on pp. 17, 28).

[32] Peter Cronin et al. ‘Dietary fibre modulates the gut microbiota’. In: *Nutrients* 13.5 (2021), pp. 1–22 (cit. on p. 68).

[33] Lawrence A David et al. ‘Diet rapidly and reproducibly alters the human gut microbiome’. In: *Nature* 505.7484 (2014), pp. 559–563 (cit. on p. 69).

[34] Cindy D Davis and John A Milner. ‘Nutrigenomics, vitamin D and cancer prevention’. In: *Lifestyle Genomics* 4.1 (2011), pp. 1–11 (cit. on p. 16).

[35] Francesca De Filippis et al. ‘High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome’. In: *Gut* 65.11 (2016), pp. 1812–1821 (cit. on p. 68).

[36] Rebecca A Deek and Hongzhe Li. ‘A zero-inflated latent dirichlet allocation model for microbiome studies’. In: *Frontiers in Genetics* 11 (2021), pp. 1–10 (cit. on pp. 82, 83).

[37] Mei Dong et al. ‘Predictive analysis methods for human microbiome data with application to Parkinson’s disease’. In: *PLoS One* 15.8 (2020), e0237779 (cit. on pp. 84, 88).

[38] Sarah Downer et al. ‘Food is medicine: Actions to integrate food and nutrition into healthcare’. In: *the BMJ* 369 (2020) (cit. on p. 36).

[39] Alejo Efeyan, William C Comb and David M Sabatini. ‘Nutrient-sensing mechanisms and pathways’. In: *Nature* 517.7534 (2015), pp. 302–310 (cit. on p. 16).

[40] Jane F Ferguson et al. ‘Nutrigenomics, the microbiome, and gene-environment interactions: new directions in cardiovascular disease research, prevention, and treatment: a scientific statement from the American Heart Association’. In: *Circulation: Cardiovascular Genetics* 9.3 (2016), pp. 291–313 (cit. on p. 106).

[41] Andrew D Fernandes et al. ‘Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis’. In: *Microbiome* 2.1 (2014), pp. 1–13 (cit. on pp. 101, 103).

[42] Chris Fraley and Adrian E Raftery. ‘How many clusters? Which clustering method? Answers via model-based cluster analysis’. In: *The Computer Journal* 41.8 (1998), pp. 578–588 (cit. on p. 81).

[43] Brendan J Frey and Delbert Dueck. ‘Clustering by passing messages between data points’. In: *Science* 315.5814 (2007), pp. 972–976 (cit. on p. 43).

[44] Sylvia Fruhwirth-Schnatter, Gilles Celeux and Christian P Robert. *Handbook of Mixture Analysis*. CRC Press, 2019 (cit. on pp. 11, 74, 75).

[45] Lizhi Fu et al. 'Interaction between metformin and leucine in reducing hyperlipidemia and hepatic lipid accumulation in diet-induced obese mice'. In: *Metabolism* 64.11 (2015), pp. 1426–1434 (cit. on p. 52).

[46] Atsushi Fukushima. 'DiffCorr: an R package to analyze and visualize differential correlations in biological networks'. In: *Gene* 518.1 (2013), pp. 209–214 (cit. on p. 45).

[47] Marta Garaulet and Juan Antonio Madrid. 'Chronobiological aspects of nutrition, metabolic syndrome and obesity'. In: *Advanced Drug Delivery Reviews* 62.9-10 (2010), pp. 967–978 (cit. on p. 36).

[48] Fengxia Ge et al. 'Insulin-and leptin-regulated fatty acid uptake plays a key causal role in hepatic steatosis in mice with intact leptin signaling but not in ob/ob or db/db mice'. In: *American Journal of Physiology-Gastrointestinal and Liver Physiology* 299.4 (2010), G855–G866 (cit. on p. 31).

[49] Sara Gerhardt and M Hasan Mohajeri. 'Changes of colonic bacterial composition in Parkinson's disease and other neurodegenerative diseases'. In: *Nutrients* 10.6 (2018), pp. 1–23 (cit. on pp. 98, 100).

[50] Dirk Gevers et al. 'The Human Microbiome Project: a community resource for the healthy human microbiome'. In: *PLoS Biology* 10.8 (2012), e1001377 (cit. on p. 1).

[51] David E Golberg. *Genetic Algorithm in Search, Optimization, and Machine Learning*. Addion Wesley, 1989 (cit. on p. 21).

[52] J Goldberger et al. 'Neighbourhood components analysis'. In: *NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems*. Ed. by L K Saul, Y Weiss and L Bottou. MIT Press Cambridge, MA, USA. 2004, pp. 513–520 (cit. on p. 32).

[53] Richard C Grandison, Matthew DW Piper and Linda Partridge. 'Amino-acid imbalance explains extension of lifespan by dietary restriction in Drosophila'. In: *Nature* 462.7276 (2009), pp. 1061–1064 (cit. on p. 15).

