

Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN  
NANOSCIENZE PER LA MEDICINA E PER L'AMBIENTE

Ciclo 34

**Settore Concorsuale:** 03/B1 - FONDAMENTI DELLE SCIENZE CHIMICHE E SISTEMI INORGANICI

**Settore Scientifico Disciplinare:** CHIM/03 - CHIMICA GENERALE E INORGANICA

A TAKE ON COMPLEXITY: BIO-MOLECULES AND HUMAN METABOLISM  
INTERACTION MODELLING FOR HEALTH AND NUTRITION WITH MACHINE  
LEARNING

**Presentata da:** Carlo Mengucci

**Coordinatore Dottorato**

Dario Braga

**Supervisore**

Francesco Capozzi

**Co-supervisore**

Daniel Remondini

**Esame finale anno 2022**



## Abstract

The advent of omic data production has opened many new perspectives in the quest for modelling complexity in biophysical systems. With the capability of characterizing a complex organism through the patterns of its molecular states, observed at different levels through various omics, a new paradigm of investigation is arising. In this thesis, we investigate the links between perturbations of the human organism, described as the ensemble of crosstalk of its molecular states, and health. Machine learning plays a key role within this picture, both in omic data analysis and model building. We propose and discuss different frameworks developed by the author using machine learning for data reduction, integration, projection on latent features, pattern analysis, classification and clustering of omic data, with a focus on  $^1\text{H}$  NMR metabolomic spectral data. The aim is to link different levels of omic observations of molecular states, from nanoscale to macroscale, to study perturbations such as diseases and diet interpreted as changes in molecular patterns. The first part of this work focuses on the fingerprinting of diseases, linking cellular and systemic metabolomics with genomic to assess and predict the downstream of perturbations all the way down to the enzymatic network. The second part is a set of frameworks and models, developed with  $^1\text{H}$  NMR metabolomic at its core, to study the exposure of the human organism to diet and food intake in its full complexity, from epidemiological data analysis to molecular characterization of food structure.





# Contents

<b>Contents</b>	<b>iii</b>
<b>1 From Nanoscale to Macroscale and Back: the Complexity of Compartmentalization</b>	<b>1</b>
1.1 Introduction : A Complex Picture . . . . .	1
1.2 Omics Sciences and Complexity: We Are Our Molecular State . . . . .	3
1.3 Tools of the Trade : Metabolomics and Machine Learning . . . . .	7
1.3.1 NMR Spectroscopy in a nutshell . . . . .	7
1.3.2 NMR-based metabolomics data analysis . . . . .	9
1.3.3 Classifiers . . . . .	15
<b>2 Fingerprinting Perturbations and Dynamics of Metabolic States</b>	<b>19</b>
2.1 Fingerprinting Perturbations : From Cells and Systems to Enzymatic Networks . . . . .	19
2.1.1 An introduction to fingerprinting in Acute Myeloid Leukemia . . . . .	19
2.1.2 Study design and methods summary . . . . .	20
2.1.3 Results . . . . .	21
2.1.4 Study conclusions . . . . .	39
2.2 Modelling Dynamics of Metabolic States . . . . .	40
2.2.1 Metabolomic evaluation of therapeutic response in breast cancer . . . . .	40
2.2.2 Early-stage results and discussion . . . . .	42
2.3 Chapter Conclusions . . . . .	47
<b>3 Macroscopic Exposures: Epidemiological Data Analysis and Physiological Outcomes</b>	<b>49</b>
3.1 Studying health, lifestyle and diet: a major challenge . . . . .	49
3.2 Correlations between inadequate Energy/Macronutrient intake and clinical alterations: the importance of stratification and model selection . . . . .	50
3.2.1 Introduction to MetS . . . . .	50

---

3.2.2	Experimental Design and Statistical Methods . . . . .	51
3.2.3	Results . . . . .	55
3.2.4	Study Conclusions . . . . .	62
3.3	Chapter Conclusions . . . . .	63
<b>4</b>	<b>A Closer Look: Modelling the Impact of Chemical Composition</b>	<b>65</b>
4.1	Unravelling the Complexity of Food: a Framework for Foodomics . . . . .	65
4.1.1	Introduction to Foodomics . . . . .	66
4.1.2	Challenges and Novel Strategies . . . . .	69
4.1.3	Remarks . . . . .	75
4.2	Multi-Omic Model of the Impact of a Bio-Active Compound . . . . .	76
4.2.1	The role of microbiome sciences in animal production . . . . .	77
4.2.2	Study design and methods summary . . . . .	78
4.2.3	Results . . . . .	80
4.2.4	Discussion and conclusions . . . . .	91
4.3	Multi-Compartmental Model of Complex Compounds . . . . .	93
4.3.1	Interindividual Variability in Bioavailability . . . . .	93
4.3.2	The Bateman equations . . . . .	95
4.3.3	Dataset and modelling . . . . .	95
4.3.4	Early-stage results and discussion . . . . .	99
4.4	Chapter Conclusions . . . . .	105
<b>5</b>	<b>Beyond Composition: the Role of Structure in Physiological Interactions</b>	<b>107</b>
5.1	Food Structure, Function and Artificial Intelligence . . . . .	107
5.1.1	What is the structure of a food? . . . . .	108
5.1.2	How to measure food structure . . . . .	109
5.1.3	Properties affected by food structure: sensory, stability, digestibility and bioaccessibility . . . . .	112
5.1.4	Structure and functional food design . . . . .	114
5.1.5	Predictive models and structure design: how do we feed AI? . . . . .	117
5.1.6	Structure and in-silico simulators . . . . .	121
5.2	Case Study: Modelling with Texture Analysis and Raw Data . . . . .	123
5.2.1	Cooking and water-matrix interaction . . . . .	123
5.2.2	Toward the automatization of water-matrix interactions and structure characterization . . . . .	124
5.2.3	Is learning from raw data and general descriptors promising? . . . . .	125
5.3	Chapter Conclusions . . . . .	128

---

<b>6</b>	<b>Conclusions</b>	<b>131</b>
<b>A</b>	<b>Experimental Designs and Methods Additional Info</b>	<b>135</b>
A.1	Simonetti, Mengucci et al., 2021, Springer Nature, Materials and Methods	135
A.2	Biagi, Mengucci et al. 2020, MDPI, Materials and Methods . . . . .	138
	<b>Bibliography</b>	<b>141</b>

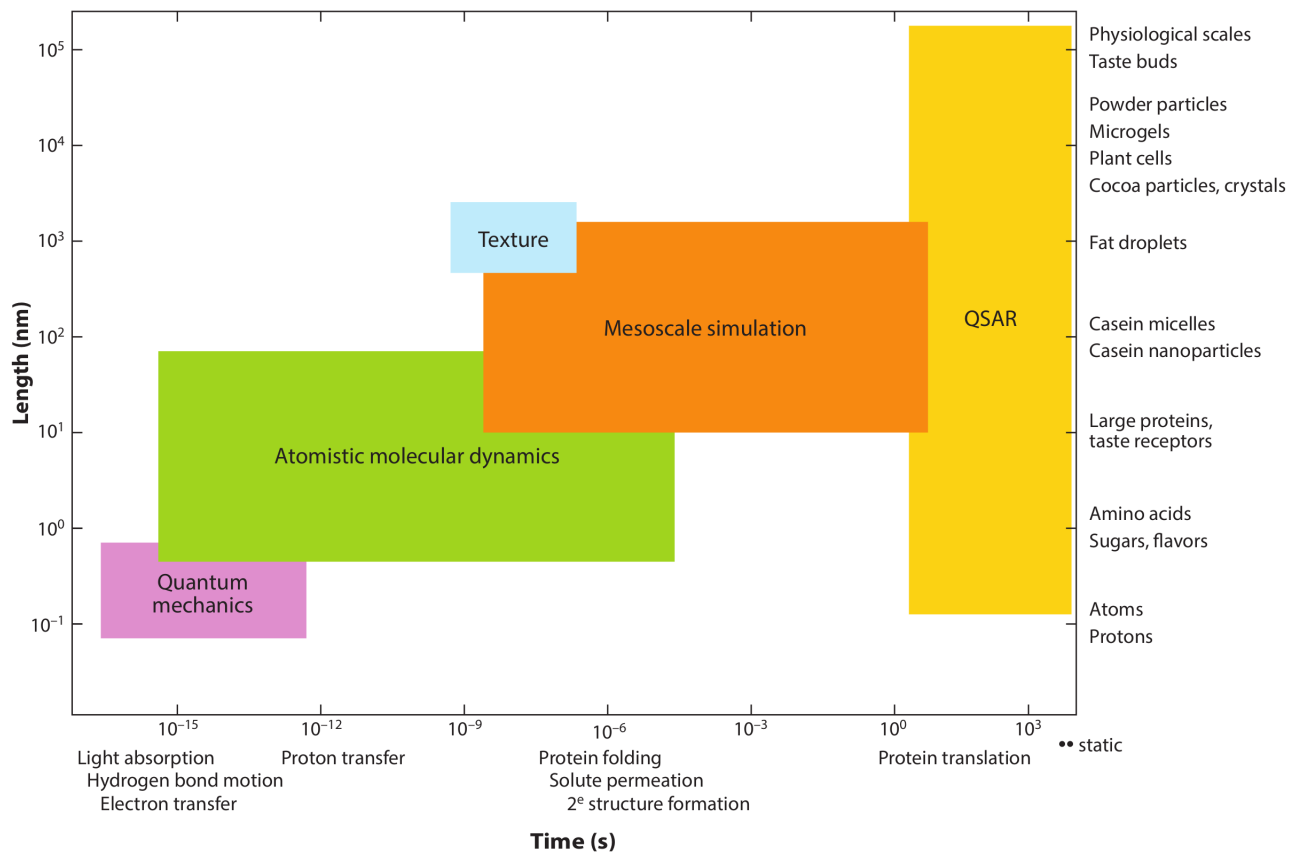


## From Nanoscale to Macroscale and Back: the Complexity of Compartmentalization

### 1.1 Introduction : A Complex Picture

The quest for modelling, simulating and solving complexity in biophysical and biochemical systems has been going on for many years. Boosted by increasing computational power, data generation and analysis together with machine learning and simulation techniques have become the foundations of many modelling routines for such systems. As a matter of fact, linking phenomena happening at various length scales and time scales, becomes a difficult task when solely relying upon predictions obtained from analytical models (i.e. canonical differential equations systems). For instance, let's try to think of all the length scales of structures and time scales of phenomena involved in the description of something apparently ordinary, such as food. The final structure, mechanical and sensory properties of real life food, which can be described in terms of soft matter, arise from a complex series of interactions at molecular levels, physicochemical transformations at the mesoscale and structural properties at the macroscale (1). As such, small changes at the nanoscale (e.g. folding and unfolding of proteins at various interfaces) can lead to drastic changes in macroscale appearance and the stability. The multiscale nature of the phenomena involved in food science is well pictured in figure 1.1. Each time and length scale is bound to require its dedicated description paradigm and simulation framework, with few overlaps between scales. While a pure simulation based approach to predict food properties is theoretically possible, the computational complexity of linking the various level of simulation frameworks for the complete multiscale model poses a practical problem.

Furthermore, interactions with human physiological functions must be added to this already complex canvas, if we want to consider also the final purpose of food: human nutrition. If modelling transformation, processing and molecular interactions that constitute



**Figure 1.1:** Schematics of molecular interactions in food science phenomena across different time and length scales, with appropriate particle based simulation methods. QSAR: quality structure–activity relationships . Adapted from da Silva et al. 2020, AnnuRev

the macrostructure of food and its properties leads to solving a multiscale problem, the same yields for the various stages at which food is digested and metabolized by the human organism. Digestion is a process unfolding in many steps, starting with oral breakdown of ingested substances as the igniter for a cascade of interconnected kinetic processes in the Gastro Intestinal Tract (**GIT**). Modelling this first key step already poses a non-trivial challenge and many *in-silico* simulations techniques and models are being proposed (2). Food fragment structure and size influences enzymatic hydrolysis and gastric emptying (3), which are key regulators of the overall nutrient absorption kinetics. The rest of the steps are a cluster of feedback interconnected kinetics processes of molecular (amino acids, protein, lipids etc.) transport and absorption, that ultimately sink in the circulatory and excretory systems. These processes, happening at different timescales and length scales, cannot be modeled using classic pharmacokinetic and simple mass action laws, as they are not taking into account the structure activated feedback affecting digestion.

Another layer of complexity is added by the physiological variability of individuals.

Metabolic functions are affected by a series of factors such as life habits, health conditions, age and genetic background. Several studies, collected by Walther et al. (4), point out that these factors are responsible for inducing variability in individual functionality of processes and phenomena in the GIT. During many stages of the digestion, protein assimilation, intestinal peptidome composition, protease activity, fat processing, carbohydrate processing and absorption, glucose transport, vitamin absorption, mineral absorption can be altered in different individuals. Furthermore, differences in functionality and activation of digestive processes, starting at the molecular level, leads to variability in systemic components of the human organism, such as gut microbiota composition. Given the many aspects involved in studies of human metabolism and its interaction with bio-molecules and compounds, which require the linking of a large number of heterogeneous descriptive paradigms and modelling tools, one can argue if the goal of a holistic physiological outcomes predictor based on molecules and metabolic functions interactions can be reached.

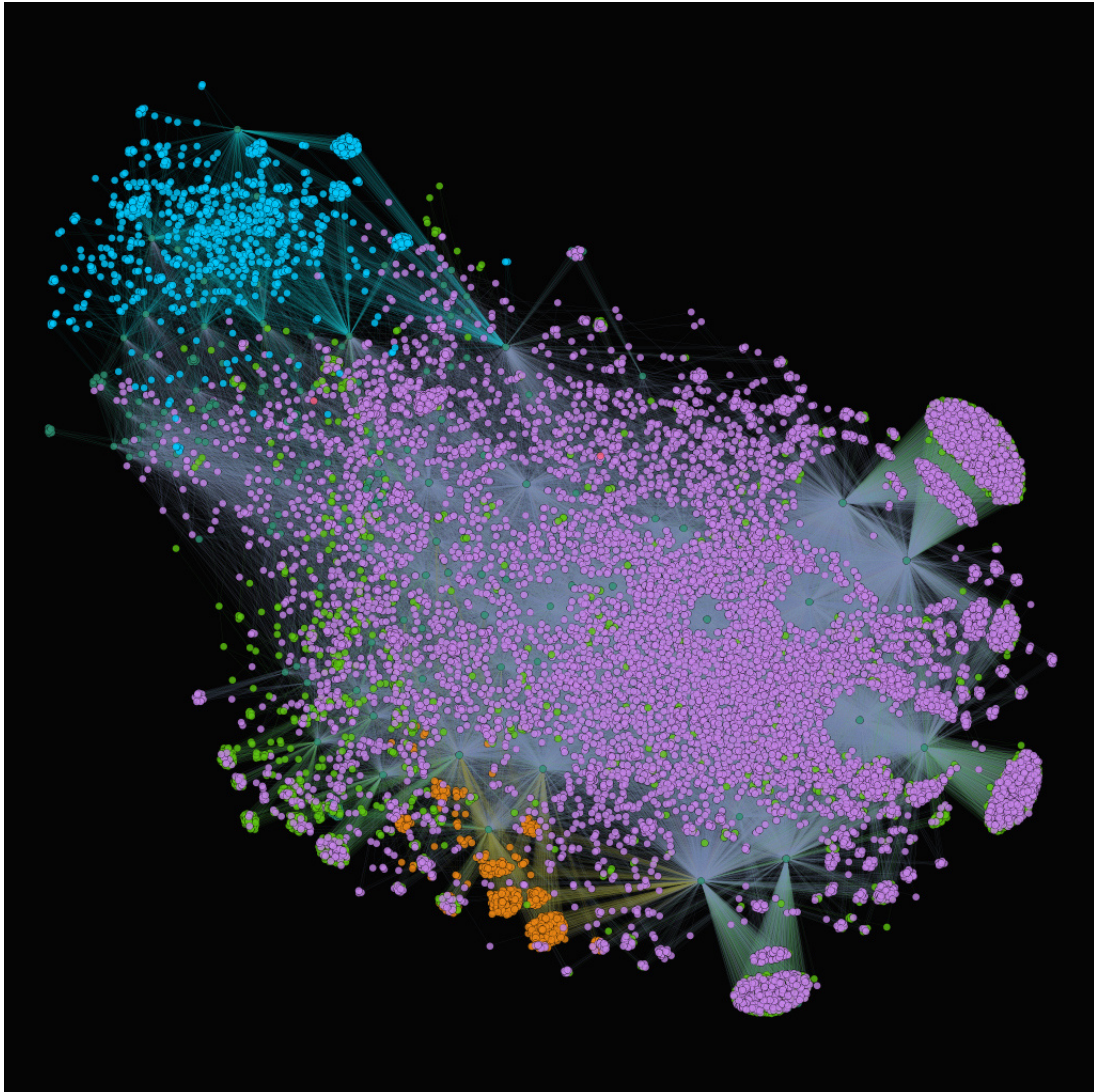
## 1.2 Omics Sciences and Complexity: We Are Our Molecular State

In physical terms, it is clear that the the human organism, as many other biochemical systems, is a nested complex system. From molecules and enzymes, to cells, to organs and tissues, each observational level that we can think of when modelling the human organism is itself a complex systems or its elementary constituents behave like one. A system is considered complex when its modelling is intrinsically difficult, due to the interactions between its component. These kind of systems, are (not easily) defined by the properties that arise from these interactions and how the system ultimately forms relationship with its environment (5). While no exact definition of complexity is generally accepted yet, one of the features that surely defines complex systems is the presence of emergent behaviors. A system is said to exhibit emergence when it is characterized by properties or phenomena that are impossible to predict from studying its isolated components. Emergence is very common in biochemical systems: one of the earliest (and most groundbreaking) example of this notion in the field of physicochemical sciences is the Hartree-Fock method for the computation of molecular structures (6), which introduces the idea that is impossible to exactly determine the properties arising from certain molecular structure from the quantum states of its isolated atomic constituents. The same holds true for molecules constituting cells, which are themselves complex systems, cells constituting tissues and organs, organs constituting anatomic systems which in turn constitutes individuals that exhibit social behaviors. Each one of such levels has to be treated as a complex system, that can be characterized by emergence, openness (energy dissipation and distance from energetic equilibrium), critical transitions (abrupt transition between system states), memory (as opposed to Markovian systems, for which each state depends exclusively on the state reached during a previous event), non-linearity and feedback loops (the effect of a component behavior is fed back at a certain time point, modifying said elementary

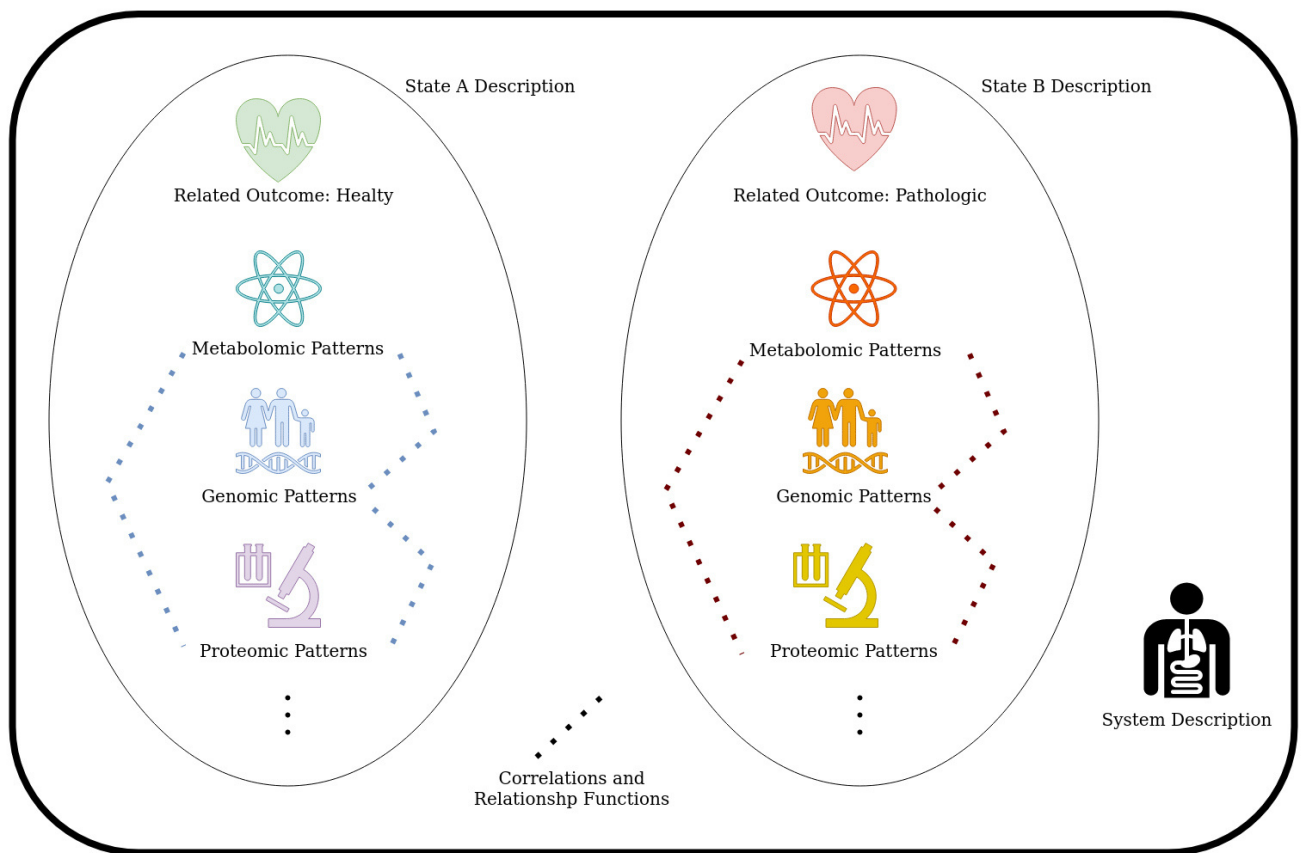
behavior). The study of the human organism as a nested complex system is being greatly impacted by the advent of the so-called *Omic Sciences* era. Since the success of the initiative of mapping and sequencing human genome, a great deal of technologies have been developed to obtain huge amount of molecular data from cells, tissues and bio-fluids. Examples include proteomics, the global analyses of proteins, transcriptomics, the analysis of RNA, genomics, the analysis of genes, metabolomics, the analysis of metabolite patterns, and epigenomics, the analysis of DNA methylation and modified histone proteins in chromosomes (7). This high-throughput ways of obtaining molecular information can be applied to characterize different elements of a biological system, as a snapshot of the underlying complex biological interactions at various level of resolution (Fig.1.2). With such quantities of information and levels of resolution, a new paradigm of obtaining a comprehensive understanding of complex biological systems is on the way. The reference state A of a generic biological system can be represented as the whole set of omics measurements of their components (molecules) at different resolution levels and their correlations and relationship. Let's now suppose that the same system goes from the state A to a state B, associated to a physiological outcome (example: a patient goes from its healthy state A to a pathological state B). The altered state B is defined by the changes of patterns of omics measurements and by the perturbations of how measurements of various omics might be related. In other word, if a way of integrating various omic data exists, we are theoretically able to use each single measurement as a parameter to define the state of our system, through different levels of resolution and elementary component definitions (Fig.1.3). This aspect is crucial in biological complexity understanding: it is a way of connecting emergent properties to elements of the system at finer resolutions. This in turn means being able to understand the etiology and causality of phenomena in a biological system, from its finest scale to the macro, and explain complex behaviors that cannot be predicted by isolating the system elements, like pathological states in the human organism.

The possibilities emerging thanks to omic high-throughput techniques in the description of biochemical complex systems and in complexity science in general are astounding. However, when trying to describe a system as an ensemble of heterogeneous patterns and pattern variations of the molecular state, many practical aspect must be considered. The high-throughput nature of omic techniques implies the creation of datasets with very large number of parameters (features). This in turn generates the need for non-trivial statistical and heuristic modelling, based upon machine learning and deep learning for data integration, parameter and model selection. Each omic technique has its own array of suitable tools for processing, dependent on the physical nature of the measurements and the mathematical background of the subsequent data generation. In the next chapters, a closer look to metabolomics and results obtained with originally developed metabolomic-based frameworks is proposed, to assess the real capabilities of studying altered states of the human organism with different levels of molecular descriptors patterns.





**Figure 1.2:** An example of interacting and correlated information from various omics sciences of possible target genes underlying COVID-19 spike protein, represented as a graph. **Magenta nodes:** genes; **light green nodes :** single-nucleotide polymorphisms; **dark green nodes :** related diseases and comorbidities; **cian nodes :** phenotypes; **orange nodes :** metabolic pathways. Original rendering from the author, with force atlas computation from the CHIMeRA project. Nico Curti & Carlo Mengucci, CHIMeRA : Complex Human Interactions in MEDical Records and Atlases, Conference on Complex Systems, 2019, Trento, Italy



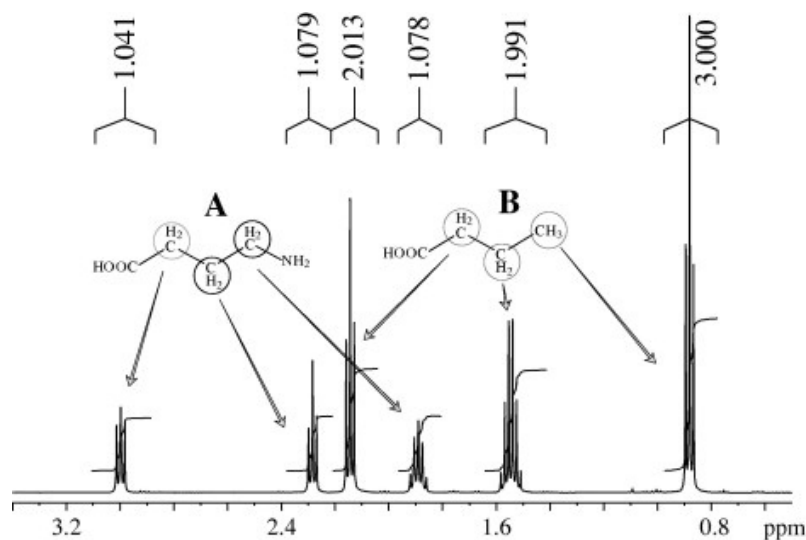
**Figure 1.3:** Interpreting the description of a biochemical complex system (the human organism) with omic data. The crosstalk between different levels of molecular descriptors patterns, linked to different omic techniques allows the description of the system in terms of its states, defined by the correlations with a physiological outcome.

### 1.3 Tools of the Trade : Metabolomics and Machine Learning

Metabolomics has been defined as "*the quantitative measurement of the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification*" (8). This definition already gives an idea of what kind of snapshots the metabolomic analysis of cells or biofluids can offer: a molecular-level ensemble of all up-stream genomic, transcriptomic and proteomic feedbacks of an organism to a given perturbation (9). As a matter of fact, metabolites are the end product of cellular pathways and are a direct result of protein and enzymatic activities. This in turn means that metabolites serve as good proxies of phenotypic exposures or perturbations such as diseases. A change in the expression level of a gene or protein does not necessarily correlate with a variation in the activity level of a protein, but an alteration in metabolite concentrations only occurs through such a change (10). For this reason, information from metabolites is generally more adjacent to perturbations such as diseases than information from genomics or proteomics, making metabolites ideal biomarkers of exposures to many factors such as drugs, diet, environmental chemical exposures etc. Thus, the overall purpose of metabolomics is to identify a subset of molecular features that can define the (possibly perturbed) state of a system, against a gargantuan and complex background of metabolites and their chemical surroundings that constitutes the system itself (i.e. samples from a tissues, cells, biofluids...). The attainment of such purpose is obviously hindered by a series of problems. Among the others, stands the lack of a complete catalogue of the human metabolome and the metabolome of several organisms (11). This means encountering unknown signals when analyzing spectra, making the interpretation of metabolic changes often incomplete. Furthermore, metabolomic data analysis is complicated by the fact that all biological systems are easily perturbed by any number of experimental or environmental factors, such as age, diet, pH, sex. These background perturbations are often cause of variability in samples, which can hide and confound the effect of the exposure (disease, intake etc.) to be modelled. For a complete overview on the topic, see (12). Overall, metabolomic analyses require robust methodologies to discover latent trends in complex datasets with variance coming from many confounding sources.