[54] Michael Greenacre. ‘Compositional data analysis’. In: *Annual Review of Statistics and its Application* 8.1 (2021), pp. 271–299 (cit. on p. 47).

[55] David R Hardoon, Sandor Szedmak and John Shawe-Taylor. ‘Canonical correlation analysis: An overview with application to learning methods’. In: *Neural computation* 16.12 (2004), pp. 2639–2664 (cit. on p. 32).

[56] Steven Haussmann et al. ‘FoodKG: a semantics-driven knowledge graph for food recommendation’. In: *International Semantic Web Conference*. Springer. 2019, pp. 146–162 (cit. on p. 106).

[57] Tobias Hegelmaier et al. ‘Interventional influence of the intestinal microbiome through dietary intervention and bowel cleansing might improve motor symptoms in Parkinson’s disease’. In: *Cells* 9.2 (2020), pp. 1–19 (cit. on p. 69).

[58] Erin M Hill-Burns et al. ‘Parkinson’s disease and Parkinson’s disease medications have distinct signatures of the gut microbiome’. In: *Movement Disorders* 32.5 (2017), pp. 739–749 (cit. on p. 86).

[59] Andrew J Holmes et al. ‘Diet-microbiome interactions in health are controlled by intestinal nitrogen source constraints’. In: *Cell Metabolism* 25.1 (2017), pp. 140–151 (cit. on pp. 11, 68, 69).

[60] Alexander J Hose et al. ‘Excessive unbalanced meat consumption in the first year of life increases asthma risk in the PASTURE and LUKAS2 birth cohorts’. In: *Frontiers in Immunology* 12 (2021) (cit. on p. 69).

[61] Riley L Hughes et al. ‘The role of the gut microbiome in predicting response to diet and the development of precision nutrition models. Part II: results’. In: *Advances in Nutrition* 10.6 (2019), pp. 979–998 (cit. on p. 69).

[62] David R Hunter and Kenneth Lange. ‘A tutorial on MM algorithms’. In: *The American Statistician* 58.1 (2004), pp. 30–37 (cit. on p. 78).

[63] Franziska Jannasch et al. ‘Exploratory dietary patterns: a systematic review of methods applied in pan-European studies and of validation studies’. In: *British Journal of Nutrition* 120.6 (2018), pp. 601–611 (cit. on p. 69).

[64] Susan J de Jersey et al. ‘An observational study of nutrition and physical activity behaviours, knowledge, and advice in pregnancy’. In: *BMC Pregnancy and Childbirth* 13.1 (2013), pp. 1–8 (cit. on p. 7).

[65] Miao Jin et al. ‘Analysis of the gut microflora in patients with Parkinson’s disease’. In: *Frontiers in Neuroscience* 13 (2019), pp. 1–19 (cit. on pp. 86, 98, 99).

[66] Jessica L Johnston, Jessica C Fanzo and Bruce Cogill. ‘Understanding sustainable diets: a descriptive analysis of the determinants and processes that influence diets and their impact on health, food security, and environmental sustainability’. In: *Advances in Nutrition* 5.4 (2014), pp. 418–429 (cit. on p. 104).

[67] Michael I Jordan and RI Jacobs. ‘Supervised learning and divide-and-conquer: A statistical approach’. In: *Proceedings of the Tenth International Conference on Machine Learning*. 2014, pp. 159–166 (cit. on p. 11).

[68] Minoru Kanehisa and Susumu Goto. ‘KEGG: kyoto encyclopedia of genes and genomes’. In: *Nucleic Acids Research* 28.1 (2000), pp. 27–30 (cit. on pp. 40, 46).

[69] Yongbo Kang et al. ‘Gut Microbiota and Parkinson’s Disease: Implications for Faecal Microbiota Transplantation Therapy’. In: *ASN Neuro* 13 (2021), pp. 1–10 (cit. on p. 100).

[70] Iara Karise et al. ‘Metformin enhances mitochondrial biogenesis and thermogenesis in brown adipocytes of mice’. In: *Biomedicine & Pharmacotherapy* 111 (2019), pp. 1156–1165 (cit. on p. 50).

[71] Hisanori Kato, Shoko Takahashi and Kenji Saito. ‘Omics and integrated omics for the promotion of food and nutrition science’. In: *Journal of Traditional and Complementary Medicine* 1.1 (2011), pp. 25–30 (cit. on p. 1).

[72] Christine Keribin. ‘Consistent estimation of the order of mixture models’. In: *Sankhyā: The Indian Journal of Statistics, Series A* (2000), pp. 49–66 (cit. on p. 81).

[73] Abbas Khalili. ‘New estimation and feature selection methods in mixture-of-experts models’. In: *Canadian Journal of Statistics* 38.4 (2010), pp. 519–539 (cit. on p. 81).

[74] Abbas Khalili and Jiahua Chen. ‘Variable selection in finite mixture of regression models’. In: *Journal of the American Statistical Association* 102.479 (2007), pp. 1025–1038 (cit. on pp. 80, 81).

[75] Leila Khorraminezhad et al. ‘Statistical and machine-learning analyses in nutritional genomics studies’. In: *Nutrients* 12.10 (2020), pp. 1–20 (cit. on p. 2).

[76] Kook Hwan Kim and Myung-Shik Lee. ‘FGF21 as a stress hormone: the roles of FGF21 in stress adaptation and the treatment of metabolic diseases’. In: *Diabetes & Metabolism Journal* 38.4 (2014), pp. 245–251 (cit. on p. 9).

[77] Vladimir Yu Kiselev et al. ‘SC3: consensus clustering of single-cell RNA-seq data’. In: *Nature Methods* 14.5 (2017), pp. 483–486 (cit. on pp. 39, 40, 43, 47).

[78] Robert A Koeth et al. ‘Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis’. In: *Nature Medicine* 19.5 (2013), pp. 576–585 (cit. on p. 68).

[79] Ling Chun Kong et al. ‘Dietary patterns differently associate with inflammation and gut microbiota in overweight and obese subjects’. In: *PloS one* 9.10 (2014), e109434 (cit. on p. 33).

[80] Julia H Kreznar et al. ‘Host genotype and gut microbiome modulate insulin secretion and diet-induced metabolic phenotypes’. In: *Cell reports* 18.7 (2017), pp. 1739–1750 (cit. on p. 104).

[81] Eric S Lander. ‘Initial impact of the sequencing of the human genome’. In: *Nature* 470.7333 (2011), pp. 187–197 (cit. on p. 16).

[82] Peter Langfelder and Steve Horvath. ‘WGCNA: an R package for weighted correlation network analysis’. In: *BMC Bioinformatics* 9.1 (2008), pp. 1–13 (cit. on pp. 10, 37, 40, 54).

[83] Peter Langfelder, Bin Zhang and Steve Horvath. ‘Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R’. In: *Bioinformatics* 24.5 (2008), pp. 719–720 (cit. on p. 43).

[84] N Latruffe and J Vamecq. ‘Peroxisome proliferators and peroxisome proliferator activated receptors (PPARs) as regulators of lipid metabolism’. In: *Biochimie* 79.2-3 (1997), pp. 81–94 (cit. on p. 51).

[85] Kim-Anh Lê Cao, Emmanuelle Meugnier and Geoffrey J McLachlan. ‘Integrative mixture of experts to combine clinical factors and gene markers’. In: *Bioinformatics* 26.9 (2010), pp. 1192–1198 (cit. on pp. 71, 72).

[86] David G Le Couteur et al. ‘The impact of low-protein high-carbohydrate diets on aging and lifespan’. In: *Cellular and Molecular Life Sciences* 73.6 (2016), pp. 1237–1252 (cit. on p. 23).

[87] David G Le Couteur et al. ‘Nutritional reprogramming of mouse liver proteome is dampened by metformin, resveratrol, and rapamycin’. In: *Cell Metabolism* 33.12 (2021), pp. 2367–2379 (cit. on pp. 6, 8, 11, 13, 35–37, 39, 50, 51, 57, 61).

[88] Kwang Pum Lee et al. ‘Lifespan and reproduction in Drosophila: new insights from nutritional geometry’. In: *Proceedings of the National Academy of Sciences* 105.7 (2008), pp. 2498–2503 (cit. on p. 17).

[89] François Leulier et al. ‘Integrative physiology: at the crossroads of nutrition, microbiota, animal physiology, and human health’. In: *Cell Metabolism* 25.3 (2017), pp. 522–534 (cit. on p. 17).

[90] Hongzhe Li. ‘Microbiome, metagenomics, and high-dimensional compositional data analysis’. In: *Annual Review of Statistics and Its Application* 2 (2015), pp. 73–94 (cit. on pp. 1, 68).

[91] Shijie Li, Michael S Brown and Joseph L Goldstein. ‘Bifurcation of insulin signaling pathway in rat liver: mTORC1 required for stimulation of lipogenesis, but not inhibition of gluconeogenesis’. In: *Proceedings of the National Academy of Sciences* 107.8 (2010), pp. 3441–3446 (cit. on p. 58).

[92] Wen Li. ‘The era of nanotechnology and omics sciences’. In: *European Journal of BioMedical Research* 1.1 (2017), pp. 1–2 (cit. on p. 1).

[93] Xiaoling Li et al. ‘PEX11α is required for peroxisome proliferation in response to 4-phenylbutyrate but is dispensable for peroxisome proliferator-activated receptor alpha-mediated peroxisome proliferation’. In: *Molecular and Cellular Biology* 22.23 (2002), pp. 8226–8240 (cit. on p. 62).