#### 1.3.1 NMR Spectroscopy in a nutshell

Nuclear Magnetic Resonance (NMR) spectroscopy is a quantitative and non-destructive experimental technique broadly employed in chemistry, providing information on the molecular structure of compounds and on the chemical surroundings of complexes. While many nuclei can be detected with this technique ( $^{13}\text{C}$ ,  $^{13}\text{P}$ , ...), the studies presented in this work are based on the detection of  $^1\text{H}$  nuclei, due to their abundance in organic compounds and biological samples. An NMR spectrum is essentially a plot of the radio frequency applied against absorption, in which each signal is referred to as resonance. The frequency of a signal is determined from its chemical shift  $\delta$ , defined in absolute terms as



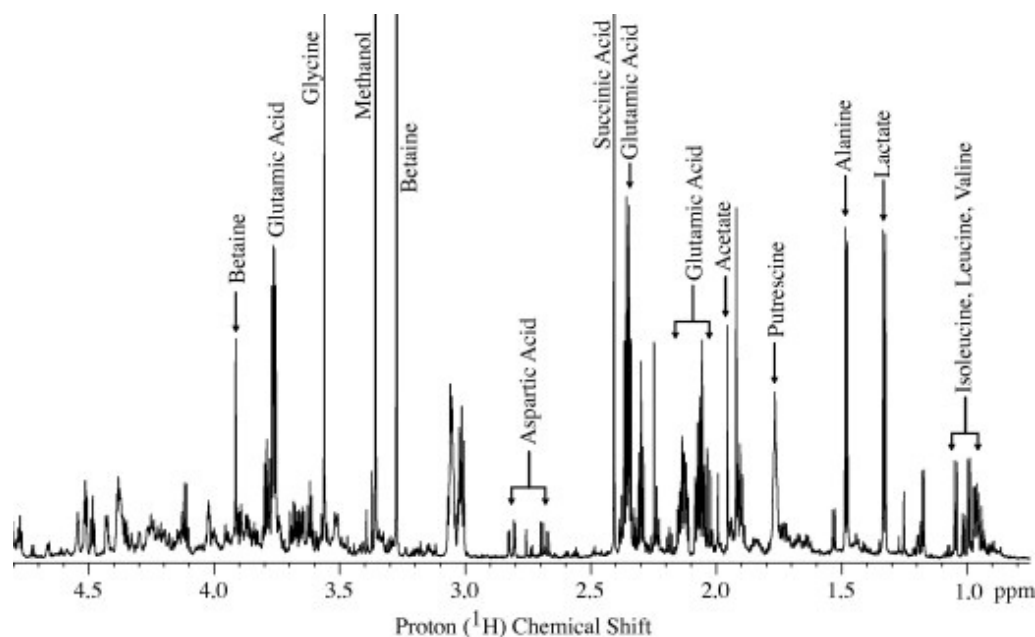
**Figure 1.4:** Proton NMR spectrum of a mixture of (A)  $\gamma$ -amino butyric acid and (B) *n*-butyric acid showing relative integral values of intra-molecular resonances that depend upon the number of nuclei per resonance and their relative concentration. Adapted from Santosh Kumar & Raja, 2012, Elsevier

the relative position of a frequency of resonance with reference to a standard compound, expressed in ppm (part per millions).  $\omega_S$  = frequency of signal,  $\omega_R$  = frequency of reference,  $\omega_{spec}$  = spectrometer frequency.

$$\delta = \frac{\omega_S - \omega_R}{\omega_{spec}} \times 10^6 \quad (1.1)$$

The detection of  $^1H$  nuclei contained in compounds provides information about the quantitative relationship between intra-molecular and inter-molecular resonances, through chemical and coupling constants (13), Figure 1.4. Translating to complex mixtures (e.g., cell extracts, tissue extracts, body fluids, natural-product isolates and drug formulations),  $^1H$  spectroscopy is capable of grasping quantitative information of their components without separating them from the chemical environment of the sample. As such, the full spectra of a biological sample provide absolute and relative quantification of several metabolites, molecules in which  $^1H$  nuclei are abundant (Figure 1.5). Specifically, signals acquired in such a way provide a representation of the distribution of proton nuclei within the molecules and the different concentration levels of the corresponding metabolites in the complex mixture (14). This type of spectroscopy allows the characterization of liquid and semi-solid biological specimens through a molecular fingerprint.

One dimensional  $^1H$  spectra have gained popularity in metabolomic studies thanks to the fact that NMR is a highly automatable, highly reliable and fast technique. This allowed NMR-based metabolomic studies to rely on collections of large number of spectra, making this technique particularly suitable for complex data analysis mostly based



**Figure 1.5:** Expansion of one-dimensional single pulse  $^1\text{H}$  NMR spectrum (0.50–4.80 ppm) of aqueous extract of bacterial cells. It represents a number of metabolites detected using a single pulse NMR experiment. The intensities of resonances depend on their respective concentrations in the extract. Adapted from Santosh Kumar & Raja, 2012, Elsevier

on machine learning. Though not as sensitive as mass spectroscopy, which is capable of identifying metabolites in the order of the thousands with concentrations of  $> 10$  to  $100$  nM, NMR-based acquisitions can retrieve information on tens to hundreds of metabolites at a time; a sufficient number for fingerprinting approaches. Metabolites can be automatically or semi-automatically matched using spectral databases such as HMDB (15) or proprietary catalogues such as Chenomx (*Chenomx Inc.*), which is the spectral reference source of choice for all the studies presented. Overall, NMR-based metabolomics has become a preferred tool in large-scale studies thanks to reproducibility and high speed analytical capabilities over large number of samples.

### 1.3.2 NMR-based metabolomics data analysis

$^1\text{H}$  NMR spectra contain the convolution of molecular ensembles of a great number of metabolites. The goal of metabolic fingerprinting experiments is to determine the relative differences between the metabolomes of two or more systems to infer a biological relationship. Thus, NMR-based metabolomics datasets can be seen as collections of spectral information, in which each spectral feature is interpreted as a variable of the underlying statistical model. In other words, metabolomic datasets are high-dimensional datasets with many sources of variance, nominally the perturbations that can occur in the system (diseases, drug intake, but also experimental factors, sample storage, pH etc.). For this reason, the most popular approaches for fingerprinting are based on the projection



of spectral features into lower-dimensional latent features space, as a way to summarize and interpret results more easily. Moreover, these methods are necessary to solve known statistical problems of datasets with a large number of variables with respect to the number of samples. The amount of raw spectral features is usually extremely excessive with respect to the samples, leading to huge amount of collinearity in the dataset and singularity of the data matrix  $X_{N \times K}$ , where  $N$  = number of samples,  $K$  = spectral features in each sample. This section is intended to provide an overview on methods of dimensionality reduction, classification and features interpretation of metabolomics data, that the author implemented for the various frameworks of the studies presented in this thesis. For an exhaustive perspective on the topic of data analysis for metabolomic, including preprocessing aspects such as normalization, binning, scaling, baseline correction and signal-to-noise ratio optimization, see (12).

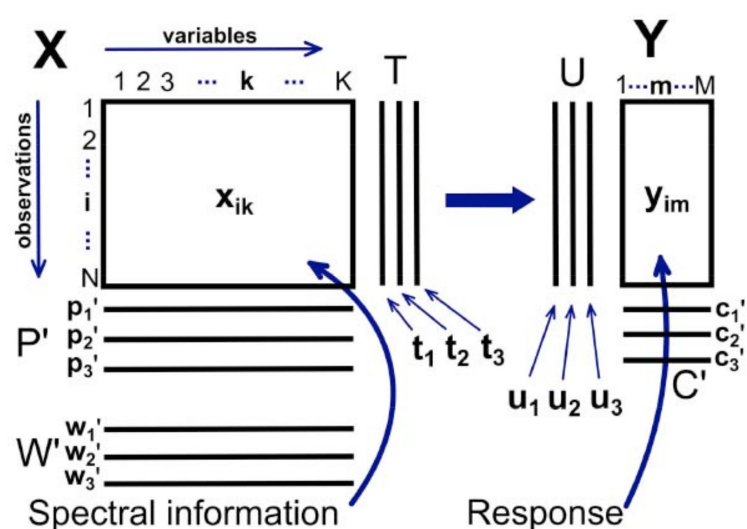
### Dimensionality reduction with single value decomposition of the covariance matrix

The most popular methods for dimensionality are arguably those based on the single value decomposition of the covariance (correlation for centered data) matrix, which is basically a diagonalization due to the properties of symmetry and positive-definitiveness of such matrix. The goal of this methods is to find a matrix  $A_{K \times P}$  representing an optimal linear transformation of the original data matrix  $X_{N \times K}$ , with  $N$  = samples,  $K$  = variables, into a new matrix of  $P < K$ -dimensional scores  $T = XA$ , preserving different kinds and portions of variance in the dataset, as a function of the problem solved by the selected method. Method selection is obviously bound to the purpose of the study and to experimental design: Principal Component Analysis (PCA) is useful when trying to explore sources of variances without assumptions, Factor Analysis is suitable when strong hypothesis about variance sources in the dataset are available (i.e. in a follow up study with spectra acquired over time, to extract time dependent latent components), Partial Least Square Discriminant Analysis (PLSDA) is useful to extract latent variables specifically tied to a class separation. A general overview of how dimensionality reduction is achieved by these methods is shown in Figure 1.6.

### Principal Component Analysis

Principal Component Analysis is widely used in fingerprinting and chemometric studies, to attain unbiased and unsupervised dimensionality reduction. PCA reaches the linear transformation that preserves the maximal variance of the original dataset in a lower dimensional space. This is obtained through an eigendecomposition of the sample covariance matrix (if non-singular). In the PCA problem the matrix  $A$  yielding the dimensionality reduction

$$T_{N \times P} = A_{K \times P} X_{N \times K}; P < K \quad (1.2)$$



**Figure 1.6:** Canonical example of the data ( $X$ ) and response ( $Y$ ) matrices (for supervised decomposition studies) and decompositions thereof used by projection-based multivariate analysis algorithms. In metabolic fingerprinting applications, the data matrix will contain spectral information on its rows, such that every column will represent a single spectral frequency or bin. For supervised projections, each row of data is paired with a corresponding row in the response matrix that holds either continuously varying outputs or binary ( $n$ -ary) class memberships. The data is then decomposed into a small number of score vectors ( $t$ ) and loading vectors ( $p$ ), with a corresponding weight vector ( $w$ ) used to transform rows of  $X$  to scores space. Responses are similarly decomposed into scores ( $u$ ) and loadings ( $c$ ), where  $t$  is an effective estimator of  $u$ . Adapted from (12), 2013, Bentham Science

that projects data into the directions of maximum variance is a matrix formed by the first  $P$  eigenvectors of the sample covariance matrix  $S_{N \times N}$ , decomposed by the following:

$$S = \frac{1}{N-1} X^T H X = Q \Lambda Q^{-1} \quad (1.3)$$

where  $H_{N \times N}$  is the centering matrix to center each feature about their sample mean,  $Q$  is the matrix of the eigenvector of  $S$  and  $\Lambda$  is the diagonal matrix of the corresponding eigenvalues. It is a well known fact that the eigenvalues in  $\Lambda$  computed from the unscaled quadratic form  $X^T H X$  equals the variance of the new transformed data in  $T$ . Thus, the amount of variance in  $X$  preserved by the  $i$ -th principal component, as a ratio of the total original variance, is given by:

$$R_i^2 = \frac{\Lambda_{ii}}{\sum_{j=1}^N S_{jj}} \quad (1.4)$$

PCA is useful when looking for an unbiased and unsupervised reduction of the data. Furthermore, sources of variances can be investigated by relating latent components to physiological outcomes or classification tasks, to model the effect of perturbations in the metabolome.

## Factor Analysis

Factor analysis is used to describe variance among observed and presumably correlated variables (in the case of metabolomics, the spectral features) in terms of a potentially lower number of gaussian latent variables called factors. The observed variables are modelled as linear combinations of latent factors and an error term. Factors are assumed to be independent from the error term and uncorrelated. The coefficients of the linear combinations, called loadings, are proportional to the extent of how a variable is related to a given factor (16). In matrix notation, we look for a loading matrix  $L_{K \times P}$  such that:

$$X - M = L F + \epsilon; \text{ with conditions } : \begin{array}{l} E(F) = 0 \\ Cov(F) = I \end{array} \quad (1.5)$$

where  $M$  is the matrix of observed variables sample mean,  $F_{N \times P}$  is the factor matrix with  $P < K$ ,  $\epsilon_{N \times K}$  is the error matrix (independent from factors by assumptions),  $I$  is the identity matrix and  $E(F)$  is the matrix of expected values for  $F$ . The condition  $Cov(F) = E$  ensures that factors are uncorrelated, without loss of generality. By solving such a problem, latent components obtained with factor analysis maximize the shared portion of variance underlying independent factors. While factor analysis formulation and solution sounds somewhat similar to PCA and can sometimes yield similar results, there is a deep conceptual difference between the two. Since any rotation of a FA solution is itself a solution, interpreting factors without an external hypothesis is difficult.



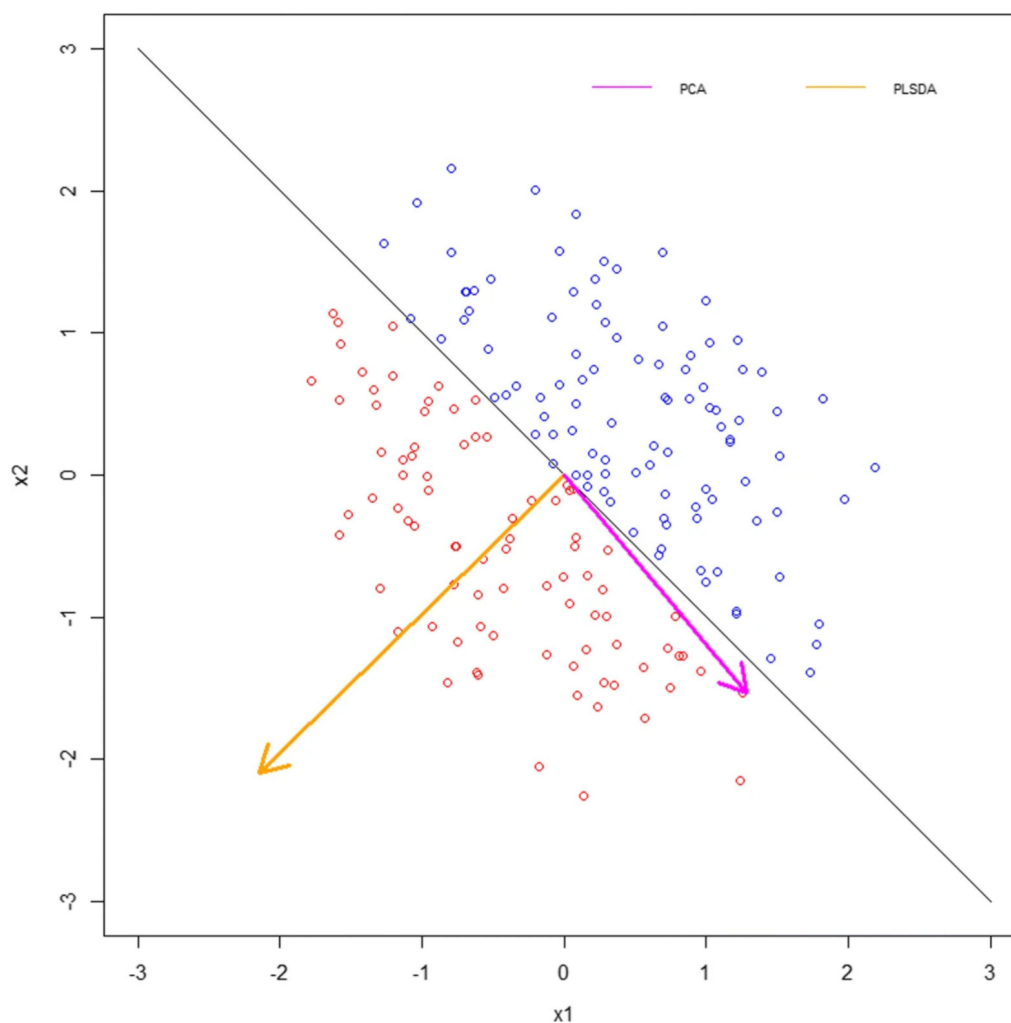
This means that FA is useful to test hypotheses about variance in datasets introduced by structured unobserved factors: as an example if we know in advance from experimental design that metabolomes in our study can be perturbed by two different independent factors, we can be confident that a dimensionality reduction based on factor analysis will yield a representation of which spectral features are most related to said factors.

### Partial Least Square Discriminant Analysis

While PCA offers an unbiased dimensionality reduction, it may underperform in the detection of group and cluster structures in samples when within-group variation is too large with respect to inter-group variation. In these kind of situations, it maybe helpful to build latent components that are not exclusively based on the variance contained in the data matrix. PLSDA offers a framework for supervised projection on latent components, based on the univariate weight that each variable contained in  $X$  has on the prediction of a certain outcome contained in an outcome matrix  $Y$ . Given a matrix of outcome  $Y$  (that contain in example the labels of classes of our samples), with the constraint that each partial least square latent component is orthogonal with respect to the others, the matrix  $A$  in 1.2 that is the solution of the dimensionality reduction problem, is given by the matrix made by the first  $P$  eigenvectors of the quadratic form:

$$S = S_{xy}S_{yx} = \frac{1}{(N-1)^2} X^T H Y Y^T H X \quad (1.6)$$

where  $H$  is the centering matrix and  $S$  is the matrix of covariances between  $X$  and  $Y$ . Thus, by solving this problem, PLSDA seeks components that have high variance and have high correlation with the response, in contrast to principal components regression/analysis which keys only on high variance (17). A complete derivation of partial least squares solutions from regressions by successive orthogonalizations, to ensure orthogonality of the linear combinations computed thorough PLS is given by Hastie et al. (18). The difference in solutions obtained with PCA and PLSDA is visualized in Figure 1.7. While PLSDA is useful to extract latent components for classification tasks, it must be noted that the method is extremely prone to overfitting under certain conditions (high variables-to-samples-ratio).



**Figure 1.7:** Different direction obtained for projection with PCA (magenta) and PLSDA (orange). PCA projects the 2-dimensional feature space into the direction of maximum variance. PLSDA projects the 2-dimensional feature space into a direction of high variance AND high correlation with the outcome (in this example, the separation of blue dots from red ones). Adapted from Ruiz, 2020, BioMed Central

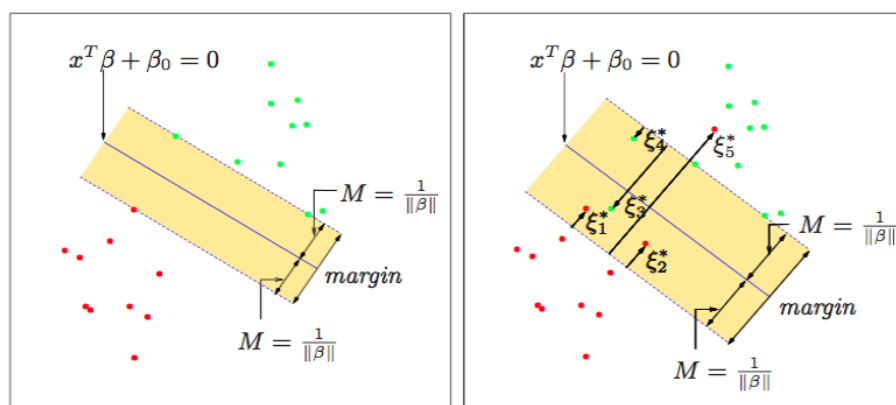
### 1.3.3 Classifiers

This section provides an overview of classification and clustering algorithms that are exploited in pipelines and frameworks by the authors. The role of classification algorithms in the studies presented in this thesis are:

- Evaluation of the robustness of fingerprinting features in classification problems
- Selection of minimal sets of latent features extracted from spectra to optimize classification problems and interpret correlation with outcomes
- Integration and clustering of features from various omic sources

### Support Vector Machines

A support vector machine can be seen as a generalization of linear decision boundaries for classification. It is a method to assess optimal separating hyperplanes for non-completely separable classes problems (18), while preserving an interpretation akin to linear decision boundaries. This is achieved by representing nonlinear boundaries as a linear boundary projected in a larger-dimensional, transformed version of the feature space. A representation of how a support vector classifier operates to solve non-separable problems is reported in fig.(1.8)



**Figure 1.8:** Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width  $2M = \frac{2}{\|\beta\|}$ . The right panel shows the nonseparable (overlap) case. The points labeled  $\xi_j^*$  are on the wrong side of their margin by an amount  $\xi_j^* = M\xi_j$ ; points on the correct side have  $\xi_j^* = 0$ . The margin is maximized subject to a total budget  $\sum \xi_i \leq C$ . Hence  $\sum \xi_j^*$  is the total distance of points on the wrong side of their margin. Adapted from Hastie et al. 2001, Springer

Support vector machines are capable of great generalization in classification performances, even with complex datasets. Furthermore, many available implementations (such as *sklearn* for the Python environment) support feature importance scores, making SVM

suitable to assess which sets of latent component can achieve optimal classification performance. This feature is particularly useful when trying to connect latent variables to possible sources of variances in the dataset.

### Classification with boosted ensembles of classifiers

Metabolomic and in general multi-omic datasets are prone to contain a very large number of features (variables). Especially for multi-omic scenarios, it is often difficult to make assumption about features and the relationships between different types of features (spectroscopic, transcriptomic, relative abundances of microbial species, peptides sequence lengths...). It is also often difficult to predict how the integration of features may translate to dimensionality reduction and separability in lower-dimensional spaces. The boosting approach proves useful in these situations, by building a meta-estimator based on an ensemble of simple classifiers that are trained on the same dataset, but are iteratively adjusting the weights of misclassified samples so that successive instance of classification focus on more difficult cases. Specifically, the AdaBoost algorithm is an iterative procedure that tries to approximate the ideal and unbiased Bayes classifier by combining many weak classifiers. Starting with the unweighted training sample, the AdaBoost builds a classifier, for example a decision tree, that produces class labels. If a training data point is misclassified, the weight of that training data point is increased (boosted). A second classifier is built using the new weights, which are no longer equal. Again, misclassified training data have their weights boosted and the procedure is repeated. A score is assigned to each classifier, and the final classifier is defined as the linear combination of the classifiers from each stage (19), Figure 1.9. To put the idea in simply terms, boosting is an iterative method for complex classification problems that relies on the contributions of many simple classifiers instead of a single, complex classifier with a large number of hyper-parameters. This approach has many advantages when using simple but highly general classifiers such as decision trees (20). Decision trees operate by seeking the group of features that better separate the elements of a dataset in various nodes, until the highest number of elements of the same class end up in the same node (supervised technique). Thanks to their easy interpretation, ensembles of decision trees can serve for simultaneously grouping heterogeneous features and trying to achieve good classification performances. Furthermore, an easy interpretation of how features interact together with respect to a classification outcome, if the outcome is tied to a perturbation of the system such as a disease, is useful to draw etiologic conclusion from the study.

1. Initialize the observation weights  $w_i = 1/n$ ,  $i = 1, 2, \dots, n$ .
2. For  $m = 1$  to  $M$ :

(a) Fit a classifier  $T^{(m)}(\mathbf{x})$  to the training data using weights  $w_i$ .

(b) Compute

$$err^{(m)} = \sum_{i=1}^n w_i \mathbb{I}(c_i \neq T^{(m)}(\mathbf{x}_i)) / \sum_{i=1}^n w_i.$$

(c) Compute

$$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}}.$$

(d) Set

$$w_i \leftarrow w_i \cdot \exp\left(\alpha^{(m)} \cdot \mathbb{I}(c_i \neq T^{(m)}(\mathbf{x}_i))\right),$$

for  $i = 1, 2, \dots, n$ .

(e) Re-normalize  $w_i$ .

3. Output

$$C(\mathbf{x}) = \arg \max_k \sum_{m=1}^M \alpha^{(m)} \cdot \mathbb{I}(T^{(m)}(\mathbf{x}) = k).$$

**Figure 1.9:** The AdaBoost algorithm, as originally formulated by (21). At each step, the exponential loss function is minimized to achieve an approximation of an ideal Bayes classifier through an ensemble of weaker classifiers. adapted from Hastie et al., 2009, International Press of Boston

## Hierarchical Clustering

Clustering is useful to group samples and features alike, to assess the similarities between samples and the patterns of interacting features that contribute to characterize them. Hierarchical bi-clustermaps might be used as a fast and informative visualization. As in (22), hierarchical clustering performances are metric and linkage functions dependent. The Nearest point linkage is one of the most common. Suppose there are  $|u|$  original observations ( $u[0], \dots, u[|u| - 1]$ ) in cluster  $u$  and  $|v|$  original objects ( $v[0], \dots, v[|v| - 1]$ ) in cluster  $v$ . Let  $v$  be any remaining cluster in the forest that is not  $u$ .

The Nearest Point Algorithm assigns:

$$d(u, v) = \min(\text{dist}(u[i], v[j])) \quad (1.7)$$

for all points  $i$  in cluster  $u$  and  $j$  in cluster  $v$ .

Other possible linkage algorithm are:

- **Farthest Point Algorithm or Voor Hees Algorithm**, assigns:

$$d(u, v) = \max(\text{dist}(u[i], v[j])) \quad (1.8)$$

for all points  $i$  in cluster  $u$  and  $j$  in cluster  $v$ .

- **UPGMA algorithm**, assigns:

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)} \quad (1.9)$$

for all points  $i, j$  where  $|u|, |v|$  are the cardinalities of clusters  $u, v$  respectively.

- **Ward algorithm**, assigns:

$$d(u, v) = \sqrt{\frac{|v| + |s|}{T} d(v, s)^2 + \frac{|v| + |t|}{T} d(v, t)^2 - \frac{|v|}{T} d(s, t)^2} \quad (1.10)$$

where  $u$  is the newly joined cluster consisting of clusters  $s$  and  $t$ ,  $v$  is an unused cluster in the forest,  $T = |v| + |s| + |t|$ , and  $| * |$  is the cardinality of its argument.

Further weighted methods are described in (22).

## Fingerprinting Perturbations and Dynamics of Metabolic States

### 2.1 Fingerprinting Perturbations : From Cells and Systems to Enzymatic Networks

This section is based on the published work by *Simonetti, Mengucci et al. (23), 2021, Springer Nature.*

The framework and results presented are an example of how to treat, integrate and model data from various omics sciences and compartments of the same omic (in this case, metabolomics) to obtain a fingerprint of a pathological perturbation (acute myeloid leukemia, AML) that defines the system at different levels of resolution. In this work, the cellular and the systemic level of the metabolome are linked with data from different biofluids and intracellular measures. Spectral metabolomic features are then integrated with genomic data, to fingerprint upstream and downstream perturbations and describe the system through the crosstalk of different molecular descriptors. The resulting observed integrated signatures are then fed to a flux variability models (FVA) (24) of the full blood cell metabolic network described at enzymatic level, to evaluate the impact of systemic perturbations linked to AML and predict possible therapeutic targets.

#### 2.1.1 An introduction to fingerprinting in Acute Myeloid Leukemia

Current personalized therapeutic approaches in acute myeloid leukemia (AML) are generally restricted to those patients with identifiable and target genomic lesions (25; 26; 27). However, these approaches do not target interactions between cancer-related features and homeostatic mechanisms that define the leukemic phenotype. The metabolome is the result of genome- and proteome-wide interactions and is shaped by microenvironmental factors. The biofluid metabolome has been extensively investigated to identify predictive signatures in cardiovascular disorders (28), diverticular disease (29) and diabetes (30),

and specific metabolic profiles have been associated with cancer risk (31). In oncology, metabolomics is a valuable approach for diagnosis, prognostication, and disease monitoring (32). A paradigmatic example in AML is the accumulation of serum, urine, and intracellular 2-hydroxyglutarate (2-HG) in IDH1/2-mutated (mut) cases (33; 34). 2-HG is an oncometabolite (35; 36), predicts clinical outcome (37), and is a noninvasive biomarker of disease activity (34).

Aberrant enzymatic activity drives cancer metabolic reprogramming and cooperates with mutations of tumor suppressors and oncogenes in pathogenesis. For example, AML cells reduce both host insulin sensitivity and secretion to increase glucose availability for malignant cells (38). The glycolytic pathway sustains leukemia maintenance and progression. AML cells have a higher mitochondrial (mt) mass and oxygen consumption rate than normal hematopoietic cells (39). Moreover, leukemia stem cells (LSCs) are addicted to oxidative phosphorylation (OXPHOS) for energy production (40). OXPHOS is sustained by elevated amino acid metabolism in LSC from de novo AML (41), with cysteine playing a crucial role (42), and is controlled by glutamine levels (43). Targeted inhibition of these pathways, among others, induces cell death and/or differentiation of AML cells (39; 41; 42; 43; 44; 45). However, the specific response of AML molecular subtypes to agents targeting metabolism has been rarely investigated (46; 47; 48; 49). The reported integrated genomic-metabolic study in AML identified, based on intracellular and the biofluid metabolic profile, a specific NPM1-mut AML subgroup characterized by mutations of genes involved in DNA damage response and/or chromatid cohesion (NPM1/cohesin-mut) and high levels of serum choline+trimethylamine-N-oxide, and leucine. In silico modeling of the intracellular metabolome based on transcriptomic data highlighted perturbations in the purine and NAD metabolic pathways as NPM1/cohesin-mut-specific alterations.

### 2.1.2 Study design and methods summary

This section is a short summary, focusing on study design and methods applied to metabolomics, to guide the reader through the results of the study. **A complete and detailed report of all experimental materials and methods is provided in Appendix A.** The full supplementary material is available at <https://www.nature.com/articles/s41375-021-01318-x#Sec17>.

#### Cohort and Study design

Participants were included if they were free from infective, autoimmune, celiac, or metabolic diseases such as diabetes and dyslipidaemia. Kidney and liver integrity were also checked. Subjects with acute or chronic renal or hepatic disease, renal or hepatic impairment, cardiovascular disease or a history of neoplasia were excluded from the control cohort. Serum samples from 119 AML and 145 healthy subjects and urine samples of 103 AML and 139 controls were collected in the fasting state (in the morning). All partici-



pants were Caucasian except for 5 (3.4%) healthy controls and six (5%) AML patients. To reduce potential bias and variation unrelated to AML pathogenesis and to ensure that the observed metabolic differences were not due to external confounders, we collected, when possible, two independent serum and urine samples from each patient (more than 50% of cases). Moreover, information on age, gender (the cohorts were balanced for gender), race, health status, diet, drug intake, physical exercise was collected along with specimens and used to filter nuclear magnetic resonance (NMR) spectra during the quality control procedures.

### NMR spectroscopy and metabolomics

Serum and urine samples were analyzed by NMR spectroscopy (acquisition details in Appendix A). A stochastic GridSearch was implemented to select the best combination of parameters for dimensionality reduction and classifier performances. Unsupervised and supervised dimensionality reduction were performed using principal component analysis (PCA) and partial least squares discriminant analysis (PLSDA)-sparse(s)PLSDA, respectively. For subset extraction, weights were obtained after signal smoothing via signal-to-noise ratio threshold (which was essential due to unavoidable use of data scalers for dimensionality reduction). The latent components of spectra containing maximum information related to molecular features were identified by a genomic-guided semisupervised approach. This means that the combination of urine and sera latent components used for clustering is extracted with classifiers-derived scores, from classifiers trained with the purpose of discriminating TP53-mut/aneuploid, NPM1-mut and chromatin/spliceosome-mut samples. Signals in the spectra corresponding to loadings and weights emerging from different tasks were checked for alignment. To minimize the possibility of confounding effects, every step of each classification and clustering task was crossvalidated through suitable k-folds, stratified for gender and age when possible depending on class sizes and sample sizes for the tasks. Three different machine learning algorithms/classifiers were used for each task to perform cross-validated predictive modelling using latent components as inputs: Linear Kernel SVM, Random Forest Classifier, Ada Boosting Tree Classifier, with AdaBoost being the best performer across most of the tasks. Pipelines and algorithm scripts were implemented using Python 3.6 and SciKit.Learn module for machine learning routines. Feature-related scores from SVM classifier were used to extract the best subsets of latent components for plotting. The assumption is that the combination of the top 3 features contributing to SVM classification yields the best possible 3D space where group linear separability emerges. A similar approach is used when selecting the best latent components to be investigated for a given task.

#### 2.1.3 Results

### The combined analysis of serum and urine profiles improves AML metabolic characterization

Given that the metabolite composition of biofluids reflects the real-time activity of all biochemical processes in the body and that leukemic cells alter systemic physiology (38), we compared the profile of blood and urine metabolites of AML patients (serum: 88 at diagnosis and 31 at relapse, urine: 80 at diagnosis and 23 at relapse) and healthy controls (CTRL, serum: 145, urine: 139). The metabolomic profile provided efficient discrimination between patient and CTRL both at serum and urine level, with an accuracy of 83% (Figure 2.2A) and 85% (Figure 2.2B), respectively. Since patient and CTRL cohorts were not age-matched (median age: AML, 67-years (18–90), CTRL, 57-years (23–75)), we verified that age had no significant effects on classification. Notably, the integration of serum and urine data yielded an average accuracy of 90% in the separation of AML and CTRL (Figure 2.2C), by using a reduced number of features compared with the analysis of each biofluid per se. In serum, PC2–3 space gave the best 2D combination for AML-CTRL separation, with 13 metabolites showing significantly different levels ( $p < 0.05$ , Figure 2.2D and Table 2.1). These metabolites were not significantly correlated with age or gender. Amino acid and tricarboxylic acid cycle (TCA) cycle byproducts, that had increased concentration in AML except for glutamine and threonine, mainly represented variance in PC3, while lactate and fatty acid metabolism compounds accounted for variance in PC2 (Figure 2.2D). When looking at sample distribution along serum PC3, that provided a good discrimination between AML and CTRL, we observed that all AML subgroups were significantly different from normal cases, independently of bone marrow or peripheral blast percentage (Figure 2.2E, F). Moreover, a low bone marrow blast percentage (20–49%, Figure 2.2E) and a high peripheral blood blast percentage ( $\geq 75\%$ , Figure 2.2F) resulted in a reduced and increased distance from CTRL, respectively.

Moreover, we detected increased concentration of 3-aminobutyrate and phenylalanine in the urine of AML patients compared with CTRL (Figure 2.2G, Table 2.1). Citrate, creatinine, and hippurate, which are among the most abundant urine components, showed low levels in AML, suggesting reduced excretion. Similarly, decreased glycine was indicative of reduced catabolism. Notably, two groups of patients were distinguished by serum metabolites in PC4 ( $p < 0.001$ ), and one of them included 70% of the mutated tumor suppressor gene TP53 (TP53-mut/deleted) AML (Figure 2.2H). When comparing TP53-mut/del and wild-type (wt) AML, we found lower levels of threonine and glucose in TP53-mut/del cases (Figure 2.2T), that suggested an increased cellular uptake, likely aimed at satisfying macromolecule biosynthesis and bioenergetic requirements (50), with reduced lactate excretion (51). Overall, integration of serum and urine metabolomics improved the prediction accuracy with respect to single biofluid classification.

Metabolite	Biplot name	Changes in AML vs. CTRL	Biofluid	Kruskal-Wallis <i>p</i> -value
3-Hydroxybutyrate	3HB	↑	Serum	<0.001
Glycerol of lipids	Gl-Lipids	↑	Serum	<0.001
Glucose	Glucose	↑	Serum	0.015
Glutamine	Glutamine	↓	Serum	<0.001
Lactate	Lactate	↑	Serum	0.033
Low density/very low density lipids1	LDL1/VLDL1	↓	Serum	<0.001
Low density/very low density lipids2	LDL2/VLDL2	↑	Serum	0.019
N-acetylglycoproteins (1 and 2)	NAC1/NAC2	↑	Serum	<0.001
Polysaturated fatty acids	Poly-UFA	↑	Serum	0.008
Pyruvate + Succinate	Pyruvate + Succinate	↑	Serum	0.033
Threonine	Threonine	↓	Serum	<0.001
Valine	Valine	↑	Serum	0.022
Phenylalanine	Ph-Alanine	↑	Serum/urine	0.008/<0.001
3-Aminobutyrate	3-Aminobutyrate	↑	Urine	0.006
Citrate	Citrate	↓	Urine	<0.001
Creatinine	Creatinine	↓	Urine	<0.001
Glycine	Glycine	↓	Urine	0.007
Hippurate	Hippurate	↓	Urine	<0.001

↑ : up; ↓ : down.

**Figure 2.1:** Upregulation and downregulation in metabolites responsible for AML and CTRL discrimination. Springer Nature, 2021

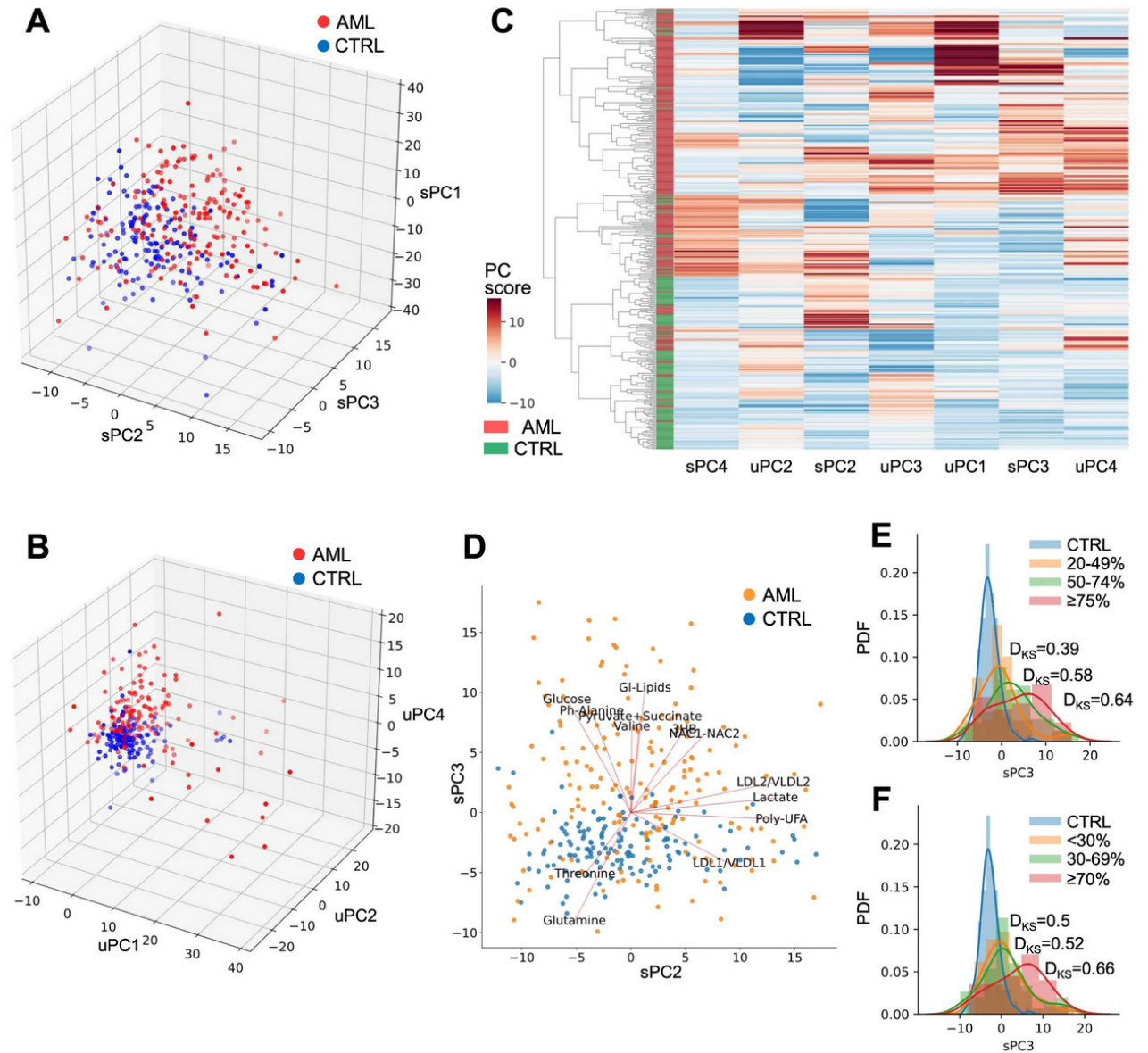
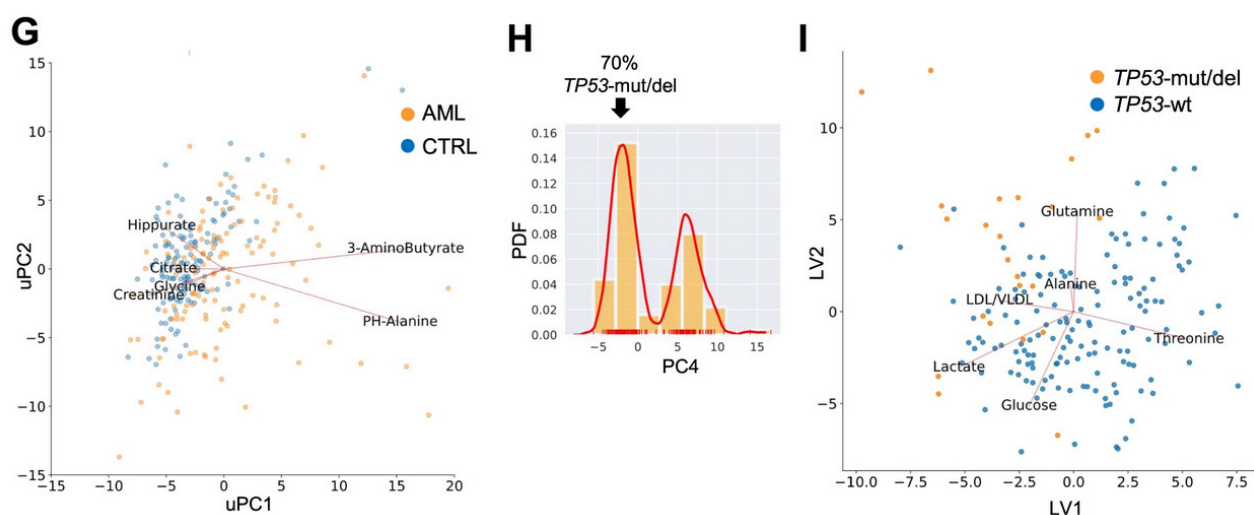


Figure 2.2: Serum and urine metabolomic profile of AML. Springer Nature, 2021



**Figure 2.2:** **A** 3D representation of principal component PC1, PC2, and PC3 projection of serum NMR data of AML and healthy controls (CTRL), which accounted for 53% of the total explained variance. **B** 3D representation of PC1, PC2, and PC4 projections of urine NMR data of AML and healthy controls. **C** Hierarchical clustering of AML and controls using integrated serum ( $n=3$ ) and urine ( $n=4$ ) PCs, selected as the best combination of predictive features by comparing an AdaBoost Classifier and a SVM Classifier. The integration yielded an enhanced coherence in adjacency between AML and controls compared with single biofluid analysis. Each component contains linear combinations of signature metabolites shown in biplots for both sera and urine samples. Colors indicate the score on each PC. **D** BiPlot on PCA reduced space of serum NMR data. Metabolites showing significant alterations ( $p < 0.05$ ) were plotted along their maximum variance direction in the PCA score space. Only completely template-matched signals were reported. **E** Estimated probability density functions (PDFs) of serum PC3 scores of AML cases according to bone marrow blast percentage (20-49%:  $p = 1.66e^{-4}$ ; 50-74%:  $p = 6.47e^{-10}$ ;  $\geq 75\%$ :  $p = 1.67e^{-15}$ ) and **F** peripheral blood blast percentage ( $< 30\%$ :  $p = 8.85e^{-8}$ ; 30-69%:  $p = 5.98e^{-10}$ ;  $\geq 70\%$ :  $p = 8.33e^{-15}$ ). The similarity between each AML blast count class and CTRL was computed using the score distribution of serum PC3, which is the latent variable best separating AML and CTRL in the metabolic latent space (DKS: absolute value of the maximal difference between the cumulative function of two distributions, representing the maximal distance between them, according to Kolmogorov–Smirnov statistics). **G** BiPlot on PCA reduced space of urine NMR data. Metabolites were plotted as in (D). **H** Serum PC4 scores in AML patients (median value: group 1, -1.94 and group 2, 6.35). **I** BiPlot on PLSDA reduced space (from a 5-PLSDA-component AdaBoost classification) for TP53-wt and TP53-mut/del AML. Metabolites were plotted along their maximum variance direction in the PLSDA score space (LV latent variables).



### **CD34<sup>+</sup> and CD33<sup>+</sup> AML cells have disrupted lipid, amino acid, nucleotide, and bioenergetic metabolism**

To obtain a complete metabolic fingerprint of AML, we performed intracellular metabolic profiling of leukemic cells (35 CD34<sup>+</sup> and 15 CD33<sup>+</sup> isolated bone marrow (BM) blasts) and compared them with 21 normal cord blood (CB) CD34<sup>+</sup> and 21 normal CD33<sup>+</sup> peripheral blood (PB) samples from healthy subjects. CD34<sup>+</sup> AML and CD33<sup>+</sup> AML segregated from their normal counterparts (Figure 2.3A, B), with a predictive accuracy of 85.7% and 94.4%, respectively, but not from each other. Among the 300 detected metabolites, 66 and 35 were down and upregulated, respectively, in CD34<sup>+</sup> AML cells, while 102 and 19 showed reduced and increased levels, respectively, in CD33<sup>+</sup> AML compared with their control group. No significant differences in metabolite levels were detected between CD34<sup>+</sup> and CD33<sup>+</sup> AML.

The top scored 30 biochemicals that distinguished CD34<sup>+</sup> AML from CD34<sup>+</sup> CB cells were primarily involved in bioenergetics, amino acid, and lipid metabolism (Figure 2.3C). Overall, 41 pathways were dysregulated in CD34<sup>+</sup> AML, with TCA cycle, D-Arginine and D-ornithine and linoleic acid metabolism showing the strongest impact (Figure 2.3D). When comparing CD33<sup>+</sup> AML and CD33<sup>+</sup> PB, the top discriminating 30 biochemicals included lipids, nucleotides, and amino acid metabolism (Figure 2.3E), with alanine, aspartate and glutamate, cysteine and methionine, purine and sphingolipid metabolic pathways showing the strongest impact (Figure 2.3F). Lipid, amino acid, nucleotide, and bioenergetic metabolism were confirmed as the most widely altered pathways when comparing the whole AML and CTRL cohorts, which were separated with a predictive accuracy of 89.1%.

### **Integrated intracellular and biofluid metabolomics highlighted alterations in the metabolism of polyamine, purine, keton bodies and polyunsaturated fatty acids and in the TCA cycle in AML**

After obtaining a distinct metabolomic profile for leukemic compared to normal CD34<sup>+</sup> or CD33<sup>+</sup> cells, we next focused on the significantly dysregulated metabolic pathways. We observed decreased arginine, methionine, and proline in leukemic cells, that suggested elevated polyamine biosynthesis (S-adenosylmethionine, 5-methylthioadenosine, and N1-acetylpermidine in CD33<sup>+</sup> and CD34<sup>+</sup> cells, respectively, Figure 2.4A), which in turn supports cell proliferation. Accordingly, the low levels of purine nucleotides (Figure 2.4B) may indicate enhanced production of adenosine 5'-triphosphate and guanosine 5'-triphosphate that are crucial for providing cellular energy and intracellular signaling, respectively (52). Tumor growth was also supported by elevated N-acetylaspartate levels in leukemic cells (Figure 2.5A) (53). Of note, in the CD33<sup>+</sup> cohort, NPM1-mut AML scored as outliers for their high levels of the N-acetylaspartate derivative N-acetyl-aspartyl-glutamate (90.9% and 25.0% of NPM1-mut AML among outliers and

non-outliers, respectively,  $p = 0.03$ ). Moreover, increased 3-hydroxybutyrate in the serum of patients and of 3-hydroxybutyrylcarnitine in leukemic cells reflected heightened ketogenesis in AML (Figure 2.5A). Polyunsaturated fatty acids (Figure 2.5B) and glucose (Figure 2.5A) were elevated in the serum of patients but reduced in CD33+ and/or CD34+ leukemic cells compared with normal ones, suggesting the need for a constant energy reservoir that is rapidly consumed by cells. The reduced levels of intracellular TCA

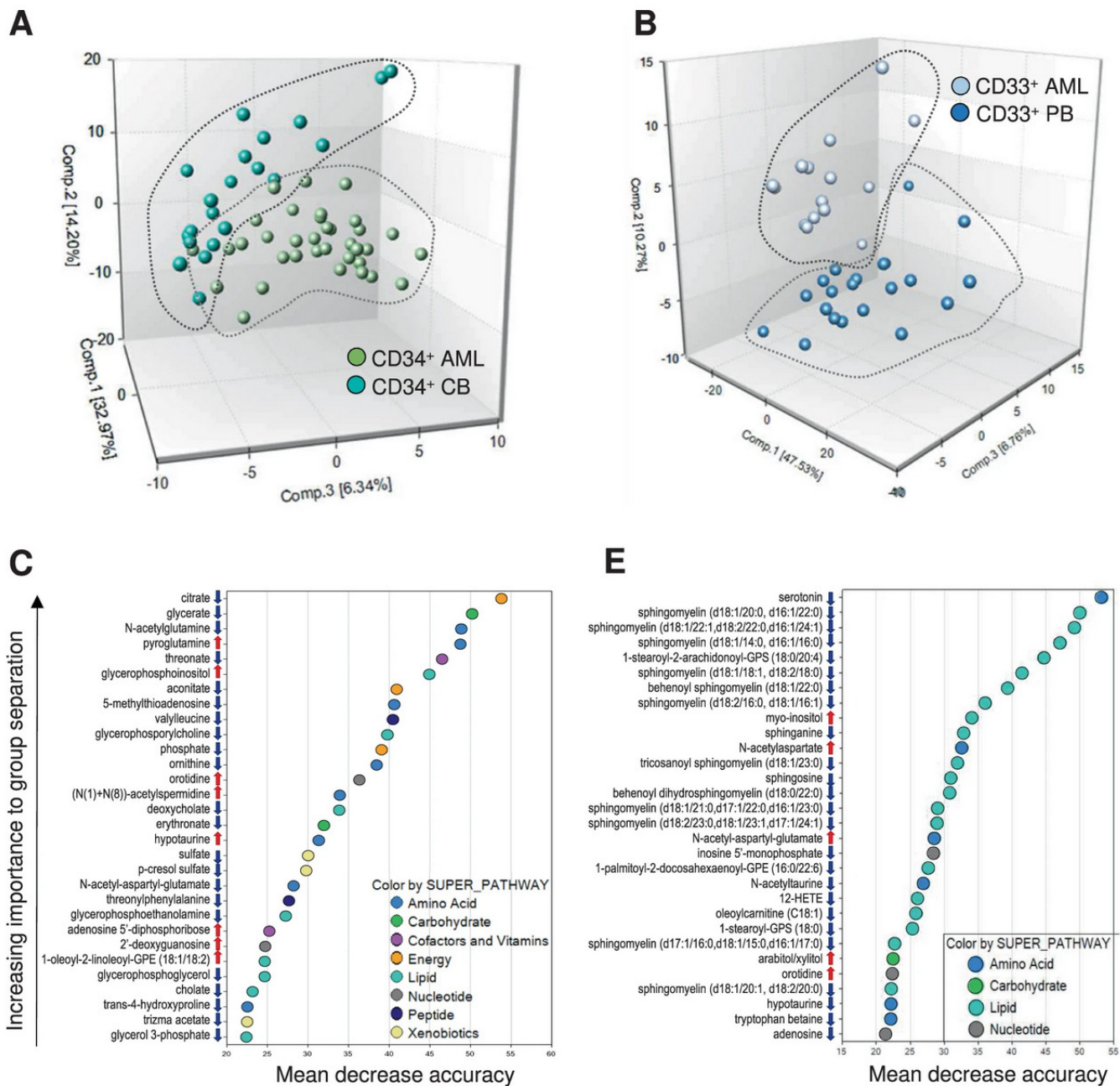
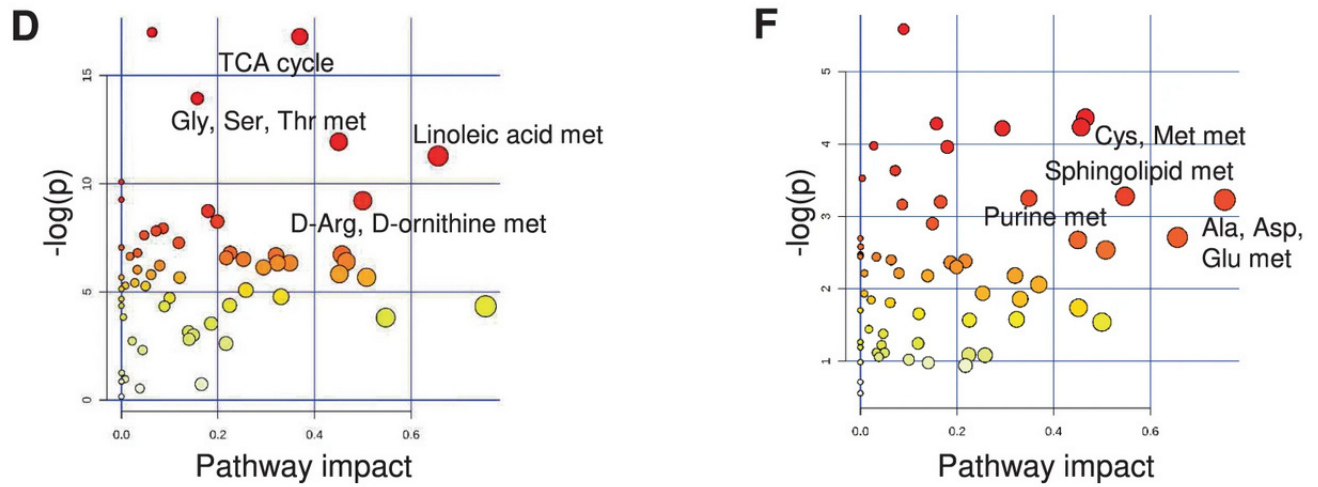