[94] Xunhan Li, Muthukumaran Jayachandran and Baojun Xu. ‘Antidiabetic effect of konjac glucomannan via insulin signaling pathway regulation in high-fat diet and streptozotocin-induced diabetic rats’. In: *Food Research International* 149 (2021), pp. 1–9 (cit. on p. 58).

[95] Dachao Liang et al. ‘Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities’. In: *Gut Pathogens* 10.1 (2018), pp. 1–9 (cit. on p. 69).

[96] Xiaomin Liu et al. ‘Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome’. In: *Nature genetics* 54.1 (2022), pp. 52–61 (cit. on p. 106).

[97] Zhigang Liu et al. ‘High-fat diet induces hepatic insulin resistance and impairment of synaptic plasticity’. In: *PloS One* 10.5 (2015), e0128274 (cit. on p. 58).

[98] Michal Lubomski et al. ‘Parkinson’s disease and the gastrointestinal microbiome’. In: *Journal of Neurology* 267.9 (2020), pp. 2507–2523 (cit. on pp. 12, 68).

[99] Michal Lubomski et al. ‘Nutritional intake and gut microbiome composition predict Parkinson’s disease’. In: *Frontiers in Aging Neuroscience* (2022), pp. 1–16 (cit. on pp. v, vi, 7, 8, 12, 13, 69, 86, 87, 89, 98, 101).

[100] Michal Lubomski et al. ‘The Gut Microbiome in Parkinson’s Disease: A Longitudinal Study of the Impacts on Disease Progression and the Use of Device-Assisted Therapies’. In: *Frontiers in Aging Neuroscience* (2022), pp. 1–24 (cit. on pp. v, vi, 7, 12, 69, 86, 87, 89, 91).

[101] Michal Lubomski et al. ‘The impact of device-assisted therapies on the gut microbiome in Parkinson’s disease’. In: *Journal of Neurology* 269.2 (2022), pp. 780–795 (cit. on pp. v, vi, 7, 12, 13, 87, 98, 100).

[102] Siyuan Ma et al. ‘Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease’. In: *bioRxiv* (2020), pp. 1–46 (cit. on p. 88).

[103] Jay Magidson, Jeroen K Vermunt and John P Madura. *Latent Class Analysis*. SAGE Publications Limited Thousand Oaks, 2020 (cit. on p. 11).

[104] Ashok Makkuva et al. ‘Learning in gated neural networks’. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3338–3348 (cit. on p. 74).

[105] Nathan Mantel. ‘The detection of disease clustering and a generalized regression approach’. In: *Cancer research* 27.2_Part_1 (1967), pp. 209–220 (cit. on p. 33).

[106] Kate A Marsh et al. ‘Effect of a low glycemic index compared with a conventional healthy diet on polycystic ovary syndrome’. In: *The American Journal of Clinical Nutrition* 92.1 (2010), pp. 83–92 (cit. on p. 52).

[107] Michael I McBurney et al. ‘Establishing what constitutes a healthy human gut microbiome: state of the science, regulatory considerations, and future directions’. In: *The Journal of Nutrition* 149.11 (2019), pp. 1882–1895 (cit. on pp. 67, 68).

[108] Geoffrey J McLachlan, Sharon X Lee and Suren I Rathnayake. ‘Finite mixture models’. In: *Annual Review of Statistics and its Application* 6 (2019), pp. 355–378 (cit. on p. 11).

[109] M Nathaniel Mead. ‘Nutrigenomics: the genome–food interface.’ In: *Environmental Health Perspectives* 115 (2007), pp. 583–589 (cit. on p. 16).

[110] Nicolai Meinshausen. ‘Hierarchical testing of variable importance’. In: *Biometrika* 95.2 (2008), pp. 265–278 (cit. on p. 43).

[111] Maria Chiara Mentella et al. ‘Nutrition, IBD and gut microbiota: a review’. In: *Nutrients* 12.4 (2020), p. 944 (cit. on p. 1).

[112] J Bernadette Moore. ‘From personalised nutrition to precision medicine: the rise of consumer genomics and digital health’. In: *Proceedings of the Nutrition Society* 79.3 (2020), pp. 300–310 (cit. on p. 35).

[113] Dariush Mozaffarian, Irwin Rosenberg and Ricardo Uauy. ‘History of modern nutrition science—implications for current research, dietary guidelines, and food policy’. In: *the BMJ* 361 (2018), pp. 1–6 (cit. on p. 36).

[114] Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012 (cit. on p. 10).

[115] Shinichi Nakagawa et al. ‘An assessment of statistical methods for nonindependent data in ecological meta-analyses: Comment’. In: *Ecology* 103.1 (2022), e03490 (cit. on p. 8).