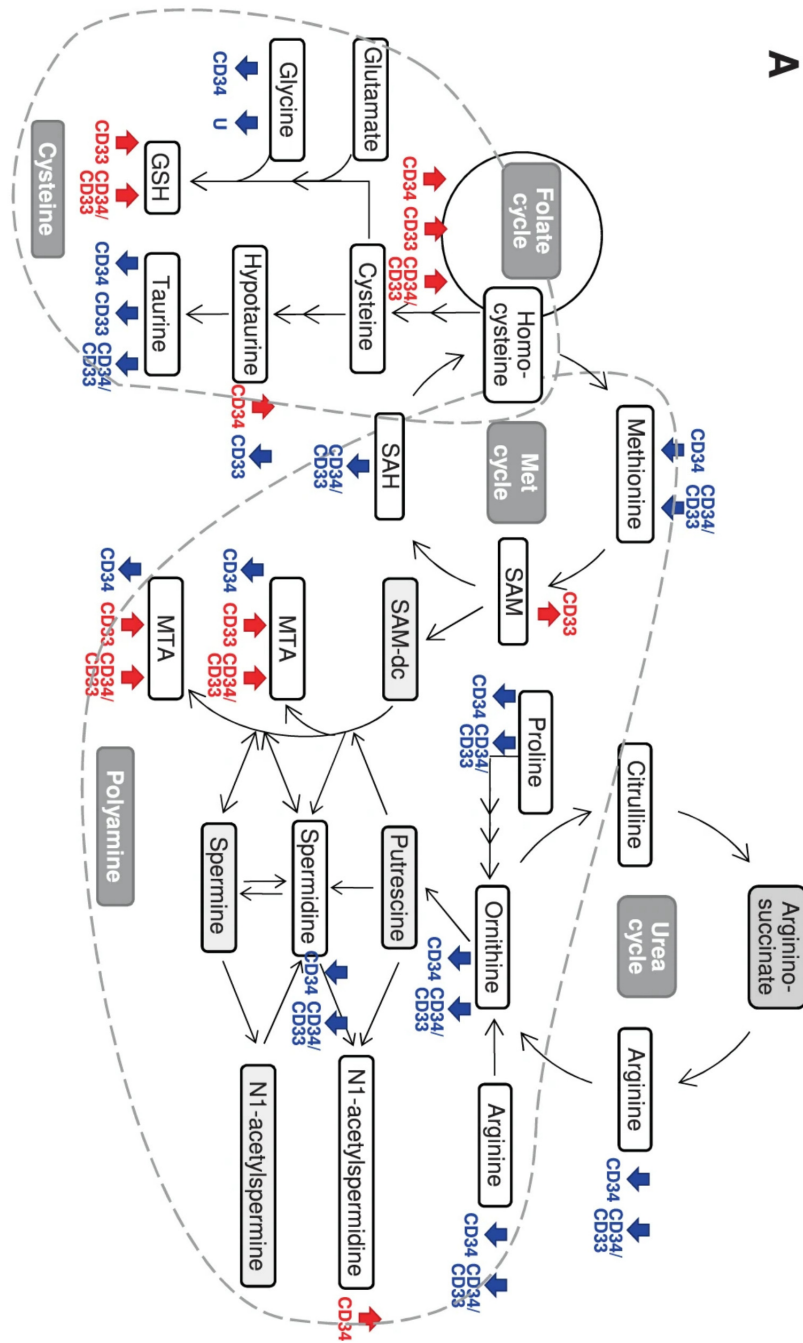
Figure 2.3: intracellular metabolomics of AML Springer Nature, 2021



**Figure 2.3:** PCA of the metabolic profile of **A** CD34+ and **B** CD33+ AML cells compared to their healthy control populations (CD34+ CB and CD33+ PB cells). **C** Biochemical importance plot of the top 30 metabolites contributing to group separation between CD34+ AML and CD34+ CB stem-progenitor cells. Red and blue arrows indicate increased or decreased metabolite levels in AML cells compared with CTRL cells ( $|\text{fold change}| \geq 2$ ,  $q \leq 0.05$ ), respectively. **D** Altered metabolic pathways in CD34+ AML cells. The most significant pathways with the strongest impact on CD34+ AML cells are shown. **E** Biochemical importance plot of the top 30 metabolites contributing to group separation between CD33+ AML blasts and CD33+ PB cells from CTRL (red and blue arrows as in (C)). **F** Altered metabolic pathways in CD33+ AML cells. The most significant pathways with the strongest impact on CD33+ AML cells are shown.

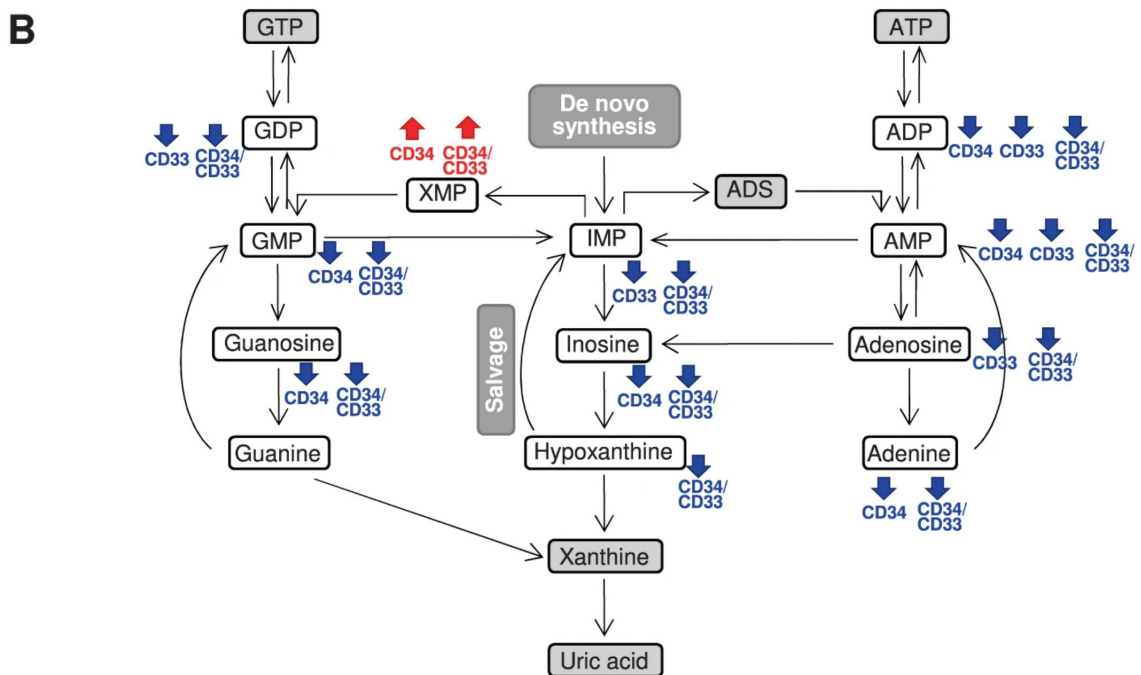
intermediates and of serum glutamine were also indicative of increased bioenergetics requirement, especially in the CD34 compartment (Figure 2.5A). This requirement was further supported by decreased levels of amino acid sources of pyruvate (e.g. threonine, glycine, serine, alanine), with a significant increase of serum lactate, an end-product of glycolysis and glutaminolysis (Figure 2.5A). In parallel, intracellular lactate levels were lower in both CD34+ and CD33+ AML than normal cells, thus suggesting a high excretion capacity .



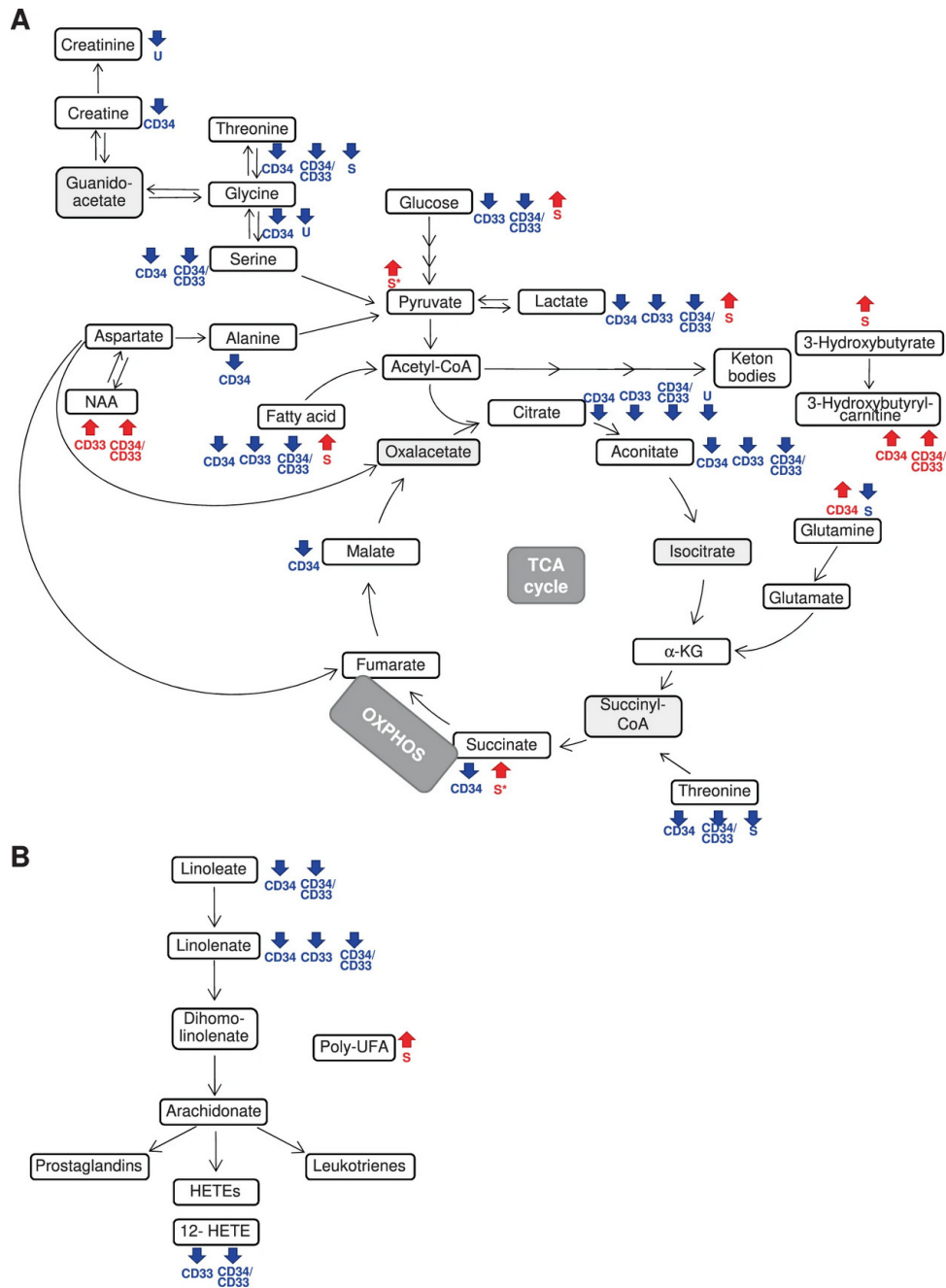


A

Figure 2.4: Schematic representation of polyamine, cysteine, and purine metabolic pathways integrating intracellular and biofluid metabolomic data. Springer Nature, 2021



**Figure 2.4:** **A** Polyamine and cysteine metabolic pathway and urea cycle. **B** Purine metabolism. Red and blue arrows/text indicate increased or decreased metabolite levels in AML cells compared with their CTRL, respectively ( $|\text{fold change}| \geq 2$ ,  $q \leq 0.05$ ) and in the urine of patients compared with CTRL ( $p < 0.05$ ). Gray metabolite boxes indicate the ones that were not detected by MS analysis. The schemes report the most relevant metabolites in the pathway according to metabolomic data (ADP adenosine 5'-diphosphate, ADS adenosine, AMP adenosine 5'-monophosphate, ATP adenosine 5'-triphosphate, dcSAM decarboxylated S-adenosylmethionine, GDP guanosine 5'-diphosphate, GMP guanosine 5'-monophosphate, GTP guanosine 5'-triphosphate, GSH reduced glutathione, IMP inosine 5'-monophosphate, MET methionine, MTA 5-methylthioadenosine, SAH S-adenosylhomocysteine, SAM S-adenosylmethionine, XMP xanthosine 5'-monophosphate).



**Figure 2.5:** **A** TCA cycle and related amino acid pathways. **B** Linoleic acid metabolism. Red and blue arrows/text indicate increased or decreased metabolite levels in AML cells versus their CTRL, respectively ( $|\text{fold change}| \geq 2$ ,  $q \leq 0.05$ ) and in the serum or urine of patients compared with CTRL ( $p < 0.05$ ). Gray metabolite boxes indicate the ones that were not detected by MS analysis. The schemes report the most relevant metabolites in the pathway according to metabolomic data (HETE hydroxyeicosatetraenoic acid, NAA N-acetylaspartate, poly-UFA polyunsaturated fatty acids, TCA tricarboxylic acid cycle). Springer Nature, 2021

### Metabolic clusters define AML subgroups with different genomic features

We classified AML cases according to their intracellular metabolic profile. Unsupervised hierarchical clustering clearly defined 3 clusters (Figure 2.6A). The top 15 metabolites that better distinguished the 3 clusters included amino acids and their derivatives (e.g. tyrosine, phenylalanine, tryptophan, threonine, lysine), intermediates of purine and pyrimidine metabolism (e.g. hypoxanthine, adenosine 5'-monophosphate, uridine) and lipids (e.g. palmitoyl sphingomyelin, cholesterol), that showed high, intermediate and low levels in cluster 1, 2 and 3, respectively (Figure 2.6B). In order to integrate genomics and metabolomics, we assigned each sample to a molecular class (27). Cluster-1 was enriched for NPM1-mut AML (50.0%), cluster-2 for cases with altered chromatin/spliceosome genes (37.5%), and cluster-3 for TP53-mut/aneuploid AML (34.4%,  $p = 0.023$ , Figure 2.6C). We then investigated differences at serum and urine level across genetic categories (chromatin/spliceosome-mut, NPM1-mut, TP53-mut/aneuploid AML,  $n=71$ ) and identified 4 NMR clusters (Figure 2.6D). Genomic categories associated with specific biofluid metabolic cluster (clusters 2, 3, and 4,  $p = 0.040$ , Figure 2.6E), in accordance with the intracellular metabolic profiles. High levels of serum tyrosine, threonine, and citrate correlated with the cluster enriched for chromatin/spliceosome-mut. Viceversa, low levels of these metabolites were detected in the cluster enriched for NPM1-mut (Figure 2.6F). The cluster associated with TP53-mut/aneuploid AML displayed intermediate threonine and tyrosine levels and high citrate in the serum compared to the other two clusters. Notably, tyrosine and threonine showed high intracellular levels in the NPM1-mut enriched cluster compared with the other clusters (mean decrease accuracy = 0.010 and 0.005, respectively, Figure 2.6B), suggesting an increased intracellular need and/or uptake leading to serum depletion.

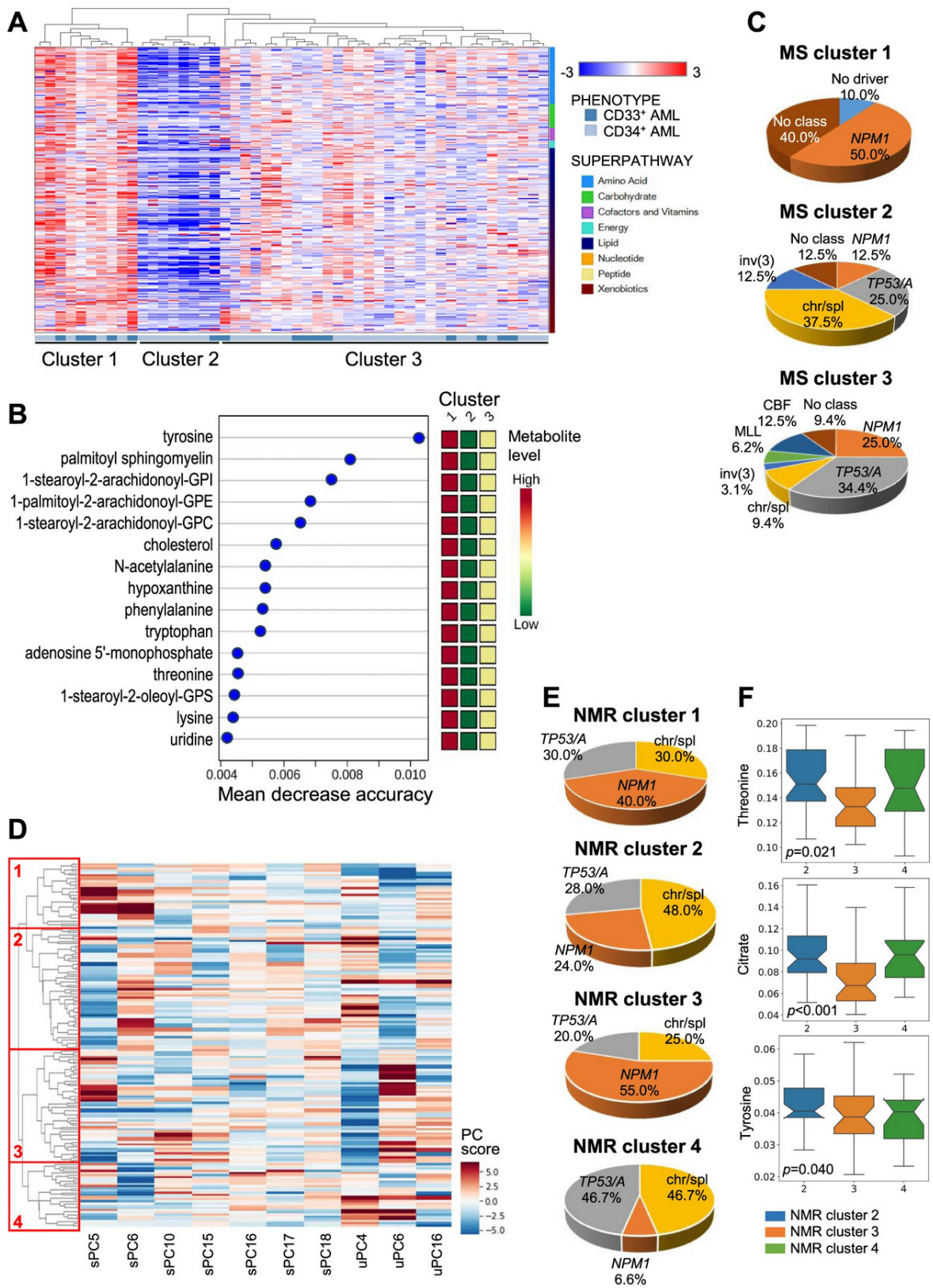


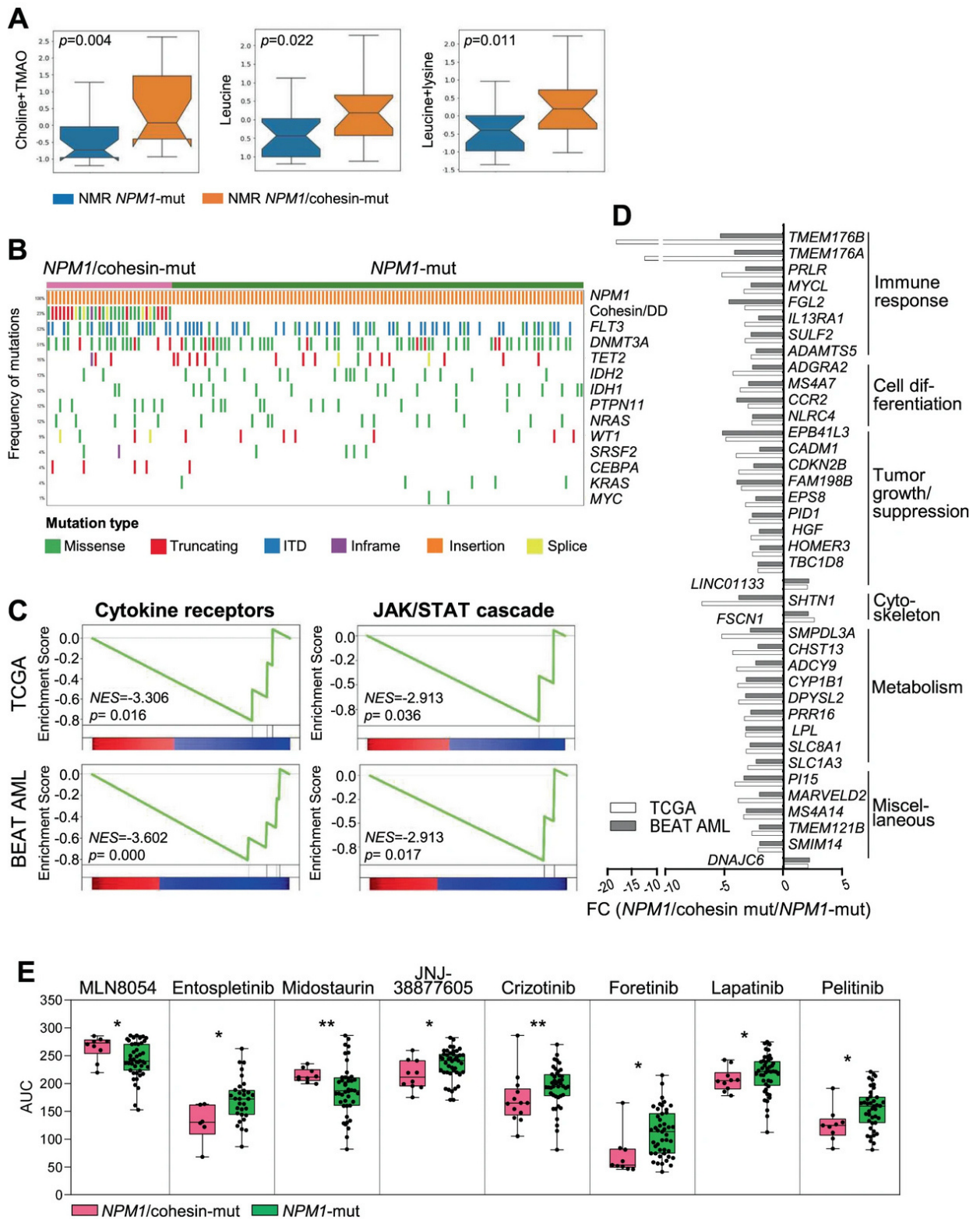
Figure 2.6: Intracellular and biofluid metabolomics show association with AML molecular classification. Springer Nature, 2021



**Figure 2.6:** **A** Unsupervised hierarchical clustering of AML according to intracellular metabolomic profiles (MS, each row denotes a metabolite, each column a sample). **B** Top 15 metabolites contributing to separation of the three MS metabolic clusters (1, 2, 3). The metabolites belong to the following superpathways: amino acids and their derivatives (tyrosine, N-acetylalanine, phenilalanine, tryptophan, threonine, lysine), intermediates of purine and pyrimidine metabolism (hypoxanthine, adenosyne-5'-monophosphate, uridine) and lipids (sphingolipid, phosphatidylinositol, phosphatidylethanolamine, phosphatidylcholine, cholesterol, phosphatidylserine). Colored squares on the right indicate metabolite levels in each cluster. **C** Molecular classification of MS metabolic clusters (27). Due to the low number of t(8;21) and inv(16)/t(16;16) cases, they were grouped in the core-binding factor (CBF) category and a t(6;9) patient with complex karyotype was included in the TP53/aneuploidy category (NPM1 NPM1-mut, chr/spl chromatin/spliceosome-mut, TP53/A TP53-mut/aneuploidy, inv(3) inv(3)/t(3;3), KMT2A KMT2A-rearranged). **D** Hierarchical clustering of AML patients belonging to the NPM1-mut, chromatin/spliceosome-mut or TP53/aneuploidy molecular classes according to biofluid metabolomic profile (NMR). These components were selected as the combination of urine and serum spectral features that best described the above mentioned genomic stratification. Of the ten features selected via stochastic gridsearch, seven came from serum spectra, indicating serum as the principal vector of information for this particular stratification. Colors indicate the score on each PC. **E** Molecular classification of NMR metabolic clusters. **F** Top scoring serum metabolites separating NMR clusters 2, 3, and 4. Signature metabolites were extracted from sera samples by selecting the highest scoring signals in terms of presence amongst the sera PC responsible for the best separation of molecular subgroups and the average of absolute values of their loadings. Statistical significance was obtained with SciPy.Stats Kruskal–Wallis H-test using stepdown Sidak correction. Notch width corresponds to the confidence interval of the median.

### NMR-driven metabolic classification identifies two subgroups of NPM1-mut patients

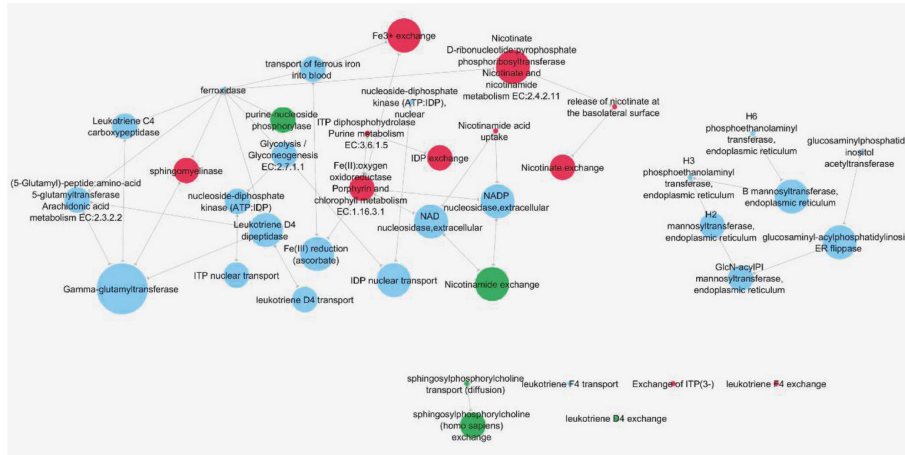
Our data so far described a significant association between genomic and metabolic profiles. However, even within the same genomic category, different subgroups can be identified according to combinatorial mutation patterns and consequently they may show metabolic differences. This hypothesis was confirmed in patients carrying NPM1 mutations, in whom the metabolic profiles defined two distinct subgroups. NPM1-mut patients with higher serum levels of choline+trimethylamine N-oxide, leucine and leucine+lysine ( $p < 0.05$ , Figure 2.7A) were enriched for co-occurring mutations in cohesin complex and DNA damage genes (SMC1A, SMC3, RAD21, STAG2, ATM, ATR, BRCA2, named NPM1/cohesin-mut), compared with NPM1-mut patients from the other metabolic group (60.0% versus 9.1% of cases, respectively,  $p = 0.024$ ). To gain insights into molecular mechanisms associated with the metabolic differences between NPM1/cohesion-mut and NPM1-mut AML, we analyzed paired exome and transcriptome data from the TCGA and BEAT AML datasets for the same genetic subgroups. Twenty-three percent of NPM1-mut AML (32/137) also carried at least one alteration in recurrently mutated genes (25) belonging to the cohesin complex or DNA damage pathways. Compared with NPM1-mut AML, NPM1/cohesin-mut cases were characterized by a lower white blood cell count (39.7 vs. 64.1 cells/mm<sup>3</sup>,  $p = 0.006$ ) and a significantly higher mutation load (average mutation number: 15 vs. 9,  $p < 0.001$ ), with lower frequency of IDH1-2/TET2 mutations (21.9% vs. 46.7% of NPM1-mut,  $p = 0.014$ , Fig. 6B). FLT3 alterations were evenly distributed between the two groups (Figure 2.7B) and no differences were observed in clinical outcome. At transcriptional level, signatures of cellular response to cytokines and JAK-STAT cascade were significantly downregulated in NPM1/cohesin-mut AML (Figure 2.7C). Accordingly, NPM1/cohesin-mut AML showed reduced expression of genes involved in the regulation of immune and inflammatory response, along with others related to cell differentiation and metabolism (Figure 2.7D). We then compared the ex vivo response of NPM1/cohesin-mut and NPM1-mut AML to a panel of targeted agents (n=122, BEAT AML (25)). NPM1/cohesin-mut AML showed decreased sensitivity to the Aurora kinase A inhibitor MLN8054 and the FLT3/JAK inhibitor Midostaurin but responded better to SYK, MET, and EGFR inhibitors (Entospletinib, JNJ-38877605, Crizotinib, Foretinib, Lapatinib, Pelitinib, Figure 2.7E). These data suggest that the co-occurrence of different mutations with altered NPM1 may confer a distinct metabolic, transcriptomic, and drug sensitivity profile to the leukemic cells.