[116] Michael A Newton and Zhishi Wang. ‘Multiset statistics for gene set analysis’. In: *Annual Review of Statistics and its Application* 2 (2015), pp. 95–111 (cit. on pp. 40, 48).

[117] Natalie C Palavra et al. ‘Increased added sugar consumption is common in Parkinson’s disease’. In: *Frontiers in Nutrition* 8 (2021), pp. 1–11 (cit. on p. 86).

[118] Victoria Pascal et al. ‘A microbial signature for Crohn’s disease’. In: *Gut* 66.5 (2017), pp. 813–822 (cit. on p. 68).

[119] Ellis Patrick et al. ‘A multi-step classifier addressing cohort heterogeneity improves performance of prognostic biomarkers in three cancer types’. In: *Oncotarget* 8.2 (2017), pp. 2807–2815 (cit. on p. 69).

[120] G Pei, L Chen and W Zhang. ‘WGCNA application to proteomic and metabolomic data analysis’. In: *Methods in Enzymology*. Vol. 585. Elsevier, 2017, pp. 135–158 (cit. on pp. 35, 37).

[121] Matthew DW Piper et al. ‘Dietary restriction and aging: a unifying perspective’. In: *Cell Metabolism* 14.2 (2011), pp. 154–160 (cit. on pp. 2, 15, 17).

[122] Claudia Plant and Christian Böhm. ‘Inconco: interpretable clustering of numerical and categorical objects’. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011, pp. 1127–1135 (cit. on p. 37).

[123] Gordon Qian and Joshua WK Ho. ‘Challenges and emerging systems biology approaches to discover how the human gut microbiome impact host physiology’. In: *Biophysical Reviews* 12.4 (2020), pp. 851–863 (cit. on p. 3).

[124] Junjie Qin et al. ‘A metagenome-wide association study of gut microbiota in type 2 diabetes’. In: *Nature* 490.7418 (2012), pp. 55–60 (cit. on p. 68).

[125] Christian Quast et al. ‘The SILVA ribosomal RNA gene database project: improved data processing and web-based tools’. In: *Nucleic Acids Research* 41.D1 (2012), pp. D590–D596 (cit. on pp. 87, 98).

[126] Xinyu Que et al. ‘Scalable community detection with the louvain algorithm’. In: *2015 IEEE International Parallel and Distributed Processing Symposium*. IEEE. 2015, pp. 28–37 (cit. on p. 43).

[127] Nimrod Rappoport and Ron Shamir. ‘Multi-omic and multi-view clustering algorithms: review and cancer benchmark’. In: *Nucleic Acids Research* 46.20 (2018), pp. 10546–10562 (cit. on p. 66).

[128] D Raubenheimer and S J Simpson. 'Nutritional ecology and human health.' In: *Annual Review of Nutrition* 36 (2016), pp. 603–626 (cit. on pp. 8, 9, 15–17, 88).

[129] Mark N Read and Andrew J Holmes. 'Towards an integrative understanding of diet–host–gut microbiome interactions'. In: *Frontiers in Immunology* 8 (2017), pp. 1–9 (cit. on p. 68).

[130] Xinmin Ren and Xiangdong Li. 'Advances in research on diabetes by human nutriomics'. In: *International Journal of Molecular Sciences* 20.21 (2019), pp. 1–15 (cit. on p. 1).

[131] Stefano Romano et al. 'Meta-analysis of the Parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation'. In: *NPJ Parkinson's Disease* 7.1 (2021), pp. 1–13 (cit. on pp. 98, 100).

[132] Gökhan Sadi et al. 'Resveratrol improves hepatic insulin signaling and reduces the inflammatory response in streptozotocin-induced diabetes'. In: *Gene* 570.2 (2015), pp. 213–220 (cit. on p. 58).

[133] Rodrigo San-Cristobal et al. 'Contribution of macronutrients to obesity: implications for precision nutrition'. In: *Nature Reviews Endocrinology* 16.6 (2020), pp. 305–320 (cit. on p. 50).

[134] Serena Sanna et al. 'Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases'. In: *Nature genetics* 51.4 (2019), pp. 600–605 (cit. on p. 104).

[135] Filip Scheperjans et al. 'Gut microbiota are related to Parkinson's disease and clinical phenotype'. In: *Movement Disorders* 30.3 (2015), pp. 350–358 (cit. on p. 86).

[136] Christina-Alexandra Schulz, Kolade Oluwagbemigun and Ute Nöthlings. 'Advances in dietary pattern analysis in nutritional epidemiology'. In: *European Journal of Nutrition* 60.8 (2021), pp. 4115–4130 (cit. on p. 69).

[137] Matthias B Schulze et al. 'Food based dietary patterns and chronic disease prevention'. In: *the BMJ* 361 (2018), pp. 1–6 (cit. on p. 69).