**Figure 2.7:** Metabolic, genomic, transcriptomic and drug response differences between *NPM1*/cohesin-mut and *NPM1*-mut AML. Springer Nature, 2021



**Figure 2.7:** *A* Serum metabolites separating NPM1/cohesin-mut and NPM1-mut AML (TMAO trimethylamine-N-oxide). *B* Oncoprint of mutations in AML-related genes (frequency > 3% in the overall population) in NPM1/cohesin-mut and NPM1-mut AML. WES data were obtained from the TCGA (n=13 NPM1/cohesin-mut, n=33 NPM1-mut) and BEAT AML (n=19 NPM1/cohesin-mut, n=72 NPM1-mut, including 7 relapse cases) cohorts. Rows denote genes or groups of genes (cohesin/DD cohesin/DNA damage-related genes). Columns represent frequency of mutations and single patients (ITD internal tandem duplication). *C* Signatures of cytokine receptors and JAK-STAT cascade from GSEA showing significance in both datasets (TCGA, left to right: cytokine–cytokine receptor binding, regulation of JAK-STAT cascade, n=9 NPM1/cohesin-mut, n=25 NPM1-mut; BEAT AML, left to right: cytokine receptor activity, JAK/STAT cascade, n=14 NPM1/cohesin-mut, n=47 NPM1-mut, including 3 relapse cases). *D* Genes involved in immune response, cell differentiation, tumor growth regulation, cytoskeleton, metabolism and other cellular processes, showing a significantly different expression between NPM1/cohesin-mut and NPM1-mut AML in both cohorts. *E* Area under the curve (AUC) for the drugs showing a significantly different response between NPM1/cohesin-mut and NPM1-mut AML was plotted for the two cohorts (NPM1/cohesin-mut, n=6–13; NPM1-mut, n=31–45) (25): MLN8054 (Aurora kinase A inhibitor), Entospletinib (SYK inhibitor), Midostaurin (FLT3, JAK inhibitor), JNJ-38877605 (MET inhibitor), Crizotinib (ALK, MET, ROS1, NTRK inhibitor), Foretinib (MET, KDR, TIE inhibitor), Lapatinib (ErbB-2, EGFR inhibitor), Pelitinib (EGFR inhibitor). Boxes represent the mean (horizontal line) and extend from the 25th to 75th percentiles; whiskers extend from the minimum to the maximum value and each value is plotted (\* $p \leq 0.05$ , \*\* $p \leq 0.01$ ).



**Figure 2.8:** Modeling the metabolic network of NPM1/cohesin-mut AML. NPM1/cohesin-mut specific metabolic reaction perturbation network. (red: minimum flux, green: maximum flux, light blue: no information among NPM1/cohesin-mut-specific alterations). Sizes of nodes are proportional to links originating from that node and pointing towards others (outdegree). Springer Nature, 2021

### Predicting metabolic specificities of NPM1/cohesin-mut AML

Seven downregulated genes in NPM1/cohesin-mut compared with NPM1-mut AML encoded for enzymes involved in nucleotide (ADCY9, DPYSL2), lipid (LPL) and carbohydrate (CHST13) metabolism, energy production (CYP1B1) and transporter/exchanger. We thus modeled the consequences of gene expression alterations of NPM1/cohesin-mut AML on the intracellular metabolome by reconstructing genome scale metabolic network models. Based on the analysis of diverse cellular models and our MS data, we selected a hematopoietic model derived from Recon2. The selected reconstruction was validated by modeling the effect of IDH mutations. We first predicted the changes in metabolic fluxes induced by the altered expression of enzymes between NPM1-mut and NPM1-wt AML. Interestingly, among the perturbed metabolites, experimental evidence confirmed increased N-acetylaspartate and glutamine, reduced spermidine levels (among others) in NPM1-mut compared with NPM1-wt AML. We then simulated the intracellular metabolome of NPM1/cohesin-mut AML by adding the 7 downregulated genes to the model. Eleven metabolites and 42 reactions were predicted to be specifically perturbed in the NPM1/cohesin-mut model. A metabolic network reconstruction of the altered reactions showed a major cluster centered on nicotinate, nicotinamide, and inosine exchange/modification, with connections to glycolysis and metabolism of leukotriene inflammatory molecules (Figure 2.8), that were also confirmed by pathway enrichment analysis of genes catalyzing the network reactions. Notably, NPM1/cohesin-mut AML showed lower intracellular levels of inosine-5'-monophosphate and glucose when compared with NPM1-wt and/or NPM1-mut AML. Overall, our multistep approach defined the metabolic specificities of NPM1/cohesin-mut AML.

#### 2.1.4 Study conclusions

Few studies have previously analyzed the metabolic profile of AML patient serum (54; 55; 56; 57; 58) or of a limited number of primary cells (41; 59; 60). Here, we have performed integrated genomics and metabolomics analysis in AML, which showed genetic-related differences in the metabolic profiles and defined multiple subgroups with distinct constellations of mutations and metabolic features. First, integrated serum and urine analysis accurately discriminated between AML and normal patients, suggesting a robust approach for evaluating disease metabolic subgroups and a valid, low-cost approach for noninvasive population analyses.

Second, we integrated biofluid and intracellular metabolomics. We used NMR and MS as complementary techniques for biofluids and primary cell profiling, respectively. The rationale of this approach is twofold: it allowed us to benefit from the reproducibility of NMR, which offers unbiased information and could enable a rapid translation to the clinical practice, and from the high sensitivity of MS in metabolite detection from low cell numbers. Our comprehensive view showed alterations in the TCA cycle and in the metabolism of purine nucleotides, amino acid, fatty acids, keton bodies, polyamine, glutamine and other amino acids. Notably, many of the identified pathways can be therapeutically exploited (e.g. glutaminolysis, arginine uptake, aspartate production, fatty acid oxidation, polyamine metabolism, ketogenesis) and the inhibition of some of them achieved promising results in AML (41; 61) or in cancer (62; 63) models.

When integrated with genetic features, the metabolic profiles showed association with NPM1-mut, chromatin/spliceosome-mut and TP53-mut/aneuploid AML classes. Our data also classified NPM1-mut AML carrying mutations in cohesin or DNA damage-related genes as a distinct metabolic subgroup. This group does not associate with IDH1–2-TET2 mutations, which are also frequently observed in NPM1-mut cases (25; 26; 64) but it was characterized by higher mutation burden, lower white blood cell count and downregulation of immune-related genes (65). Accordingly, in silico modeling of the NPM1/cohesin-mut-specific metabolic perturbations predicted changes in the balance of leukotrienes. Moreover, flux and network analysis based on the identified transcriptomic changes pointed at alterations in the purine and NAD superpathways as NPM1/cohesin-mut-specific ones. Inosine-5'-monosphate, an intermediate in the purine metabolism, showed low levels in NPM1/cohesin-mut compared with NPM1-mut AML cells. With regard to therapeutics, NPM1/cohesin-mut AML were more sensitive than NPM1-mut AML to EGFR inhibition, which may lead to the release of the differentiation brake (66) and to drugs targeting the tyrosine kinase receptor MET, likely due to a mild autocrine pathway activation in these cases, who express low levels of the ligand (67).

Overall, our results provide a map of the crosstalk between metabolic pathways and between genomics and metabolomics in AML, reflecting functional interactions and dependencies that could be therapeutically exploited and provide the rationale for a switch to a genomic- and phenotypic-driven personalized medicine

## 2.2 Modelling Dynamics of Metabolic States

It has been shown in the last section that metabolomics of biofluids and cells, when joined with other omic data in a framework to characterize the crosstalk between the different snapshots of the molecular states of a system, can find accurate fingerprints of perturbations that can in turn help important etiologic conclusions about the system itself. An interesting question that arises from the results achieved by fingerprinting approaches (in the study previously shown but also from the recently flourishing literature on the topic), is how to study the response of a system after a perturbation and which information can be obtained from the molecular states it crosses when trying to get back to equilibrium. For instance, can we obtain the molecular, metabolic fingerprint of a pathology from a group of individual and evaluate how these individuals respond to a therapy, that can be seen as an additional perturbation that forces the system to move from its pathological state? Answering such a question can ultimately lead to useful applications in phenotyping the response over time of different (classes of) individuals to drugs, therapies, diets, habits and environmental factors. This in turn can translate to more individually-tailored therapeutic indications, that can increase the odds of favorable outcomes especially after exposition to oncologic pathologies. In the following section, we report the preliminary result of a biofluid NMR metabolomic based framework that is being developed to address the study of the evolution of fingerprints in the metabolomic data space over time.

### 2.2.1 Metabolomic evaluation of therapeutic response in breast cancer

The data analyzed in this early-stage study come from an ongoing clinical trial designed in cooperation with the *Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori"*. The study is designed to monitor the ongoing metabolomic conditions of serum and urines from a cohort of women that underwent breast cancer surgery.

#### Samples processing and analysis

Urine and serum aliquots were stored at  $-80^{\circ}\text{C}$  until their use for the NMR analysis. 630  $\mu\text{L}$  of urine sample were centrifuged to remove debris, then 540  $\mu\text{L}$  of supernatant were placed in a clean Eppendorf containing 60  $\mu\text{L}$  of D<sub>2</sub>O-based phosphate buffer containing also trimethylsilyl propionate (TSP) as Internal Std and sodium azide (NaN<sub>3</sub>) as an antibacterial agent. A total of 590  $\mu\text{L}$  of the mixture was transferred into  $5 \times 178$  mm (7") 5 mm, outer diameter NMR tubes (for Bruker Match holder). <sup>1</sup>H NMR spectra were recorded at 298 K with an AVANCE spectrometer (Bruker BioSpin, Karlsruhe, Germany) operating at a proton frequency of 600.13 MHz, equipped with an autosampler with 60 holders. The HOD residual signal was suppressed by applying the NOESYGPPR1D sequence (a standard pulse sequence included in the Bruker library) incorporating the first increment of the NOESY pulse sequence and a spoil gradient. Each spectrum was ac-

quired using 32 K data points over a 7211.54 Hz spectral width (12 ppm) and summing up 128 transients. A 90° pulse of 12.5  $\mu$ s was set up. A delay of 5 s between transients, extending the acquisition time of 2.27 s, was chosen to provide a recycle time 5 times longer than the longitudinal relaxation time of the protons under investigation, expected to be not longer than 1.4 s. The data were Fourier transformed and phase and baseline corrections were automatically applied using TopSpin version 3.0 (Bruker BioSpin, Karlsruhe, Germany). Signals were assigned by comparing their chemical shift and multiplicity with Chenomx software data bank (version 8.1, Edmonton, Canada). Analysis of Spectra. Spectra were exported in ASCII file format and then imported into R software (version 3.3.2). Chemical shift referencing was performed by imposing the TSP signal to 0.00 ppm. The spectral regions including only noise (e.g., the spectrum edges below 0.5 and above 10 ppm), as well as the data points which are strongly affected by the residual water (between 4.95 and 4.7 ppm) and the urea signals (5.45-6.1 ppm) were removed prior to data analysis. Normalization was carried out using the PQN algorithm. Spectra were bucketed in a total of 413 spectral features.

### Cohort and dataset

The first batch of samples for which the framework has been tested refers to 60 patients and 20 healthy controls. To monitor the evolution of metabolomic fingerprints in patients, samples of serum and urine has been drawn at different time points, 3 pre and 6 post surgery:

- **Pre-intervention samples:** 14 days prior the surgery (**T14P**), 7 days prior the surgery (**T7P**), 1 day prior the surgery (**T1P**)
- **Early post-intervention samples:** 14 days after the surgery (**T14D**), 28 days after the surgery (**T28D**)
- **Late post-intervention samples:** 6 months after the surgery (**T6D**), 12 months after the surgery (**T12D**), 18 months after the surgery (**T18D**), 24 months after the surgery (**T24D**).

To avoid misinterpretation of results and enhance robustness, each classification task has been k-fold crossvalidated with stratifications to maintain proportion between sample classes and patient age ranges as much as possible. Spectra from the late-post intervention samples were not considered in the present model, to avoid possible confounding effect introduced by long-term pharmacological treatments (i.e. chemotherapy). After preprocessing, filtering and outlier checks, metabolic profiles and therapy response trajectories were computed for a total of 45 patients using serum samples and a total of 39 patients using urine samples.

### Therapy response trajectory extraction

The rationale behind the experimental design is to somehow define a therapy response trajectory in the metabolomic space, defined by the molecular signature of breast cancer, for each patient. To achieve this, we developed a framework that can be summarized in the following steps :

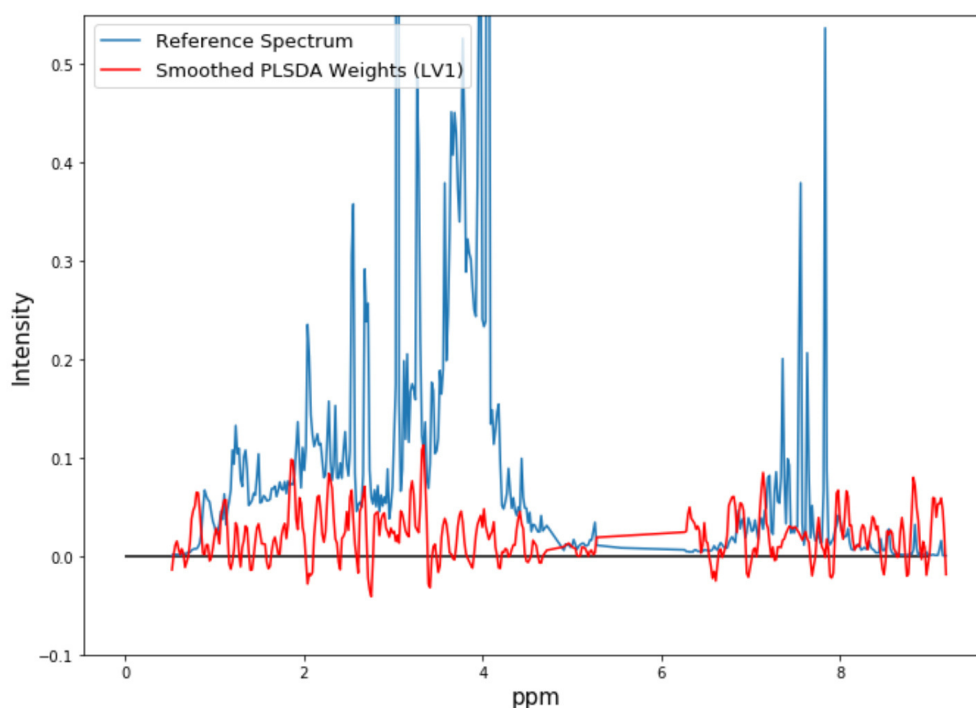
1. Define the molecular signature of breast cancer with the best possible projection on latent structures that separates healthy control spectra from pre-intervention spectra.
2. Project post-intervention spectra into this space, using the transformation coefficient of the previously trained model.
3. Compute, for each patient, the centroid of its pre-intervention spectra and post-intervention spectra in the metabolomic space. This will characterize each individual in the metabolic space, as a function of its variations from the molecular signature of its pathological state.
4. Compute the centroid of healthy control spectra to find the point in the metabolomic space that is assumed to define an ideal response to therapy. Compute the vector that joins the centroid of pre-intervention spectra with the centroid of healthy controls for each patients. This is the ideal therapy response vector.
5. Compute the vector that joins the centroid of pre-intervention spectra with the centroid of post-intervention spectra for each patient. The norm of this vector defines the intensity of the therapy response.
6. Perform cosine similarity analysis (68) between the vector computed in step 4 and the vector computed in step 5. This will yield a measure of how much the response direction of each patient is similar to its ideal response direction.
7. Cluster patients based on the features computed in step 5 and 6. The common characteristics of patients belonging to the same cluster, defined by therapy response features, can be studied to phenotype different populations of patients.

### 2.2.2 Early-stage results and discussion

#### Serum and urine lower dimensional metabolic spaces

Using an AdaBoost framework based on decision trees, many concurrent projection on latent structure models were trained. The resulting best model is a projection in a 5-dimensional PLS space, whose latent component reached an average validation fold accuracy of  $\sim 0.8$  in separating T1P spectra and healthy controls spectra (Figures 2.9, 2.10). Although limited by the sample size at the stage of the study, that did not allow for a robust identification of a great number of metabolites, interesting information emerges from



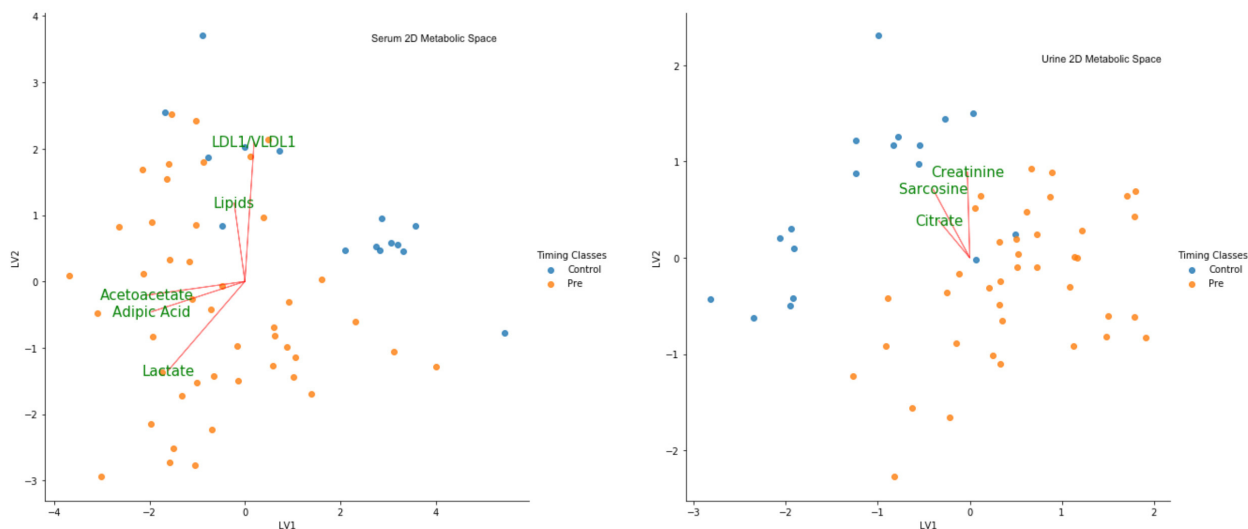


**Figure 2.9:** *Weights of the first PLS latent variable of the best classification model, highlighting importance zones of serum spectra for breast cancer fingerprinting.*

the metabolomic signature of breast cancer (Figure 2.10). Serum metabolomic space yields information about a possible perturbation of lactate metabolism, as well as on the entire fatty acid metabolism, from beta-oxidation to ketogenesis (through the presence of acetoacetate). Urine highlights a perturbation of TCA cycle and glycolysis through an higher presence of citrate in controls urine with respect to T1P urine. Perturbation of lactate and energetic metabolism are often found as common background of cancer cells, that rely on anomalous pathways for energy production (especially glycolysis) with respect to healthy ones. Interestingly, the presence of sarcosine in the metabolomic space of urine points to a possible perturbation of choline-glycine metabolism. Furthermore, sarcosine is being investigated as a possible marker of breast cancer subtypes (69).

### Therapy response analysis in the metabolomic space

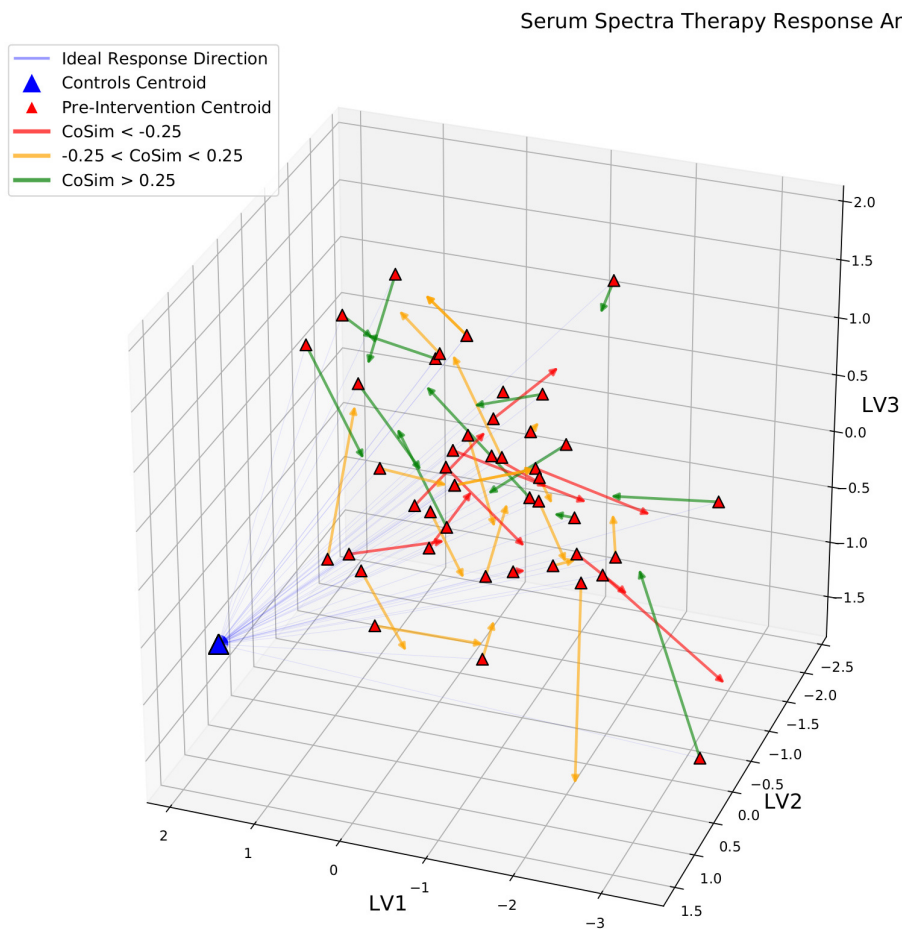
Due to urine spectra being generally more sensitive to drug intake, time of collection during the day and variability in general and given the sample size at this stage of the study, we focused on modeling therapy response based on serum spectra. Following the steps described in 2.2.1, we obtained a summary of therapy response analysis (Figure 2.11) projected in the 3-D extension of the serum space computed in 2.10. Different patients obviously exhibit difference therapy response, both in terms of direction of the response (where surgery and treatment has caused a patient's metabolomic state to move) and its



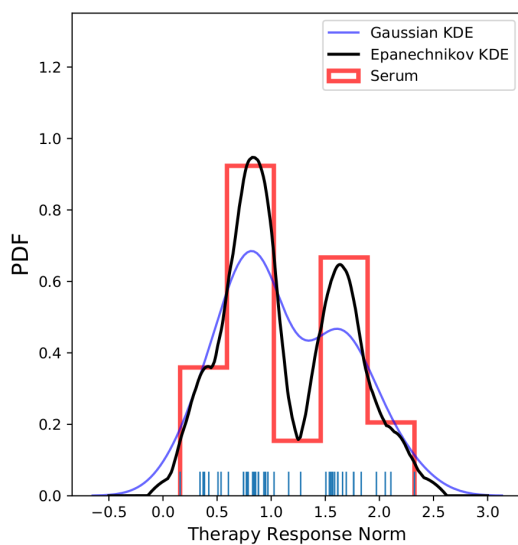
**Figure 2.10:** 2-D projection of TIP spectra and healthy controls spectra on the first two latent components of the best dimensionality reduction model. The metabolic serum space contains information about lactate and fatty acid metabolism (LDL/VLDL = low density, very low density lipids). The urine serum space contains information about TCA cycle and creatinine metabolism, as well as info on choline metabolism through the presence of sarcosine.

intensity (how much surgery and treatment has caused a patient's metabolomic state to move). To assess a clearer trend in therapy responses, we focused on the characterization of patients with their distribution analysis and clustering based on therapy response features. Just by looking at the distributions of therapy response norms in serum 2.12, a modelling of its probability density function (PDF) with two different kernel density estimation (KDE) methods, a distinct bimodal trend appears. This suggests the presence of at least two distinct type of patients, as far as the intensity of therapy response is concerned. Encouraged by these results we performed a hierarchical clustering of patients using the two surgical therapy response features 2.13. The individuation of the two clusters confirms the bimodal nature of the distribution of therapy response norms, highlighting the existence of at least two macro classes of patients that exhibit differences (both in intensity and direction) in therapy responses. The investigation of the etiology of this stratification is the next objective of this ongoing study: the aim is to find correlation of this stratification with cytokines response, clinical variables and other omic data regarding molecular state that are currently being collected from the cohort.

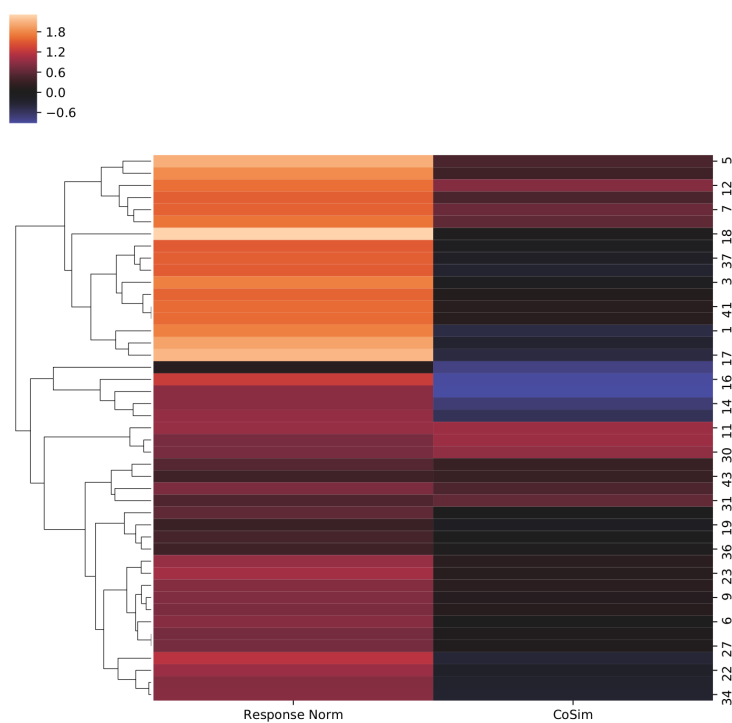




**Figure 2.11:** Visualization of different responses to the therapy of each patient. Arrow colouring is based on the categorization of direction of responses based on cosine similarity (CoSim) values. CoSim values spans from  $(-1, 1)$ , with  $\text{CoSim}=1$  denoting a parallel direction with concordant sense with respect to the ideal response vector (translucent blue lines);  $\text{CoSim} = -1$  denoting a parallel direction with opposite sense with respect to the ideal response vector;  $\text{CoSim} = 0$  denoting an orthogonal direction with respect to the ideal response vector.