[138] Alistair M Senior et al. 'Global associations between macronutrient supply and age-specific mortality'. In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30824–30835 (cit. on pp. 8, 9, 36).

[139] Vandana Sharma et al. ‘Mannose alters gut microbiome, prevents diet-induced obesity, and improves host metabolism’. In: *Cell Reports* 24.12 (2018), pp. 3087–3098 (cit. on p. 62).

[140] Stephen J Simpson and David Raubenheimer. *The Nature of Nutrition: A Unifying Framework from Animal Adaptation to Human Obesity*. Princeton, NJ, USA: Princeton University Press, 2012 (cit. on pp. 8, 16, 17, 22).

[141] Stephen J Simpson et al. ‘The geometric framework for nutrition as a tool in precision medicine’. In: *Nutrition and Healthy Aging* 4.3 (2017), pp. 217–226 (cit. on pp. 2, 3, 8, 9, 15, 17, 22, 23, 35, 63, 88).

[142] Rasnik K Singh et al. ‘Influence of diet on the gut microbiome and implications for human health’. In: *Journal of Translational Medicine* 15.1 (2017), pp. 1–17 (cit. on pp. 67, 104).

[143] Michael B Sohn and Hongzhe Li. ‘Compositional mediation analysis for microbiome studies’. In: *The Annals of Applied Statistics* 13.1 (2019), pp. 661–681 (cit. on p. 106).

[144] Samantha M Solon-Biet et al. ‘Macronutrient balance, reproductive function, and lifespan in aging mice’. In: *Proceedings of the National Academy of Sciences* 112.11 (2015), pp. 3481–3486 (cit. on p. 23).

[145] Samantha M Solon-Biet et al. ‘Defining the nutritional and metabolic context of FGF21 using the geometric framework’. In: *Cell Metabolism* 24.4 (2016), pp. 555–565 (cit. on pp. 9, 13, 17, 22, 23, 28–30).

[146] Nicolas Städler, Peter Bühlmann and Sara Van De Geer. ‘1-penalization for mixture regression models’. In: *Test* 19.2 (2010), pp. 209–256 (cit. on pp. 80, 81).

[147] Harald Staiger et al. ‘Fibroblast growth factor 21—metabolic role in mice and men’. In: *Endocrine Reviews* 38.5 (2017), pp. 468–488 (cit. on p. 9).

[148] Alexander Strehl and Joydeep Ghosh. ‘Cluster ensembles—a knowledge reuse framework for combining multiple partitions’. In: *Journal of Machine Learning Research* 3 (2002), pp. 583–617 (cit. on p. 44).

[149] KP Suresh. ‘An overview of randomization techniques: an unbiased assessment of outcome in clinical research’. In: *Journal of Human Reproductive Sciences* 4.1 (2011), pp. 8–11 (cit. on p. 6).

[150] Ai Huey Tan et al. ‘Gut microbial ecosystem in Parkinson disease: new clinicobiological insights from multi-omics’. In: *Annals of Neurology* 89.3 (2021), pp. 546–559 (cit. on p. 69).

[151] Julien Tap et al. ‘Diet and gut microbiome interactions of relevance for symptoms in irritable bowel syndrome’. In: *Microbiome* 9.1 (2021), pp. 1–13 (cit. on pp. 67, 69).

[152] Linda C Tapsell et al. ‘Foods, nutrients, and dietary patterns: interconnections and implications for dietary guidelines’. In: *Advances in Nutrition* 7.3 (2016), pp. 445–454 (cit. on p. 36).

[153] Abdellah Tebani and Soumeya Bekri. ‘Paving the way to precision nutrition through metabolomics’. In: *Frontiers in Nutrition* 6 (2019), pp. 1–10 (cit. on p. 69).

[154] Devarajan Thangadurai et al. *Nutriomics: Well-being through Nutrition*. CRC Press, 2022 (cit. on pp. 1, 11, 15, 67).

[155] Janet A Tooze et al. ‘A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution’. In: *Journal of the American Dietetic Association* 106.10 (2006), pp. 1575–1587 (cit. on p. 106).

[156] Juan de Toro-Martın et al. ‘Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome’. In: *Nutrients* 9.8 (2017), pp. 1–28 (cit. on p. 36).

[157] Courtney M Townsend et al. *Sabiston Textbook of Surgery*. Elsevier Health Sciences, 2016 (cit. on p. 7).

[158] Cassie M Tran et al. ‘Rapamycin blocks induction of the thermogenic program in white adipose tissue’. In: *Diabetes* 65.4 (2016), pp. 927–941 (cit. on p. 50).

[159] Samantha C Tyner, François Briatte and Heike Hofmann. ‘Network Visualization with ggplot2’. In: *The R Journal* (2017), pp. 27–59 (cit. on p. 49).

[160] Kathy Vagianos et al. ‘Nutrition assessment of patients with inflammatory bowel disease’. In: *Journal of Parenteral and Enteral Nutrition* 31.4 (2007), pp. 311–319 (cit. on p. 1).