**Figure 2.12:** Probability density function (PDF) estimation of the therapy response vectors norms distribution. Two different KDE methods highlight bimodality, suggesting a stratification in patients with respect to therapy response.



**Figure 2.13:** Hierarchical clustering of patients based on therapy response norm and cosine similarity from serum metabolomes. Two distinct clusters appear, suggesting a stratification of patient that must be investigated.

### **2.3 Chapter Conclusions**

In this chapter we outlined possible ways to study biochemical complex system using patterns of omic data. In particular, approaches based on fingerprinting the molecular state of a system, at various level of resolution, and their response to a perturbation such as a disease, are offering important results and opening many possibilities for modelling. Modelling the crosstalk of molecular patterns descriptors of the compartment of a system at different resolutions, often leads to important etiologic conclusions about aspects of extremely complex systems, such as the human organism. The studies reported and the frameworks implemented and discussed, highlight two important aspects in modelling the exposure of the human organism to pathologic perturbations: the possibility of linking of systemic and cellular level metabolomics to the simulation of the enzymatic level through genomics, and a possible way to study individual variability in responding to perturbations over time. In the next chapters, these premises will be exploited to study the exposure of the human organism to another key perturbation, that requires a similar holistic paradigm of description : diet.



# Macroscopic Exposures: Epidemiological Data Analysis and Physiological Outcomes

## 3.1 Studying health, lifestyle and diet: a major challenge

The fate of an inherently holistic approach, that is required to shed light upon the interaction between bio-molecules and the human organism, rooted at many levels in heuristic methodologies, is obviously tied at its most macroscopic level to the quality of epidemiological data. The drive to develop cheap and portable systems for health and lifestyle data collection, together with a growing interest in the parametrization of nutritional habits is linked to the goal of more personalized medical interventions and suggestions. To this end, a great deal of data types and their relationships are being explored in several fields of life sciences. Classical health indicators, environmental exposures, nutritional and cultural factors are all studied to stratify and classify free living populations, in an effort to phenotyping individuals and predict all sorts of clinical and physiological outcomes. However, besides for a small cluster of known factors with an almost deterministic detrimental effect on health (smoking, drinking, certain chemical exposures...) shedding light upon health risk factors and their health-affecting mechanisms is a challenging matter. The presence of a plethora of confounding factors, including genetic (both genotypic and phenotypic) and epigenetic variability of individuals, hinders the results of even the largest epidemic studies in free living populations; correlation, causality and etiology of health risk factors are often difficult to determine. This holds particularly true when trying to determine the effects of specific diet patterns or nutritional interventions on health conditions, giving rise to a sort of paradoxical outcome. On one hand, it is established that diet has a major influence on health: optimal eating is associated with increased life expectancy, reduction in lifetime risk of chronic diseases, enhanced gene expression, while bad eating habits are associated with leading causes of chronic diseases (70). On the other hand, there have been no long-term studies comparing diet patterns using rigorous statis-

tical methods for bias and confounding factors suppression. However, even without a comparison of strictly defined diet patterns, literature strongly supports a background set of healthful eating habits: minimally processed foods, predominantly plants, is associated with health promotion and disease prevention and is coherent with key elements of seemingly distinct dietary approaches (71). Moreover, the health impact of dietary patterns has a strong dependence with the metabolic state of individuals. Clinical conditions, genetic traits, sex, age, physical activity are all key determinants of metabolic features such as daily energy expenditure (72). To complete an already confused picture, there is a well documented source of concern about the accuracy of canonical methods of dietary assessment, such as food frequency questionnaires (FFQ) and 24 hour recall questionnaires (24Hr) (73; 74).

Thus, due to difficulties in stratifying individuals and taking confounding factors into account, ever-shifting paradigms in diets and dietary patterns definition, an uncertainty about how efficiently dietary assessment (FFQ, 24Hr) can support etiologic studies, the links between health and nutrition are still largely unexplored. The work proposed in the next sections is focused on taking some of the challenging aspects of epidemiological data analysis in nutritional studies on: a rigorous statistical learning model to study diets and eating habits as patterns of macronutrient intakes and their correlation to clinical outcomes, based on state of art reference for nutritional data handling and standardization, in a cohort of subjects at risk of metabolic syndrome (MetS).

## **3.2 Correlations between inadequate Energy/Macronutrient intake and clinical alterations: the importance of stratification and model selection**

This section is entirely based on the published work by Danesi, Mengucci et al. (75).

### **3.2.1 Introduction to MetS**

Metabolic syndrome (MetS) is a pathologic condition including a cluster of components such as hypertension, dyslipidaemia, insulin resistance, hyperinsulinemia, glucose intolerance, and obesity, in particular central obesity (76). MetS represents an epidemic clinical condition in countries where obesity and Western, unhealthy dietary patterns prevail, and its development is associated with both non-modifiable and modifiable risk factors as low physical activity and a poor-quality diet (77). Currently, lifestyle-based interventions aimed at normalising body weight (BW) and controlling lipid levels, glucose sensitivity, and blood pressure are the most effective preventive approaches to MetS. Although available evidence suggests certain nutrients, foods, and dietary patterns have beneficial effects on MetS, there is no definitive agreement on which nutritional strategy is the most effective (78; 79). The association between different eating patterns and the MetS components has been evaluated in several studies (80; 81); in general, adherence to the Mediterranean or Nordic diets is associated with a lower prevalence of MetS or reduction in its com-

ponents (82), while a Western dietary pattern is positively correlated with greater odds of MetS (83). Conversely, the association between the individual macronutrient intake and the components of the MetS has been analysed in a few studies (84; 85) and controversy still exists about the optimal amount and source of dietary macronutrients and their relative proportions to counteract MetS risk. Over the past decades, an impressive body of quantitative knowledge regarding how dietary changes impact various aspects of BW and metabolism has been accumulated. Integrating this knowledge to make quantitative predictions is a formidable task given the multiple nonlinear interactions between various organ systems. Such an integrative approach is required to better connect energy and nutrient intake to normal physiology as well as to derangements that underlie conditions such as obesity, diabetes, and MetS. To our knowledge, there are no available reports demonstrating the predictive role of the energy/macronutrient intake gaps (molecular level), as assessed by the difference with the dietary reference values, on the clinical parameters (macroscopic level) related to MetS. In the present retrospective study, we correlated energy and macronutrient intake to the clinical features of MetS, with the final aim to provide an additional indication about the most important dietary contributors to clinical abnormalities related to an increased risk of MetS. To grasp the role of each nutritional variable in the general frame of MetS pathological conditions, a model selection for various regression models between nutritional variables and clinical outcomes was performed. The analysis was inherently multivariate and allowed for the unveiling of how inadequate energy/macronutrient intake can predict clinical alterations leading to the MetS onset in a group of subjects at risk of the disease.

### 3.2.2 Experimental Design and Statistical Methods

#### Cohort and dietary assessment

The subjects involved in the study were men and women (age 18–80 years) at risk for MetS enrolled in the randomised, double-blind, placebo-controlled, parallel intervention trial performed in the framework of the EU project PATHWAY-27. Eligible volunteers had two to four of the MetS diagnostic criteria (86), with at least one of them being elevated fasting triglycerides (TG) or low high-density lipoprotein cholesterol (HDL-C). Exclusion criteria are reported in Table S1 (<https://www.mdpi.com/article/10.3390/nu13041377/s1>). Volunteers were recruited in four European centres: Human Nutrition Research Centre of Auvergne (Clermont-Ferrand, France), Max Rubner-Institut (Karlsruhe, Germany), University of Leeds (Leeds, UK), and St. Orsola-Malpighi Hospital (Bologna, Italy). At recruitment, blood pressure and anthropometric measurements (height, weight, and waist circumference, WC) were taken by trained staff as described in Bub et al. (2019) (87). Blood was collected and analysed, as previously reported (88). The present study addressed the intake of energy and macronutrients at baseline as possible dietary predictors of the onset of MetS. At recruitment, participants were asked to complete a validated

semiquantitative food frequency questionnaire (FFQ) that was developed in the European Prospective Investigation into Cancer and Nutrition (EPIC) study (89), and a 24-h dietary recall (24hR), which is designed to assess energy and nutrient intake. The FFQ (covering a 12-month period) and the 24hR were administered by trained personnel. Both FFQ and 24hR were completed by 281 participants (125 females and 156 males). Of the 281 dietary assessments, 66 with missing clinical information were excluded from the analysis and 215 subjects (94 females, aged 23–77 years, and 121 males, aged 24–78 years) having a complete dataset of both dietary assessment and clinical parameters were considered. After misreporting evaluation, 157 subjects were included in the analysis.

### Energy intake and misreporting

Energy and nutrient intakes from all foods and beverages were calculated using national and international databases and literature information. Dietary information by 24hR was used to corroborate energy and food intakes provided by the FFQ. Daily energy intake was derived for each subject. Daily intake of total available carbohydrates, sugars, total fat, saturated fat, and unsaturated fat was expressed as percentage of daily energy intake (% EI). Intake of protein, dietary fibre, and alcohol was reported as g per day. Based on the protocol developed by the European Food Safety Authority (EFSA) (90), energy misreporting was assessed as the ratio of reported energy intake (EI) to estimated basal metabolic rate (EI:BMR) according to the Goldberg method (91) modified by Black (92). The FFQs were used to estimate reported EI and BMR was calculated using the validated sex- and age-specific Oxford equations suitable for use in populations with a range of weight statuses (93). A moderately-active physical activity level (PAL) of 1.6 was assumed for all participants (94). Misreporters of dietary intake were identified by EI:BMR ratios  $< 0.901$  (underreporters) or  $> 2.841$  (overreporters). Fifty-eight participants were classified as misreporters (17 females and 41 males), and further statistical analysis was performed on a total of 157 subjects (77 females, aged 23–77 years, and 80 males, aged 25–76 years).

### Data tricks: comparison with diet reference values

In each subject, adequacy of intake was assessed by comparing energy and nutrient intakes with age-/sex-specific EFSA dietary reference values (DRVs) (95) or nutrient requirements and dietary guidelines of WHO/FAO (96; 97) if the former were not available. Specifically, the following daily intakes were considered adequate:

- Energy ranging between EFSA DRVs for energy calculated according to age using PAL values of 1.4 and 1.8, which approximately reflect low active (sedentary) and active lifestyles (6.8–10.1 MJ/day and 8.3–12.6 MJ/day ranges for females and males, respectively);



- Total carbohydrates ranging from 45 to 60% energy (% EI);
- Sugars (monosaccharides and disaccharides) < 10 % EI based on the WHO/FAO dietary recommendations;
- Dietary fibre intake  $\geq 25$  g/day;
- Protein between the average requirement (AR) and the population reference intake (PRI) of EFSA DRVs;
- Total fat ranging from 20 to 35% EI;
- Saturated fatty acid (SFAs) < 10%EI according to FAO;
- Total unsaturated fatty acids (UFAs), i.e., monounsaturated fatty acids (MUFAs) plus polyunsaturated fatty acids (PUFAs) ranging from 15 to 20% EI, as calculated by difference according to FAO;
- PUFAs ranging from 6 to 11% EI according to FAO.

In addition, a moderate alcohol consumption (up to one serving per day for women and up to two servings per day for men) (98) was considered acceptable. **Differences between current intake and corresponding recommended intake (mean value of recommended range for energy, total carbohydrates, protein, total fat, and total UFAs and PUFAs; limit of adequate intake for sugars, dietary fibre, SFAs, and alcohol) were calculated. The resulting delta values were then used as features of the predictive model.**

### Statistical learning: model selection and fine tuning

Data were stratified by gender. All clinical parameters were classified as normal (1) or abnormal (2) according to their overlap with the recognised normal ranges (Table S2, <https://www.mdpi.com/article/10.3390/nu13041377/s1>). The distribution of clinical parameters was evaluated using the D'Agostino–Pearson test. Student's t-test for normally distributed data and Mann–Whitney U test for non-normally distributed data were used to compare the general characteristics of the study population grouped by gender. All statistical analyses were conducted using the Python programming language, using custom scripts and the sklearn package (99). A predictive model for each clinical parameter was computed using all dietary variables via a ridge regression framework (100). To simultaneously reach the best prediction performances while learning which sets of dietary macronutrient intakes (variables) were the most important for each predicted clinical outcome (target), a multivariate model was applied. To this end, a model selection was performed in order to find the best regression model, using a stochastic grid-search of the parameters of each regression for optimization. Since no univariate effect

of a single nutritional variable on the clinical targets emerged, we assumed that multivariate techniques were the most promising methods as they are capable of simultaneously reaching the best prediction performances while learning which sets of variables are the most important for each prediction task. Indeed, penalised maximum likelihood methods (LASSO regression, ridge regression) outperformed other classes of regression models as previously shown in other nutritional studies (101; 102). In particular, the ridge regression yielded the best fit on the data and was selected as the model to be fine tuned. The complete framework used for learning in this study can be found at <https://github.com/CarloMengucci/Model-selection-and-learning-for-nutritional-data>. The ridge regression belongs to the wider class of penalised linear regressions. These types of models allow computing a regression while shrinking the coefficients of uninformative variables. The linear ridge regression minimises the function:

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p b_j x_{ij})^2 + \alpha \sum_{j=1}^p \beta_j^2 = RSS + \alpha \sum_{j=1}^p \beta_j^2 \quad (3.1)$$

RSS = residual sum of squares, with  $i$  = index of summation for observations,  $n$  = number of observations (1 to 77 for women, 1 to 80 for men),  $j$  = index of summation for variables,  $p$  = number of variables. The penalty coefficient  $\alpha$  can be tuned to optimise the bias-variance trade-off of the model, leading to a maximisation of the predictive performance as a function of the smallest set of the descriptive variables necessary to achieve said performance. The penalty term introduced by the ridge regression is useful to deal with multicollinearity and prevent overfitting; for the present case, it translated to the shrinkage of coefficients of dietary variables strongly correlated among themselves and weakly correlated to a given clinical marker. The absolute value of the regression coefficients  $\beta$  is related to the univariate effect of a given dietary variable  $x_j$  on a given clinical marker  $y$ , while the sign of coefficients is not directly interpreted as it would have been in a normal ordinary least squares (OLS) solution. Before regression, data were standard scaled. All the ridge models computed were cross-validated to optimise the parameter  $\alpha$  through 5-fold cross-validation. Train and test subsets were extracted to maintain the proportion between recruiting centres to minimise the possible confusion factor tied to dietary habits of the country of origin. To represent the statistical dependence between the rankings of dietary variables and clinical outcomes, correlation heatmaps were also computed using the Spearman rank correlation coefficient (103), which measures how well the relationship between two variables or targets can be described using a monotonic function. The Spearman rank correlation coefficient allows for nonlinear relationships to be detected, providing a good description of the relationships between features and targets.

### 3.2.3 Results

Table 3.1 summarises the characteristics of volunteers included in the study. As expected, a significant heterogeneity was evidenced between men and women, possibly due to different hormonal profiles and body fat distribution (104; 105). Ridge-type penalisation was obtained retaining all predictors and minimising collinearity amongst variables; it performed better than LASSO probably due to the complexity of interactions of all the dietary variables in defining the overall clinical picture in the subjects at risk of MetS. Indeed, ridge regression performs better when many predictors have coefficients of roughly equal size (106). The Pearson correlation coefficients of determination ( $R^2$ ) for the clinical outcomes according to the ridge regression are reported in Table 3.2. To visualize the overall complexity of the relationships between clinical parameters and nutritional variables, we computed the heatmap of correlations between them (Figure 3.3). The topology of the heatmaps for male and female study participants was slightly different, highlighting the gender-related differences in clinical and nutritional characteristics associated with MetS. Within these premises, the linear ridge regression has been chosen as the best trade-off between predictive performances and interpretation of results.

Based on the results of the predictive model (Table 3.2), we focused on clinical features that were better estimated by the adequacy/inadequacy of dietary intakes ( $R^2 > 0.4$ , as a generally accepted standard (107)), highlighting the variables characterising the prediction. To this aim, ridge regression performance and the magnitude of regression coefficients were plotted per individual clinical outcome (figures S1 and S2 in supplementals, at <https://www.mdpi.com/article/10.3390/nu13041377/s1>); an example for males BMI prediction (with protein color gradient) and ridge coefficients is reported in Figures 3.4,3.5.

	Women		<i>p</i> †
	Median (IQR) †	Median (IQR) †	
Subjects (n; %)	77; 49.0%	80; 51.0%	–
Age (years)	58 (50–66)	54 (46–63)	0.0631
BMI (kg/m <sup>2</sup> )	31.6 (27.5–35.5)	29.0 (26.1–33.0)	0.0187
WC (cm)	100.5 (93.0–111.0)	104.5 (99.0–113.5)	0.0489
TG (mg/dL)	160.5 (125.1–192.6)	188.9 (153.1–239.1)	0.0006
Total cholesterol (mg/dL)	232.6 (201.4–254.2)	202.7 (188.6–230.1)	<0.0001
HDL-C (mg/dL)	48.2 (41.7–56.6)	39.2 (35.2–42.6)	<0.0001
LDL-C (mg/dL)	164.5 (138.0–180.7)	136.0 (112.2–154.8)	<0.0001
Fasting glucose (mg/dL)	96.1 (87.4–100.8)	97.0 (89.2–102.6)	0.3454
Fasting insulin (μIU/mL)	12.4 (7.9–18.4)	12.9 (9.7–18.5)	0.5382
HbA1c (%)	5.6 (5.3–5.8)	5.3 (5.1–5.6)	0.0022
SBP (mmHg)	130.0 (120.0–145.0)	130.0 (125.0–141.5)	0.2527
DBP (mmHg)	81.0 (76.0–89.0)	85.0 (80.0–91.0)	0.0057

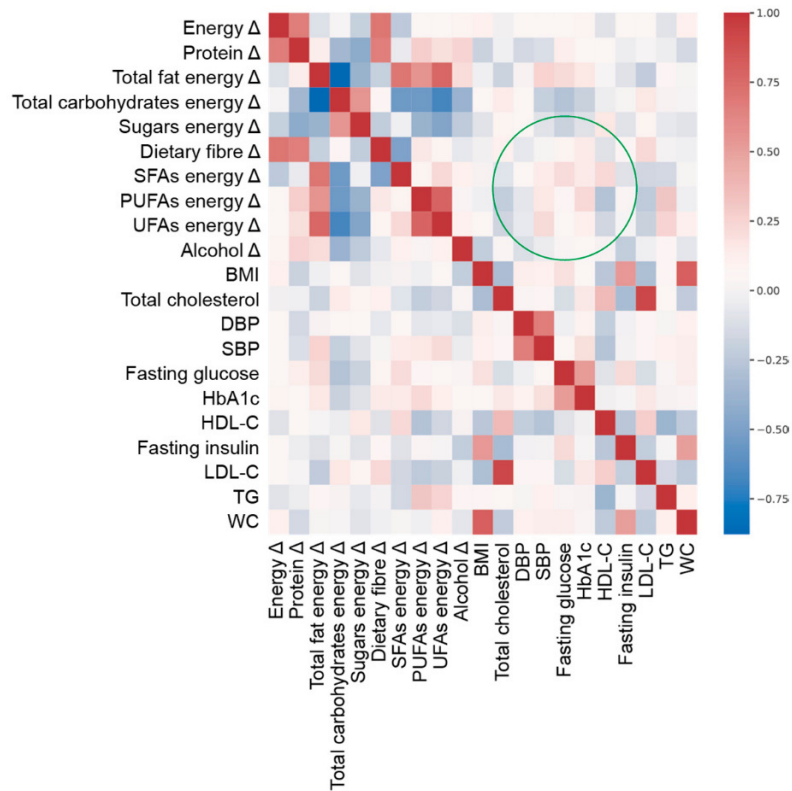
Abbreviations: BMI: body mass index; DBP: diastolic blood pressure; HbA1c: glycated haemoglobin; HDL-C: high-density lipoprotein (HDL) cholesterol; IQR: interquartile ranges; IU: international units; LDL-C: low-density lipoprotein (LDL) cholesterol; SBP: systolic blood pressure; TG: triglycerides; WC: waist circumference. † Median (IQR) for all parameters, except subjects (n; %). ‡ *p* values from Student's *t*-test for normally distributed variables (WC, total cholesterol, LDL-C, DBP) and Mann–Whitney U test for non-normally distributed variables (age, BMI, TG, HDL-C, fasting glucose, fasting insulin, HbA1c, SBP).

**Figure 3.1:** General characteristics of the study population grouped by gender (medians and interquartile ranges, IQR). MDPI, 2021

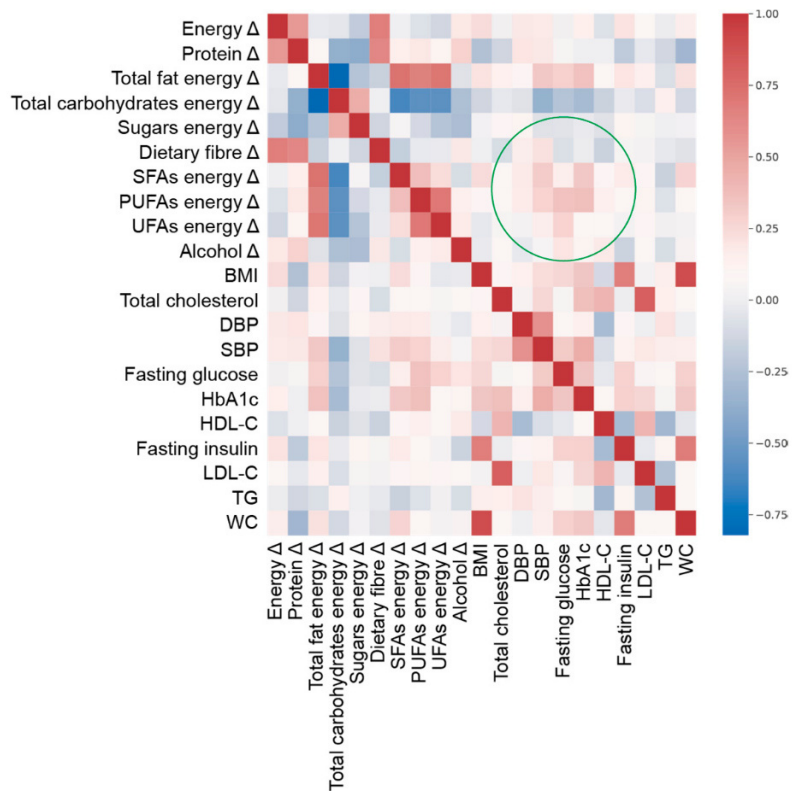
	Women	Men
	R <sup>2</sup> +	R <sup>2</sup> +
BMI	<b>0.43</b>	<b>0.78</b>
WC	0.39	<b>0.79</b>
TG	<b>0.46</b>	0.22
Total cholesterol	0.35	0.25
HDL-C	<b>0.44</b>	0.34
LDL-C	<b>0.42</b>	0.22
Fasting glucose	0.33	<b>0.44</b>
Fasting insulin	0.26	<b>0.49</b>
HbA1c	0.27	<b>0.52</b>
SBP	0.36	<b>0.52</b>
DBP	0.27	0.40

+ The Pearson correlation coefficient of determination (R<sup>2</sup>) of each clinical outcome is the average of the results of each validation fold.

**Figure 3.2:** Pearson correlation coefficients of determination R<sup>2</sup> for the clinical outcomes according to the ridge regression. R<sup>2</sup> > 0.4 are in bold. The significance of these correlation coefficients is p ≤ 0.01. This indicates that the ridge regression is a good model for predicting the respective clinical outcomes. MDPI, 2021

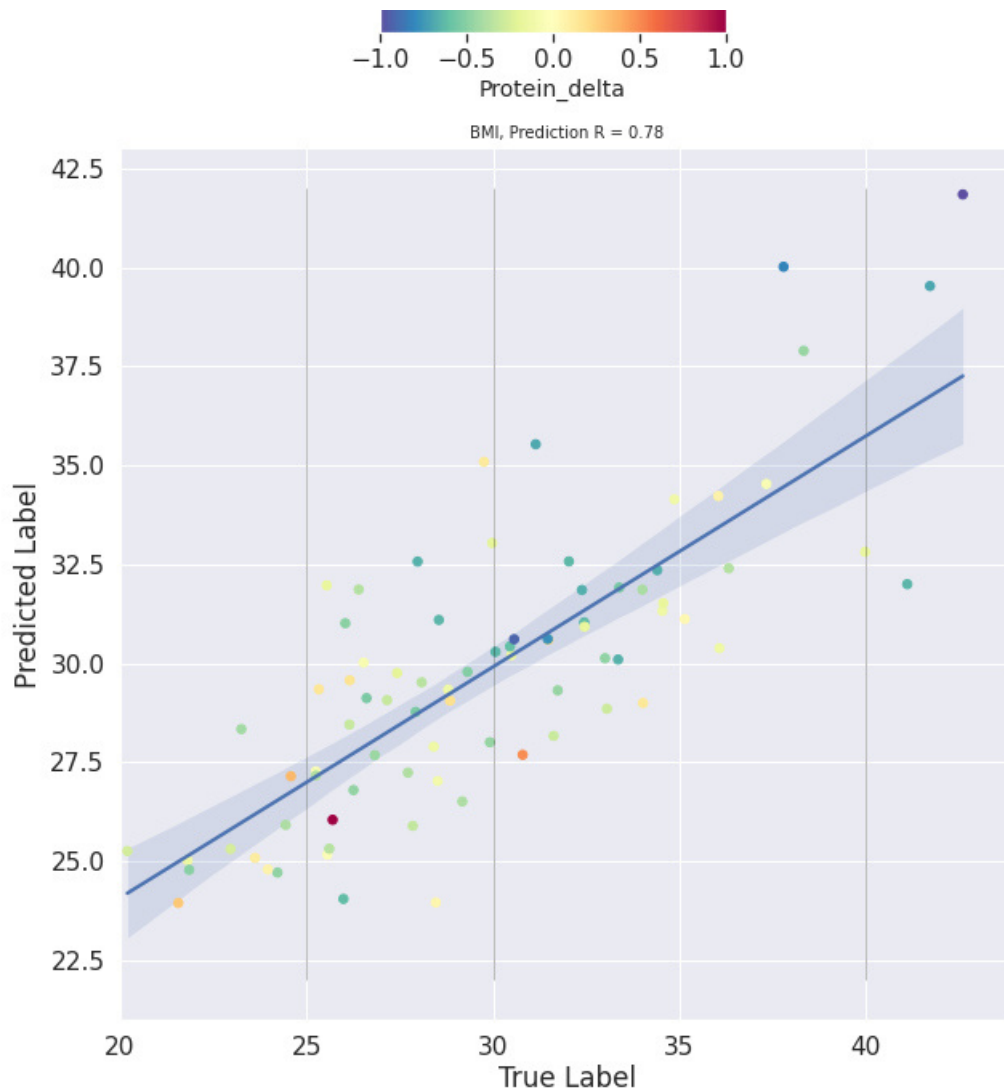


(a)

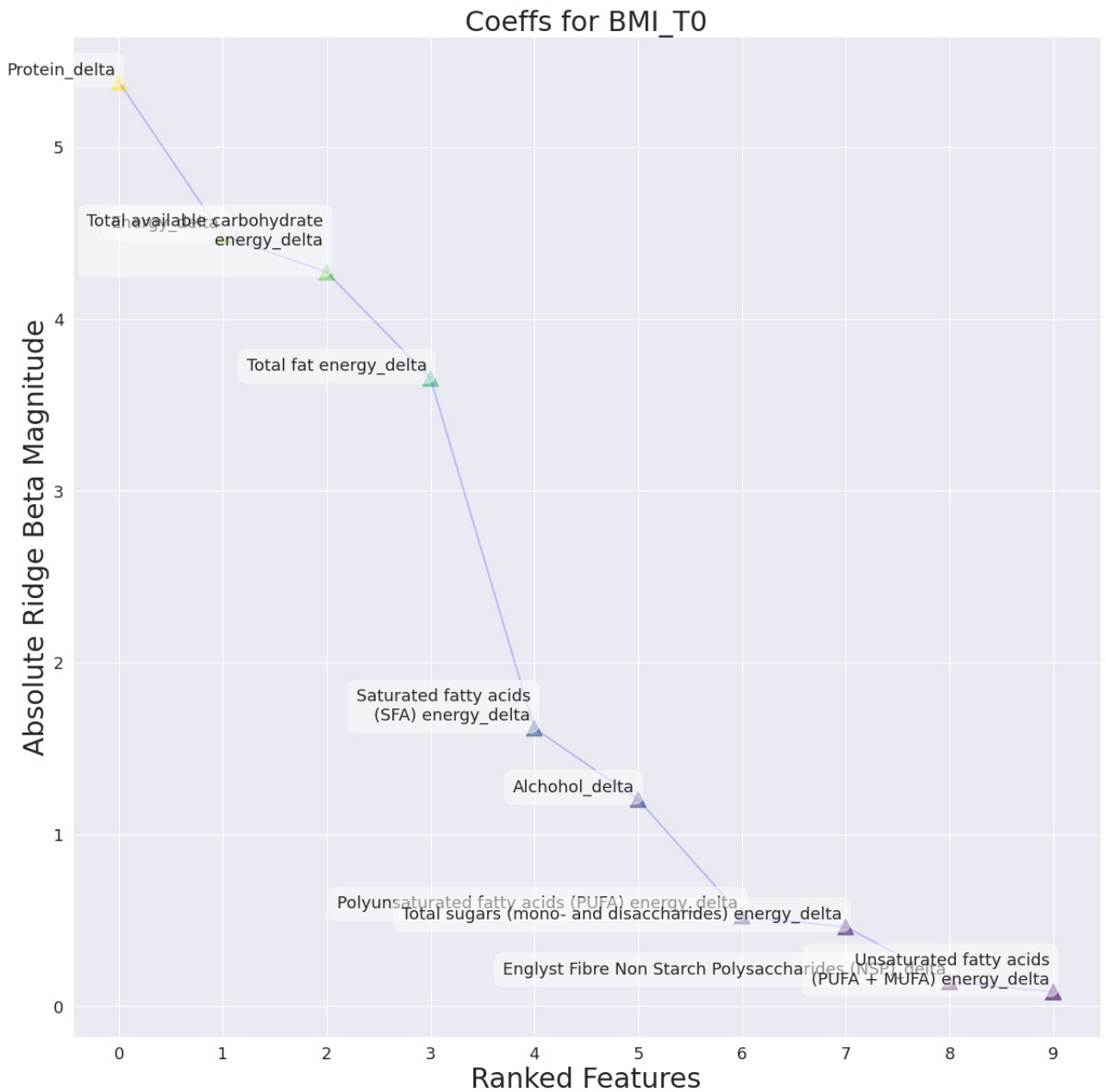


(b)

**Figure 3.3:** Spearman's correlation heat map of adequacy of energy/macronutrient intake ( $\Delta$ , delta value calculated as difference between current intake and corresponding recommended intake of each variable) and clinical parameters at enrolment (Table 3.1) in (a) females and (b) males. Green circle marks an example of a different pattern in correlation between variables and targets in males and females. , MDPI, 2021



**Figure 3.4:** Example of scatterplot of ridge regression performance per clinical outcome (BMI, males). The plot is coloured with a continuous gradient related to the position of each subject (dot) in the distribution of each dietary variable (protein intake in this example). The gradient is computed on normalised values of the intake of each variable so that an extremely scarce intake gets the value  $-1$  (indigo) and an extremely excessive intake gets the value  $1$  (dark red). Values within the recommended range are normalised accordingly, with the spectre of yellow denoting a correct intake ( $\Delta \sim 0$ ). This example shows that in males a protein-poor diet contributes to the development of obesity in subject at risk of MetS. This result is reflected in the ridge coefficient plot for this regression, where protein intake has a high-magnitude coefficient associated. MDPI, 2021



**Figure 3.5:** Magnitude of regression coefficients (BMI prediction). High magnitude is related to a significant univariate effect on the prediction of the given target., MDPI, 2021



According to the model results, inadequate dietary intakes better predicted BMI in males ( $R = 0.78$ ) than in females ( $R = 0.43$ ). In both genders, exceeding energy intake was the key predictor of a high BMI, especially for obese subjects, confirming that BW changes are associated with an imbalance between energy intake and expenditure (97). Based on our predictive model, overweight and obesity were also predicted by elevated intake of total fat ( $> 35\%EI$ ) in males and of SFAs ( $> 10\%EI$ ) in both genders. This is in accordance with outcomes from epidemiologic studies and clinical trials, which suggest that total fat (108) and SFA intake (109) are strongly linked to BW. Excessive intake of SFAs was also a characterising predictor variable of high WC, which was well estimated by inadequate dietary intakes in males ( $R = 0.79$ ). In males, low consumption of protein was a good indicator of both elevated BMI and WC. Our results suggest that consumption of high-energy, low-protein, and high-fat diets, particularly when including excessive SFAs, strongly relates to the development of obesity in men and to a lesser extent in women. This gender-related difference confirms results from long-term prospective studies evidencing a significant positive connection between weight gain and dietary fat in a cohort including males and female (110), while energy content from fat was weakly correlated with weight gain in The Nurses' Health Study, including only women (111). Interestingly, neither total carbohydrate nor sugar intake were predictors of overweight/obesity. Epidemiological evidence and results from diet intervention trials suggest that protein and carbohydrate intakes are inversely related to BMI, while excessive intake of sugars contributes to obesity (112). Although the plausibility of the mechanisms provides support for a role of sugar consumption in the epidemics of overweight/obesity, definitive studies are missing (113). In our group of subjects at risk of MetS, the predictivity of sugar intake on BMI was low and supported the conclusion that there is no clear or convincing evidence that any dietary or added sugars have a unique or detrimental impact relative to any other source of calories on the development of obesity (114).

Results from the model indicated that blood TG were better predicted by the dietary variables in women ( $R = 0.46$ ) than in men ( $R = 0.22$ ). In the female subjects, correct (6–11% EI) or slightly low PUFA intake and correct fibre intake (25 g/day) had a good predictive value of normal TG level, confirming the importance of dietary fibre in the maintenance of normal blood TG (115). Although the total PUFA intake must be considered first when examining dietary habits affecting lipemia (116), it is documented that high n-3 PUFA intake favourably impacts on blood TG (117), while excessive consumption of n-6 PUFAs may lead to negative effects (118). In this study, based on dietary questionnaires, it was not possible to accurately discriminate between n-6 and n-3 PUFA intake and it could explain why high PUFA intake did not predict normal TG level. In females, inadequate dietary intakes, mainly high consumption of available carbohydrates and fats, predicted high LDL-C, and adequate total PUFA intake was a predictive variable of normal LDL-C ( $R = 0.42$ ). In women, low HDL-C was well predicted by inadequate intakes ( $R = 0.44$ ), mainly excessive energy, SFA, and available carbohydrates. Although

the magnitude scale of the ridge coefficient was generally low for HDL-C prediction, denoting a prediction characterised by the combined effect of nutrients rather than a single dietary variable, overall, our data confirmed evidence in the literature about the negative effect of excessive carbohydrate intake on dyslipidaemia (119; 120). Total cholesterol was not well predicted by any dietary intake both in females ( $R = 0.35$ ) and males ( $R = 0.25$ ).

Clinical features related to glucose metabolism and insulin sensitivity were better predicted in men than in women. In males, fasting glucose was well predicted ( $R = 0.44$ ), and elevated total fat and SFA intakes were slightly associated with moderate fasting hyperglycaemia. These results are consistent with evidence demonstrating that excessive consumption of total fat (121) and SFAs (122; 123) favour the onset of insulin resistance. Although no univariate effect of any specific dietary variable was evidenced, fasting insulin and glycated haemoglobin (HbA1c) were moderately predicted by the combined effect of all dietary nutrients in the male group ( $R = 0.49$  and  $R = 0.52$ , respectively). DBP was not well predicted by the examined dietary variables either in women ( $R = 0.27$ ) or men ( $R = 0.40$ ). Conversely, inadequate dietary intakes well predicted high SBP in males ( $R = 0.52$ ), and excessive total fat intake concomitant to low PUFA intake was a good predictor of moderate hypertension (140–159 mmHg). Again, our results confirmed evidence from observational and epidemiological studies (124; 125; 126). Although several studies found an association between alcohol drinking and the prevalence of MetS and most of its components, in our study, alcohol consumption was not predictive of any clinical outcome. We speculate that this was related to the very low percentage of enrolled volunteers exceeding the acceptable consumption of alcoholic beverages (6.5% of females and 16.5% of males), which did not allow any stratification based on alcohol intake.

### 3.2.4 Study Conclusions

In summary, in this retrospective analysis, we focused on the predictive effect of energy/macronutrient intake on the clinical features related to the risk of MetS. We did not focus on food intake and/or dietary pattern, of which their contribution to the risk of MetS has already been addressed by several studies (see (127) for a comprehensive review). Although this approach has limitations since components other than energy and macronutrients are provided by food/diet, our results suggest that predictivity of inadequate intake of energy/macronutrients is independent of dietary patterns. Indeed, we evaluated four different cohorts with different dietary habits tied to the geographical origins of the volunteers, and train and test batches of the cross-validation were stratified with respect to the nationality of each subject to avoid biases derived from different eating habits. Overall, energy/macronutrient intake had a strong predictivity. We speculate that this relies on the intimate relationship between MetS and obesity, which is in turn strongly dependent on the unbalance of energy/macronutrients in the diet. Of note, not all clinical outcomes

were predicted with the same accuracy, and the predictivity was overall higher in men than in women. Furthermore, inadequate intake of specific nutrients was associated to abnormality of specific clinical parameters. Most of the observed intake/clinical outcome associations were consistent with previous evidence. This does not mean that our results are trivial and simply confirmatory, but rather it confirms that the proposed model is suitable for shedding light on the complexity of nutritional variables that, although responsible for impacting on clinical outcomes and, therefore, for influencing the pathological condition, have an effect that is not evident with univariate analysis and must be considered in the framework of the reciprocal influence of the other variables. The impact of physical activity and smoking was not considered in our model, and this is a limitation since they are both included among lifestyle factors predictive for MetS (128). Of note, none of the enrolled volunteers was a heavy smoker ( $\geq 5$  cigarettes per day); this minimising the confounding effect and making a stratification based on these lifestyle characteristics impossible. As well, based on exclusion criteria, none of enrolled subjects had a high level of physical activity ( $\geq 5$  h of physical activity per week). Specific information on physical activity was collected using the international physical activity questionnaire (IPAQ) only for volunteers who accepted to participate in the sub-study of the trial, so we could not use those data in the regression model. Anyway, collected IPAQ confirmed that physical activity was low-moderate.

In this retrospective study, the energy and macronutrient intake of 157 (80 males and 77 females) adult volunteers at risk of MetS from four different countries was evaluated using a validated standardised protocol to measure dietary intakes. The use of the ridge regression, which optimises prediction performances while retaining information about the role of all the nutritional variables, allowed us to assess if a clinical outcome is strongly dependent on a single nutritional variable, or if its prediction is characterised by more complex interactions between the variables. The approach appeared robust, and although our results cannot be applied to the general population, they allowed for the linking of energy/macronutrient intake to the clinical features of MetS, thus providing additional indications about the most important dietary contributors to the risk of the disease. Methods in prediction modelling have been recently growing and are becoming more relevant in the nutrition field (129). In the near future, they could be useful to healthcare professionals and policymakers to effectively counteract the risk of MetS and other diet-related diseases.

### 3.3 Chapter Conclusions

Evaluating exposure to diet, even at its most macroscopic level is no trivial matter. The heterogeneity of free living populations, from genetics to cultural and geographical habits, poses a challenge in finding links between diet and health. In addition, there is an objective technical difficulties in gathering nutrition epidemiological data and organize large

scale studies. There is also an everchanging consensus in the definition of diet itself, that in turn affects the formulation of an effective modelling paradigm. We proposed a predictive model based on rigorous standardization and handling of nutritional data, based on state of art references. The model focused on the stratification of the sample population to avoid confounding factors, and treating food intake as patterns of macronutrients. This allowed to bypass possible biases tied to diet description and quantification, while focusing on the crosstalk of macronutrients in predicting some clinical outcomes. This resulted in a non-trivial characterization of the links of metabolic syndrome and effects of diet in a multivariate fashion. However, given the complexity of food and its interaction with human physiological functions, the sole description of edipemiological aspects of diet exposure is not enough when treating the problem of health and nutrition links from a complex systems perspective. In the next chapters, we try to take a deeper look into food intake description and the simulation of how it interacts with the human organism.

## A Closer Look: Modelling the Impact of Chemical Composition

### 4.1 Unravelling the Complexity of Food: a Framework for Foodomics

This section is entirely based on the published work by Mengucci et al. (130).

Holistic methods at the basis of the foodomics approach are allowing the in-depth understanding, at molecular and supramolecular level, of the complexity of food matrix. The latter, in turn, affects the nutrient bioaccessibility, one of the crucial factors impacting on the final effect of diets. However, many levels of complexity are emerging, relating to food-human interactions, while bolus descends along the whole gastrointestinal tract. Such complexity makes in-vitro and in-silico models still unable to fully describe intertwined kinetics between food matrix and human compartments. A possible framework to unravel complexity is outlined, starting from bioaccessibility modelling all the way down to inter-compartmental kinetics. The aim is to enhance algorithms and models for the prediction of the impact of a food category on a class of individuals. The proposed framework can consider many levels of complexity, provided that time-resolved experiments, suitable for integration with food matrix description, are correctly designed for this purpose.

#### Glossary

- **Bioaccessibility:** fraction of a given food compound released from the food matrix in the gastrointestinal tract. Bioaccessibility kinetics comprises description of both release and transition to absorbable form.
- **Bioavailability:** fraction of a given food compound effectively reaching systemic circulation.

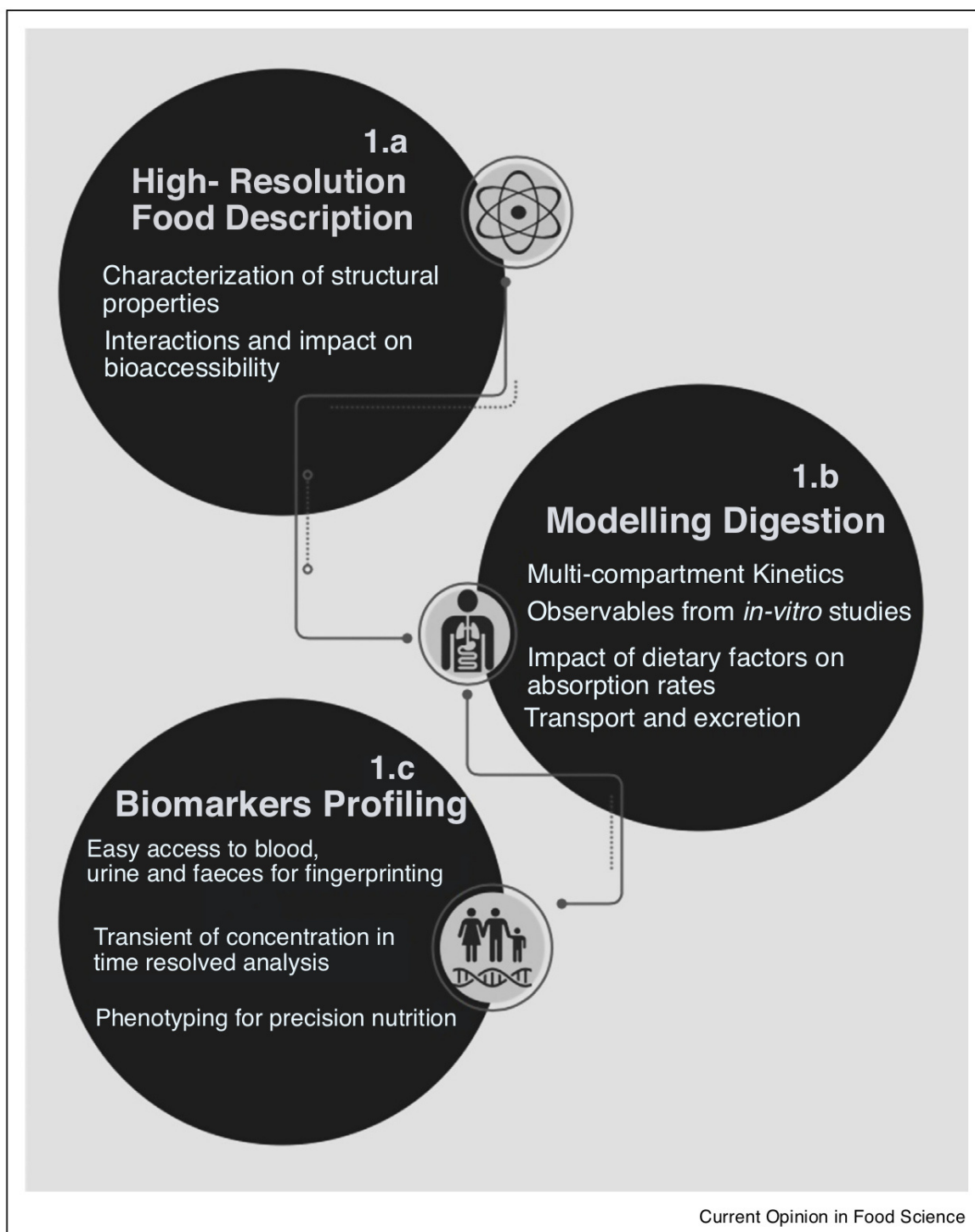
- **Data-driven approach:** machine learning or deep learning architectures applied to high-throughput data with the aim of extracting patterns of features for classification (fingerprinting, molecular profiling, untargeted studies, etc.)
- **High resolution food description:** description of the heterogeneity of the food matrix, in terms of compounds and their interactions, organizing spatially and by speciation, along different length scales.
- **In-vitro model:** proxy experiment performed with microorganisms, cells, or biological molecules, outside their normal biological context, aimed at reproducing different stages of physiological processes as close as possible to in-vivo scenarios.
- **In-silico model:** mathematical model capable of describing kinetics, using observables and parameters derived from in-vivo, in-vitro models and data-driven approaches.
- **Multi-compartmental model:** kinetic model capable of considering the propagation of the effects of bioaccessibility along different stages of the digestive process, that is, transformation of the food in different sections of the gastrointestinal tract, absorption, transport and excretion.
- **Metabotype:** the metabolic phenotype of an individual.
- **Observable:** physical and chemical quantity that can be measured.

#### 4.1.1 Introduction to Foodomics

One of the main challenges in clinical nutrition is the translation of findings emerging from basic nutrition into meaningful, tailored and clinically relevant dietary advises to prevent or counteract metabolic disorders (131). Several factors must be taken into consideration when designing efficient nutritional solutions: although those relating to individuals are generally considered to be the most important, other variables, equally important, emerge. Among them, the food structure and the interaction between food and the human gastrointestinal tract (GIT) are fundamental. Therefore, a ‘precision nutrition’ approach should consider not only individual variability (i.e. genetics, type of microbiome, metabolome, dietary habits, lifestyle) (132) but also food structure and composition, along with dynamics of digestion and absorption. At present, the evaluation of nutrient intake is mainly based on chemical composition of consumed food and does not consider bioaccessibility, that is, the amount of the food components that is released from the food matrix, and bioavailability, that is, the amount of bioaccessible components that is absorbed and delivered to tissues through the blood stream. Since the food matrix and processing have a significant impact on bioaccessibility, which in turn impacts on bioavailability, a holistic approach to food characterization is needed. The foodomics

approach offers not only a high-resolution food description, dealing with the various levels of complexity converging into food science (133), but also the in-depth description of the food metabolome. The food metabolome is the part of the human metabolome directly derived from the digestion, absorption and biotransformation of foods and their constituents (134). Thus, the food metabolome strictly depends on bioaccessibility and bioavailability kinetics. Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS), hyphenated or not to chromatographic separation methods, are optimal techniques to comprehensively characterize the food metabolome, which can be considered one of the dimensions of the foodomics space (135). The different levels of information in the food metabolome can be explored by i) targeted metabolite analysis, ii) metabolite profiling, iii) spectral fingerprinting, iv) untargeted metabolite analysis and v) metabolomics, with increasing discrimination capability. To fully understand the food metabolome, the behaviour of food and food components along the gastrointestinal tract (GIT) must be considered (fig.4.1). In-vitro and in-silico models have been developed to simulate digestion and absorption, allowing to build up predictive models (136; 137). Predictive models need validation using blood, urine and faeces obtained from carefully designed intervention trials, including data quality control protocols. Samples from well-designed intervention trials can also be used to select specific biomarkers of intake. These biomarkers reflect the interactions between the food and the human body and can be used to build up in-silico models to predict bioaccessibility and bioavailability, thus allowing the classification of foods, diets and human subjects. To this purpose, the kinetic constants that regulate mass transfers between the different body compartments (including GIT) are crucial. Therefore, to develop accurate in-silico kinetic models, the time-dependent concentrations of biomarkers in different body compartments must be assessed.





**Figure 4.1:** Framework for kinetics of bioavailability investigation. The three main stages of modelling foodomics data are highlighted: 1a) numerical descriptors for the food matrix are required to be included as input for the machine learning framework; 1b) modelling will find the right parameters for matching the food intake to the experimental time-resolved concentrations of food biomarkers; 1c) the set of output parameters, extracted upon modelling of food-biomarkers kinetics in blood and urine, are condensed signatures of metabolic phenotypes, linked to nutritional response to specific food products.

Within the framework of the FoodBALL project (138), many databases have been developed (FoodDB, Exposome Explorer) merging data obtained from samples coming from intervention studies. One of the main concerns emerged in the project is the transient concentration of food-related molecules, which makes their classification as biomarkers extremely difficult. Indeed, a food-related molecule may not be recognized as biomarker of intake depending on its absorption kinetics. In fact, its concentration at the time of sampling may not be different from baseline because it has not reached the peak yet (subjects with slow absorption kinetics) or it has already passed the peak (subjects with fast kinetics). The use of proper modelling can overcome this limitation, and it can also consider the ‘food matrix effect’. Although recent works highlight the importance of developing personalized wellness tools relying on data integration and biomarker mapping approaches (139; 140), a consensus solution is far from being accepted since the derived in-silico models are not yet validated and are still at an embryonic stage. In the following sections, we discuss the possible integration of in-vitro experiments (data sources) with machine learning approaches aiming at extracting molecular features (data-driven approaches) to give rise to in-silico modelling able to predict kinetics of biomarkers in different compartments. We outline a framework for merging different levels of complexity by discussing methodologies and challenges for food-human interactions while stressing the importance of: i) choosing proper in-vitro descriptors for the food matrix; ii) identifying in-vivo biomarkers of food intake within pattern clustering and fingerprinting techniques; iii) integrating food matrix descriptors with biomarkers kinetics.

#### 4.1.2 **Challenges and Novel Strategies**

The work by Westerman et al. (140) outlined a promising direction for nutritional recommendations based on custom biomarker correlation mapping. In that work, a set of common blood biomarkers of health was organized in a network of correlations, whose variations were studied over time. This approach allowed finding new patterns or networks of predictive biomarkers to better understand transitions between health and disease states. Such patterns resulted in valuable information about the average baseline functional complexity and a subject-dependent variability. New correlations between biomarkers emerged, such as those between Low Density Lipoproteins (LDL) and iron stores, possibly explaining perturbations in lipid metabolism in conditions of iron overload. However, causality between changes in biomarkers after dietary intervention and health improvement could not be established, except for a small subset of subjects with biomarkers ‘out-of-clinically accepted range’ at baseline. Beside the presence of confounding factors and the difficulty to treat baseline variability, one limit of the above described approach could be the attempt of connecting the intervention diet and the biomarkers without fully considering the complexity of the food and of the food-human interactions.

### Challenges in food matrix description

A high-resolution description of the food is the first step needed to unravel the complexity of the food-human interaction (fig.4.1a). Foods are highly heterogeneous materials, and food components interactions are organized physically and chemically in the space along different length scales, thus generating a structural complexity in the food matrix. The effect of food structure on food disintegration and micronutrient release has been exemplarily described in a recent work by Hiolle et al. (141). The description of food structure usually relies on data gained by several imaging techniques, including Light Microscopy (LM), Scanning Electron Microscopy (SEM) (142), and Magnetic Resonance Imaging (MRI) (143). Further details about the interactions between the food matrix and water, which is the diffusing medium for most nutrients, are also provided by nuclear magnetic relaxometry (144). Image analysis and relaxometry allow to evaluate physicochemical and rheological features of the food, assessing their impact on bioaccessibility.

A different approach is given by modelling based on machine learning and data-driven techniques, which provides a fingerprint of the food matrix by merging its chemical and physical properties. Chemical fingerprints can be obtained through various techniques ranging from spectroscopy to gas chromatography. Accordingly, chemical descriptors can be concentrations and variations of concentration in time-resolved observations, proportional to spectral features, with the advantage of not needing to formally identify each single descriptor. If the quantification is robust, a correlation pattern of descriptors, even if unidentified, can be exploited along with other outputs for hypothesis-free fingerprinting. Furthermore, the physical structure of a matrix can be described by merging quantitative measures of structural properties of the sample and multimodal imaging derived features. Techniques such as multidimensional hyperspectral imaging analysis have proven to be effective for matrix characterization and oxidative damage detection (145) and to be suitable for descriptors extraction for fingerprinting. Magnetic Resonance Imaging can also give quantitative information about properties of the food matrix, such as tortuosity and porosity (146), enhancing the array of possible multimodal descriptors for machine learning and data-driven approaches.