[161] Eric O Verger et al. ‘Dietary diversity indicators and their associations with dietary adequacy and health outcomes: A systematic scoping review’. In: *Advances in Nutrition* 12.5 (2021), pp. 1659–1672 (cit. on p. 105).

[162] Frank de Vocht et al. ‘Conceptualising natural and quasi experiments in public health’. In: *BMC Medical Research Methodology* 21.1 (2021), pp. 1–8 (cit. on p. 6).

[163] Ulrike Von Luxburg. ‘A tutorial on spectral clustering’. In: *Statistics and Computing* 17.4 (2007), pp. 395–416 (cit. on p. 32).

[164] Zachary D Wallen et al. ‘Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens’. In: *NPJ Parkinson’s Disease* 6.1 (2020), pp. 1–12 (cit. on pp. 86, 87, 98).

[165] Bo Wang et al. ‘Similarity network fusion for aggregating data types on a genomic scale’. In: *Nature Methods* 11.3 (2014), pp. 333–337 (cit. on p. 66).

[166] Tao Wang and Hongyu Zhao. ‘Constructing predictive microbial signatures at multiple taxonomic levels’. In: *Journal of the American Statistical Association* 112.519 (2017), pp. 1022–1031 (cit. on p. 85).

[167] Yi-Wei Wang et al. ‘Metformin: a review of its potential indications’. In: *Drug Design, Development and Therapy* 11 (2017), p. 2421 (cit. on p. 52).

[168] Y Watanabe et al. ‘Isoliquiritigenin Attenuates Adipose Tissue Inflammation in vitro and Adipose Tissue Fibrosis through Inhibition of Innate Immune Responses in Mice’. In: *Scientific Reports* 6 (2016), p. 23097 (cit. on p. 31).

[169] Patrick Webb, Hanqi Luo and Ugo Gentilini. ‘Measuring multiple facets of malnutrition simultaneously: the missing link in setting nutrition targets and policymaking’. In: *Food Security* 7.3 (2015), pp. 479–492 (cit. on p. 36).

[170] Roxanne Weiss et al. ‘Metformin reduces glucose intolerance caused by rapamycin treatment in genetically heterogeneous female mice’. In: *Aging* 10.3 (2018), pp. 386–401 (cit. on p. 61).

[171] Klaas R Westerterp. ‘Diet induced thermogenesis’. In: *Nutrition & Metabolism* 1.1 (2004), pp. 1–5 (cit. on p. 50).

[172] KR Westerterp, SAJ Wilson and V Rolland. ‘Diet induced thermogenesis measured over 24h in a respiration chamber: effect of diet composition’. In: *International Journal of Obesity* 23.3 (1999), pp. 287–292 (cit. on p. 50).

[173] Renger F Witkamp. ‘Nutrition to optimise human health—how to obtain physiological substantiation?’ In: *Nutrients* 13.7 (2021), p. 2155 (cit. on p. 5).

[174] Svante Wold, Michael Sjöström and Lennart Eriksson. 'PLS-regression: a basic tool of chemometrics'. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130 (cit. on p. 32).

[175] S N Wood. 'Thin Plate Regression Splines'. In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 65 (2003), pp. 95–114 (cit. on p. 8).

[176] Simon Wood and Maintainer Simon Wood. 'Package 'mgcv''. In: *R Package Version* 1.29 (2015), p. 729 (cit. on p. 43).

[177] Simon N Wood. 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1 (2011), pp. 3–36 (cit. on p. 8).

[178] Simon N Wood. 'On p-values for smooth components of an extended generalized additive model'. In: *Biometrika* 100.1 (2013), pp. 221–228 (cit. on p. 43).

[179] Simon N Wood, Natalya Pya and Benjamin Säfken. 'Smoothing parameter and model selection for general smooth models'. In: *Journal of the American Statistical Association* 111.516 (2016), pp. 1548–1563 (cit. on p. 8).

[180] Gary D Wu et al. 'Linking long-term dietary patterns with gut microbial enterotypes'. In: *Science* 334.6052 (2011), pp. 105–108 (cit. on pp. 67, 68, 95).

[181] Tianzhi Wu et al. 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data'. In: *The Innovation* 2.3 (2021), pp. 1–10 (cit. on p. 46).

[182] Yinglin Xia. 'Correlation and association analyses in microbiome study integrating multiomics in health and disease'. In: *Progress in Molecular Biology and Translational Science* 171 (2020), pp. 309–491 (cit. on pp. 1, 3).

[183] Yinglin Xia, Jun Sun, Ding-Geng Chen et al. *Statistical Analysis of Microbiome Data with R*. Vol. 847. Springer, 2018 (cit. on p. 5).

[184] Xiangnan Xu et al. 'LC-N2G: a local consistency approach for nutrigenomics data analysis'. In: *BMC Bioinformatics* 21.1 (2020), pp. 1–14 (cit. on pp. iii, 12, 13).

[185] Xiangnan Xu et al. 'NEMoE: A nutrition aware regularized mixture of experts model addressing diet-cohort heterogeneity of gut microbiota in Parkinson's disease'. In: *medRxiv* (2021), pp. 1–37 (cit. on pp. iv, 13).

[186] Zhenjiang Xu and Rob Knight. ‘Dietary effects on human gut microbiome diversity’. In: *British Journal of Nutrition* 113.S1 (2015), S1–S5 (cit. on pp. 67, 68).

[187] Shi-Bing Yang et al. ‘Rapamycin induces glucose intolerance in mice by reducing islet mass, insulin content, and insulin sensitivity’. In: *Journal of Molecular Medicine* 90.5 (2012), pp. 575–585 (cit. on p. 61).

[188] Pelin Yilmaz et al. ‘The SILVA and “all-species living tree project (LTP)” taxonomic frameworks’. In: *Nucleic Acids Research* 42.D1 (2014), pp. D643–D648 (cit. on pp. 87, 98).

[189] Danxia Yu et al. ‘Long-term diet quality is associated with gut microbiome diversity and composition among urban Chinese adults’. In: *The American Journal of Clinical Nutrition* 113.3 (2021), pp. 684–694 (cit. on p. 68).

[190] Seniha Esen Yuksel, Joseph N Wilson and Paul D Gader. ‘Twenty years of mixture of experts’. In: *IEEE Transactions on Neural Networks and Learning Systems* 23.8 (2012), pp. 1177–1193 (cit. on p. 72).

[191] David Zeevi et al. ‘Personalized nutrition by prediction of glycemic responses’. In: *Cell* 163.5 (2015), pp. 1079–1094 (cit. on pp. 11, 69).

[192] Steven H Zeisel and Kerry-Ann Da Costa. ‘Choline: an essential nutrient for public health’. In: *Nutrition Reviews* 67.11 (2009), pp. 615–623 (cit. on p. 106).

[193] Bin Zhang and Steve Horvath. ‘A general framework for weighted gene co-expression network analysis’. In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005), pp. 1–43 (cit. on pp. 24, 31).

[194] Baiyu Zhou and Wing Hung Wong. ‘A bootstrap-based non-parametric ANOVA method with applications to factorial microarray data’. In: *Statistica Sinica* (2011), pp. 495–514 (cit. on pp. 39, 40, 42).

[195] Hui Zou and Trevor Hastie. ‘Regularization and variable selection via the elastic net’. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320 (cit. on p. 73).

Acknowledgements

During my long journey toward my PhD, lots of support and involvement come from many amazing people, which makes my life colorful. Not only my knowledge and skills increased, but also the people I met, knew, and collaborated with became the most important treasure in my life.

First, I would like to say big thanks to my two wonderful supervisors. To Samuel, I thank you for your patience and great ideas during our supervision meeting. I learned a lot from you not only from your knowledge but also from how to manage and deal with the hard tasks in my life. During our meetings, you always told me to "slow down". Although I'm still not good at expressing myself, I still realize that I need to "think carefully and persuade myself first". To Jean, you provide me with a lot of brilliant ways of thinking. I still remember how you mentioned the word "big picture" when I just started my PhD. Your ability for collaboration, communication, and presentation is what I can spend a lifetime learning from you.

I would also like to thank all my excellent colleagues in the Sydney Precision Data Science Centre for every precious feedback during the Friday Lab meetings as well as our daily discussions: Adam, Andy, Anne, Carissa, Connor, Chunlei, Daniel, Dario, Di, Ellis, Garth, Hani, Hao, Harry, John, Kevin, Kitty, Mengbo, Peng, Pengyi, Sarah, Shila, Taiyun, Weichang, Yingxin, Yue and Yunwei. It's my pleasure to work with you wonderful guys. Special thanks to Peng for our interesting brainstorming during our lunchtime chats; and Peng, Mengbo and Andy for proofreading this thesis.

Thanks to all my insighful collaborators. For the nutriomics project, thanks to Dr. Alistair Senior, Dr. Samantha M. Solon-Biet, Prof. David Raubenheimer, Prof. Stephen J. Simpson and Prof. Luigi Fontan. For the microbiome project, thanks to Dr. Michal Lubomski, A/Prof. Andrew J. Holmes, Prof. Carolyn M. Sue, Dr. Ryan L. Davis.

Thanks to my friends in Carslaw, Chunlei, Mengfan and Yang. I really enjoyed our weekend trips and special thanks to Mengfan for your encouraging and patiently listening to my difficulties and providing your solutions.

Finally, I would like to thank my parents, my cousin and all my relatives. Thanks for you unconditional love.