Breaking down the challenges and the modelling aspects of food matrix effects on chemical reactivity, many levels of complexity are emerging (147): i) effects on thermal stability of bioactive compounds and micronutrients; ii) thermodynamics and kinetics of reactions; iii) reactants concentration when catalytic phenomena are present; iv) diffusivity and partitioning of reactants among different phases of a matrix and v) enzymatic interactions. As a matter of fact, a chemical reaction occurring in food will yield a rate different from the rate obtained in ideal conditions (i.e. a very diluted solution) and varying from food matrix to food matrix. Such an effect can also account for a displacement of reaction equilibrium. Food matrix can thus change thermodynamic and kinetic properties of the reaction by acting on: i) concentrations of reactants and products, ii) activity coefficients, iii) diffusivity of reactants and products, as well as on iv) the temperature

perceived by the reactants in each compartment of the system.

An exhaustive framework for the integration of fingerprinting and kinetics studies has been brought forth by Grauwet et al. (148). Although focusing on the topic of evaluating the effects of extrinsic factors, such as processing on food quality changes, this work offered a comprehensive view on the techniques and approaches to be exploited for food characterization and extensive data generation (GC and HPLC MS, NMR based approaches). Moreover, the importance of linking fingerprinting with kinetics, through multi-response observation was highlighted. Multi-response observation for food means studying transformations in food. They do not occur isolated but, rather, within a network of reactions which are consequent to a variety of combinations of processing conditions. From a mathematical point of view, this is done by translating the reactions network into a system of coupled differential equations, using all the information extracted during studies aiming at characterizing the food matrix. The result is an insight into the rate constants of specific reactions steps, and their dependence on secondary variables (i.e. temperature, pressure, time, etc., in food processing), which refers to the study of a multi-phase reaction system shaping the food matrix. The paper by Grauwet et al. (148) also outlined some basic concepts behind multivariate data analysis (MVDA) techniques, which are crucial for information extraction in frameworks of the proposed type. On the basis of the concept of multi-response kinetics, different compartments (i.e. the food, the GIT, the human metabolism) cannot be considered isolate systems. Therefore, data obtained in each compartment should be merged and integrated as part of a network of interactions. Modelling kinetics should consider complexity by building in-silico models including information from food matrix to the human body, including the GIT.

### **Challenges in the description of the impact of the food matrix on digestion**

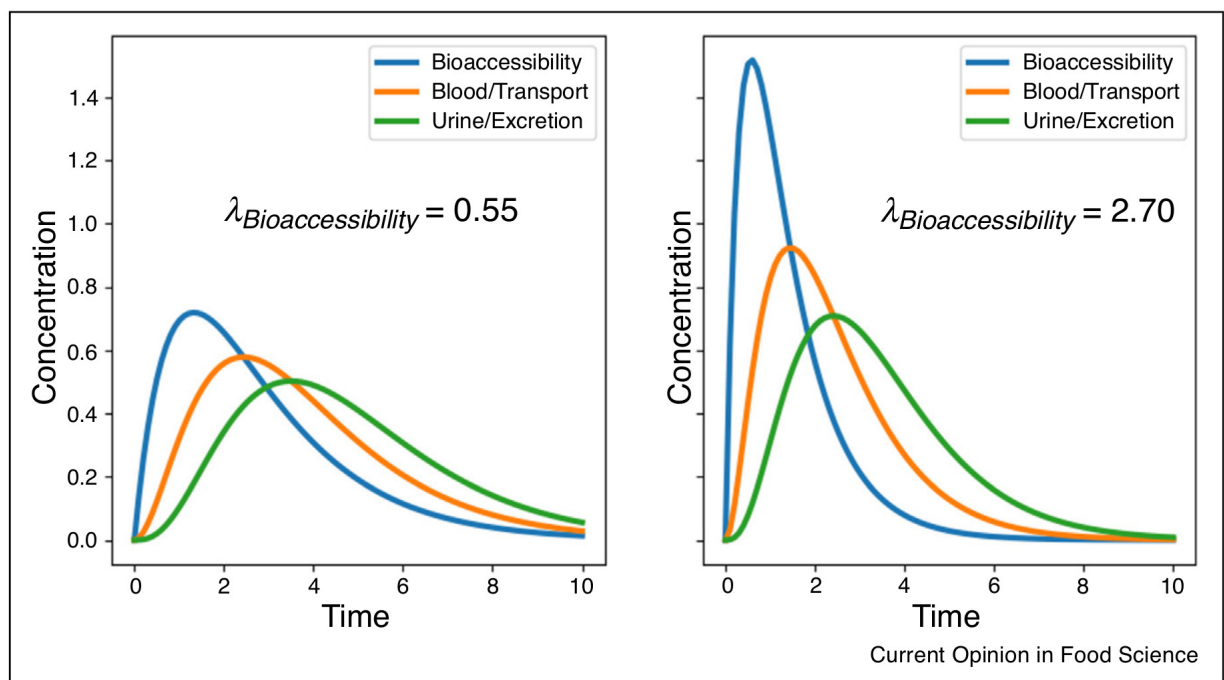
The food matrix affects food components bioaccessibility/bioavailability influencing the entity and the kinetic of the release process in the GIT. Together with the individual intrinsic variability (e.g. genetic polymorphisms) and the effect of the overall diet, the food matrix effect can lead to different digestion or absorption capacity of specific components, thus modulating the ultimate effect on physiology and health (4). Research has focused on the development of standardized food models (SFM) for in-vitro experimental set-ups and investigations on three major levels for bioavailability modelling: bioaccessibility, absorption and transformation of nutrients within the GIT (fig.4.1b). Mimicking the composition of representative diets allowed studying bioaccessibility of bioactive compounds. This aspect is affected by the heterogeneity of mixtures with different physical phases and nanostructures, in which nutrients tend to organize during digestion processes along the entire digestive apparatus. In a recent work by Zhang et al. (149), an SFM representing a typical US diet was proposed to investigate the effects of food matrices on bioaccessibility of nutraceutical and pesticides. Microstructures were characterized in each phase of the simulated in-vitro digestion using confocal fluorescence microscopy, also consid-

ering electrical properties. The work showed significant impact of the food matrix on bioaccessibility of bioactive compounds, and provided insights on the role of lipid digestion and its interaction with hydrophobic nutraceutical. Besides, it provided examples of possible important observables (i.e. physico-chemical properties) derived from in-vitro set-ups. As examples, variations in electrical properties, particle size and microstructure distribution were acquired in each single stage of the digestion, to model different levels of complexity, as they impact on the interactions of enzymes with fat droplets. Indeed, the inclusion of these variables allowed to describe and explain the different ions release from food fragments and fat droplets in the different environments of the GIT. Similar descriptors coming from in-vitro studies could play a crucial role in integrating the food matrix effect into reliable in-silico models considering the matrix-dependent complexity of digestion kinetics. Of note, modelling structural interaction terms in kinetic equations systems, by inserting in example a quadratic damping term representing diffusion under certain conditions or a sigmoidal term representing percolation dynamics, could enhance in-silico simulations capabilities. When in-silico models must predict intertwined kinetics occurring in different compartments of the GIT, a set of observables such as those discussed above can be used to estimate and model interactions.

### Challenges in integrating food matrix description and metabolomics

Observation derived from in-vitro experiments play a key role in the construction of appropriate kinetic descriptors of the food matrix effect on bioavailability. Since properties of the matrix influence the first phase of the food/human interaction, that is, bioaccessibility, they influence all the subsequent phases. Therefore, a multi-compartmental modelling is needed (fig.4.2) to account for complexity in an appropriate manner. To build multi-compartmental models, patterns of blood and urine biomarkers can be adopted as proxies of the food/human interactions evolving during digestion and link them to the description of the food matrix.

Metabolomic of blood and urine is a key tool in the identification of dietary biomarkers that can be also used to classify and quantify food intake (138). Many metabolomic studies are focused on expanding and validating Biomarkers of Food Intake (BFI). Garcia-Pérez et al.(150) suggested an analytical pipeline based on correlation maps of <sup>1</sup>H-NMR identified metabolites for evaluation of dietary intake. That work evidenced tartaric acid as a dose responsive biomarker of grape intake, while proline betaine was indicated as a marker of citrus intake in the study by Gibbons et al. (151). Clusters of biomarkers of milk, cheese and soy-based drink were identified by Münger et al. with untargeted multiplatform analysis (152), and 3-methylhistidine was confirmed as specific for white-meat intake (153).



**Figure 4.2:** *Compartments kinetics at different bioaccessibility parameters. A simple model to visualize the effect of parametrizing bioaccessibility tied to food matrix. The simulations are run at identical starting concentration values and parameters of other compartments, except for bioaccessibility, whose parametrization is given by  $\lambda$ . Differences in bioaccessibility propagates affecting the kinetics of a given observable in other compartments.*



However, the ratio between validated and putative bio-markers of food intake is still very low. A guideline for evaluating the quality of candidate biomarkers was proposed by Dragsted et al. (154). The adopted parameters included assessment of plausibility, dose response, time response, robustness, reliability, stability, analytical performance, and inter-laboratory reproducibility. The most powerful perk of metabolomics is its ability to discover untargeted patterns of metabolites for subject classification. Single diet biomarkers might offer incomplete information and do not suffice when phenotyping free-living populations or trying to understand relationships between food consumption and disease risk (fig.4.1c). Garcia-Perez et al. (155) suggested the possibility to overcome biases related to self-reported dietary intake by a discrimination based on the fingerprinting of the whole urinary spectral profiles. Specific spectral archetypes were obtained from individuals kept in controlled feeding conditions and used for classification of dietary intake in free-living individuals. It was shown that the differentiation among dietary interventions was only allowed by whole patterns of urinary biomarkers embedded in the metabolic profile, while single specific biomarkers were not able to correctly classify the diet. Considering whole patterns in place of single biomarkers can also mitigate the risk of misinterpreting metabolites concentration. As an example, urinary concentrations of TMAO can be associated with healthy, fish-rich diets; however, gut bacteria can synthesize TMAO from choline and hence high urinary and plasma concentrations can also originate from high red meat consumption, which is commonly tied to adverse health outcomes. Observation of the whole metabolome can disentangle such ambiguities. The whole spectra of identified and unidentified signals, and the modification of their correlations, can allocate individuals in different metabotypes, thus enhancing baseline modelling and providing elements for intervention-related kinetics evaluation. One of the biggest challenges in this kind of approach is that compartmental-model computing needs a large amount of time-points data for robust parameter estimation. Many studies have thus focused on breaking down and simulating single compartment kinetics, focusing on absorption, digestion, transport or excretion. A recent work from Bjornson et al. (156) highlighted the importance of evaluating interactions between absorption and transport phases. Using plasma samples, a novel non steady-state model was proposed, integrating metabolic characteristics of both apoB100 and apoB48 and the kinetics of triglycerides in response to a fat-rich meal. The model was proven to be physiologically relevant, providing information about apoB48 release in the basal and post-absorptive state, as well as about the contribution of intestine to Very Low-Density Lipoprotein (VLDL) pool size and kinetics. In a similar fashion, patterns of variation of spectral signals tied to metabolites may be used to intertwine multi-compartmental kinetics, highlighting different profiles of response for different dietary interventions, while retaining inter-individual information and variability.

Usually, kinetic parameters can be drawn from at most two compartments (transport/absorption and excretion, if both serum and urine metabolomics are available) of



the  $N$  possible macro-compartments of a model given by a chain of differential equations describing kinetics, such as the Bateman equations system. Fitting parameters for such equations becomes thus a challenge, especially when trying to model single-subject kinetics in the parameter space, unless time sampling is sufficiently high. Such a constraint should drive the experimental design of nutritional trials when kinetic information must be used for in-silico models. The high-resolution food description is also essential to conceive an informative quantification of bioaccessibility in the kinetic model. In fig.4.2 the effect of including bioaccessibility in a simple multi-compartmental model is shown as a scalar parameter  $\lambda$ , to emphasize its effect propagating to every compartment. This quantification can be improved by finding functions of different parameters, extracted with the different techniques used to describe food matrices in each experiment, and including them in the kinetic model.

### 4.1.3 Remarks

- Holistic approaches such as foodomics are allowing the in-depth understanding of food matrix characteristics at molecular and supermolecular level. This is radically changing the nutritional approach that is now starting to consider the food complexity as an important variable in the final effect of the diet.
- Responses to food intake are not only specific for each individual but largely depend on the food matrix, including its modification due to processing. It is now clear that food cannot be considered a homogeneous mixture and it is time to give the right emphasis to the organization of the matrix.
- The heterogeneous phases of the food matrix compartmentalize the biological systems and modulate the interactions among substrates and enzymes. This spatial restriction to the free diffusion of molecules may change during storage and/or processing of the food, which could be described as a dynamic system, and it is dramatically modified during digestion.
- The destiny of a food component, from raw material to human compartments, is very complex. After digestion, accessible components are absorbed in a temporal and spatial distribution, some of them being meanwhile actively metabolized by the microflora. Active metabolization of absorbed components can occur already in the enterocyte before distribution to organs through the bloodstream. To predict it, in-vitro models simulating the physiological processes are adopted to the purpose of simplifying the interpretation of the results. However, these systems must undergo complex validation before being considered reliable predictors of in-vivo phenomena. This validation is enhanced by an in-silico step, that is, the construction of mathematical models and algorithms, which simplify the description of the different phases that food undergoes. These models are based on multi-factorial kinetic

functions, whose parameters can be used to classify different categories of foods and of the corresponding individual responses.

- To tackle the goal of these models, that is the possibility to predict the impact of a food category on a class of individuals, and to overcome mathematical constraints on parameter estimation, huge amounts of data from time-resolved studies are necessary.
- The framework described considers many levels of complexity and highlights the importance of optimizing time-resolved experiments. This is a crucial step to implement robust algorithms and models based on machine learning and data-driven approaches, currently at the embryonic stage in this specific field of applications.

A time-resolved picture for the evaluation of the impact of the intake of a food with a given composition, on certain categories of individuals, is a key element to start grasping the whole set of possible intake related physiological outcomes. There are at least two important aspects, besides the evaluation of instantaneous exposures to certain compositions, that are linked to the molecular impact of intake: the study of how a long-term exposure to certain bio-active compounds modulates the molecular state of individuals; the study of the mechanism of how physiological functions linked to digestion are activated to assimilate compounds, by observing short-term kinetics after intake. In the next sections, we present and discuss two studies that focus on modelling kinetics aspects considering composition, by operating within the Foodomic framework.

## 4.2 Multi-Omic Model of the Impact of a Bio-Active Compound

The study of the impact of nutritional strategies based on the introduction of a specific bio-active compound, is considered interesting under many aspects. The endpoint of establishing causality between the intake of certain compounds and health promoting effects is the formulation of nutraceutical (157) and nutritional strategies, that can boost response to diseases, prevent inflammations, boost immunological triggers and promote healthy statuses in general. This endpoint is not easily reached: free living populations are exposed to large numbers of perturbations that can modulate the molecular state, making it difficult to find direct and robust links with induced health promoting effects (if any exists) derived from the introduction of the bio-active compound of interest. As such, an approach that is gaining popularity is to study the downstream-upstream loop changes in molecular statuses through modulation effects in gut microbiota (158). The microbiota is an excellent proxy of inflammatory status and probiotic effects, by essentially reflecting the changes in the patterns of abundances of microbial species that can produce/consume health promoting or health degrading metabolites.

In this section, based on the published work by Biagi, Mengucci *et al.* (159), 2020, MDPI, we propose a framework to investigate the modulation in microbiota and

metabolome induced by a nutritional strategy based on different doses of vitamin B2 in broilers. Apart from the possible applications in production efficiency of broiler meat, this study allowed us to develop a framework to study the crosstalk between microbiome and metabolome. The use of broilers for such an in-vivo trial also has clear experimental advantages: a large sample size in a relatively short time span, the possibility to study molecular state modulation at different growth stage of a complex organism and consider the effect of aging in the model, the availability of different tissues (caecum, ileum) from the GIT.

#### 4.2.1 The role of microbiome sciences in animal production

Microbiome science holds great promise for the future of health maintenance and performance improvement in animal production, because gut microbes are responsible for the degradation of complex substrates and energy extraction, as well as for the promotion of the animal's immune system functionality (160; 161). Indeed, gut microbiome acquisition and maturation are pivotal processes for the development of intestinal epithelium physiology, in terms of immunity, intestinal barrier integrity and nutrient digestion (162; 163; 164), possibly playing a crucial role in strategies aimed at preventing pathogen colonization and boosting weight gain (165). Therefore, a key issue in animal production, including chicken nutrition, is to understand the relationship among the effects of diet composition and the changes in microbiota and host metabolism (166). Chickens' microbiota is characterized by strong spatial variability along the gastrointestinal tract: specialized communities inhabit different sections of the animal gut, performing specific digestive functions. The most studied of these communities are those residing in the ileum, where nutrient absorption takes place, and the caeca, in which fermentation and digestion of complex polysaccharides occur (167). The caeca, typical of the avian intestinal tract, are a couple of appendages protruding from the junction of the small and large intestines, in which the feed retention time is the highest, and carbohydrate fermentation, urea recycling and water retention take place (160; 165). Indeed, 10% of the energy recovered from the food is estimated to be produced by fermentative processes occurring in the caeca. In that intestine section, the concentrations of short chain fatty acids (SCFA) and other organic acids (i.e., lactate) are higher than in other tracts (167). Such microbial products are crucial for host immunological fitness and nutritional homeostasis. Indeed, they provide energy to the epithelial cells, and can be carried to the liver and used as energy substrates for muscle tissue (161). Some of these compounds can be the subject of microbial cross-feeding, e.g., the lactate produced by *Bifidobacterium* and *Lactobacillus* members can be utilized by other anaerobic bacteria to produce butyrate (168), highlighting the complexity of the microbe–microbe and host–microbe relationships, all involved in defining the final homeostasis and health of the chicken meta-organism.

The microorganisms inhabiting the litter are of both environmental and fecal origin. Litter is continuously pecked and ingested by the animals, thus playing a relevant role in

determining the composition of gastrointestinal communities. In addition, litter can act as a reservoir of both animal pathogens and zoonotic agents (160). Studies in this field pointed out the importance of analyzing changes in the different broilers' microbiomes over time and how these are affected by intervention strategies to improve animals' performance. An important focus of such studies must be the effect of interventions on the abundance and persistence of key core microbiota players (169; 170; 171). Probiotics and prebiotics are the most accredited strategies to attempt to modify microbiome functionality and composition (172), but other dietary components and nutritional supplements can also modulate gastrointestinal functionality, the gut microbiome, the innate immune system, the intestinal barrier integrity and the intestinal enzyme activity. In this framework, vitamin B2 (riboflavin) can modulate multiple pathways important for the maintenance of the gastrointestinal functionality. There is evidence that vitamin B2 has prebiotic effects (173), affecting the microbiome's ability to regulate the innate immune system (mucosal associated invariant T cells, MAIT cells). This compound reduces intestinal inflammation and apoptosis and regulates gut protease activity (impacting animals' food behavior and growth). Moreover, vitamin B2 has been found to be most effective in synergy with antibiotics against methicillin-resistant *Staphylococcus aureus* (174). Therefore, vitamin B2 can be part of novel solutions that modulate several aspects of gastrointestinal functionality, creating the opportunity to identify additive/synergistic effects with other feed additives.

Here, we report the results of an experimental trial on Ross 308 broilers fed different amounts of vitamin B2. Caeca and ileum microbial communities were longitudinally analyzed along the 42-day broiler productive cycle, together with litter samples, in order to investigate the effects of 50 and 100 mg/kg vitamin B2 dietary supplementation on the microbiota composition and diversity, as well as on the core microbiota components that can persist over time and be shared across the different ecosystems. In addition, in order to explore the supplementation effects on microbial-host co-metabolism, a nuclear magnetic resonance (NMR)-based metabolomics approach was used for analysis of caecal contents.

#### 4.2.2 Study design and methods summary

To guide the reader through result, we report a short summary of the study design, metabolomics and machine learning for spectral and kinetics analysis. **A complete description of acquisition techniques, metabolomics, sequencing and bioinformatics for microbiota analysis is reported in *Appendix A***

#### Study design and dataset

Three groups of 120 Ross 308 female chickens each (total number of birds: 360) were housed at the Poultry Research Facility of the University of Bologna in Ozzano dell'Emilia (Italy) in three separate rooms, labelled as A, B and C. The rooms were next to one another

and were under identical environmental conditions. Birds reared in each room received a different diet/were fed with a different diet (Room A-control diet; Room B-control diet + 50 mg/kg vitamin B2; Room C-control diet + 100 mg/kg vitamin B2). To obtain diets with a medium (Room B) and a high (Room C) level of vitamin B2, the control diet, containing the standard dosage (i.e., 5 mg/kg) of vitamin B2, fitting the recommendations for the whole grow-out phase of broilers fed diets containing wheat (Ross 308 Nutrition Specifications, 2014), was supplemented with vitamin B2 (Rovimix<sup>R</sup> 80SD; DSM Nutritional Products) up to 50 mg/kg for group B and 100 mg/kg for group C. The dosages in groups B and C (10× and 20× the control, respectively) were set to ensure that a quantity of vitamin B2, largely above the recommended dosage, was able to reach the lower gut.

Each experimental group was sampled three times: at day 15 (T1), day 28 (T2) and day 42 (T3). During each sampling, a total of 40 birds/room (a total of 120 birds) were randomly selected and euthanized following ethical guidelines to minimize stress and pain. Nine litter samples of 10 g each (3 samples/room; 1 sample/pen) were also collected. The entire gastrointestinal tract was obtained from each bird. Caeca and ileum contents were collected in 2 mL sterile tubes, flash frozen in liquid nitrogen and stored at -80 °C for further investigations. Caeca contents from the 120 birds were collected in duplicate to conduct microbiome and NMR metabolome analyses separately.

### Spectral processing and machine learning

After Fourier transform and baseline correction, spectra were calibrated with reference to the chemical shift of 0.00 ppm assigned to the internal standard TSP; spectral peripheral regions, together with the water signal, were removed. After this, spectra were normalized employing the probabilistic quotient algorithm (PQN) (175) on two different regions separately (regional scaling) since this worked best for this type of sample. After normalization and prior to any possible statistical analysis, spectra were binned into intervals of 100 data-points of 0.0183 ppm each. As a result, the new spectral profile consisted of 410 binned data, which were saved as a matrix in a text file.

All statistical analyses and machine learning routines were carried out in Python 3.6, using implementations from the ScikitLearn package and custom scripts. Ten-fold cross-validation was carried out for each prediction task in order to avoid overfitting. Prediction results reported are the average of the folds. Spectra were reduced with ScikitLearn PLSRegression using the NIPALS algorithm (176), modified to suit a discrete classification problem. To select the most important features for each latent variable created, the partial least square (PLS) weights spectra were smoothed with a combination of Savitzky–Golay (SAVGOL) filters (177) and asymmetric least square smoothing and baseline correction (178). Peaks in weights spectra were furthermore filtered using a signal-to-noise ratio (SNR) threshold, to minimize the probability of selecting uninformative zones of the original NMR spectra. Sample group separation was evaluated using the SciKitLearn implementation of C-support vector classifier (SVC). The parameters for

the classifier were estimated using a stochastic grid search, with a linear kernel and a regularization parameter of 0.01 yielding the best performances.

### Kinetics fitting

The term kinetics is hereby used to emphasize the emerging time-dependent variation of the concentrations of the metabolites, of which the estimate was obtained by fitting average concentrations at each one of the three time points. The signals, proportional to the concentration, of metabolites of interest were fitted to highlight possible differences in the variations between treatment groups. For the purpose of the study, the amounts of metabolites are calculated using normalized signal arbitrary units (A.U.), proportional to their molar concentration. Signal distributions were square root transformed, in order to enhance normality and reduce fit bias. At each time point, for each group, metabolite signal was estimated as the average of the signals, while using standard error as the error bar for the plots. Time points were fitted using generalized linear models (GLM) from the statsmodels (<https://www.statsmodels.org/stable/glm.html>) module of Python 3.6, with a regression of the form:

$$Y_i = \alpha + \beta \log(X_i) + \epsilon_i \quad (4.1)$$

This allowed us to account for the non-linear relationship between variables, while preserving the linearity of the model and the solution. A fit confidence interval (CI) of 95%, represented by light-colored boundaries in each plot, was reported to assess statistical significance. The CI was computed using bootstrap resampling, to provide an estimate of the variability of the mean tied to the population for each time point, by using the distribution of the means of a sufficiently large number of resamples of the data. This estimation gives an interval where there is a 95% confidence that the true mean of the population lies for each time point. In other words, non-overlapping CI in the plot amongst different treatment groups correspond to statistically different means with  $p < 0.05$ .

### 4.2.3 Results

The aim of this trial was to assess the effects of vitamin B2 supplementation on the ileum, caeca and litter microbiota of broilers, as well as on the metabolic profile of the caecal contents.

#### Microbial communities

741 samples were analyzed, including 357 caeca samples, 357 ileum samples and 27 litter samples. For both caeca and ileum, 120 samples from the first time point (14 days, T1), 119 from the second time point (28 days, T2) and 118 from the third time point (42 days, T3) were available. A total of 4,986,865 high-quality sequences were obtained, ranging

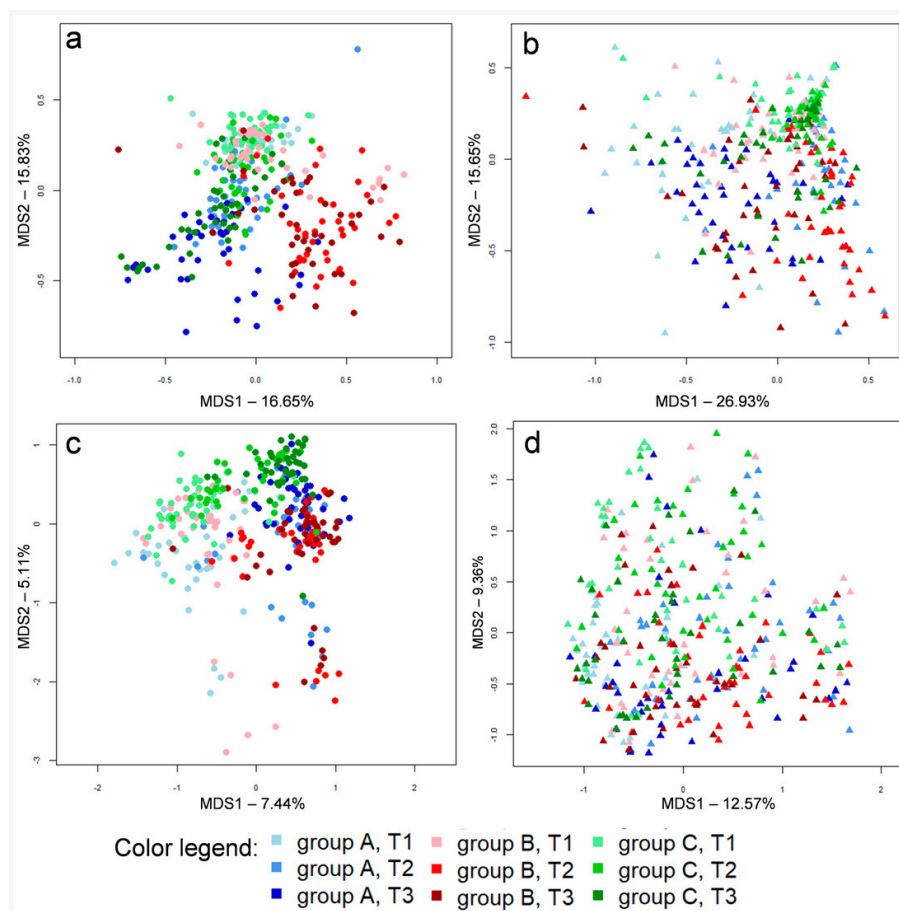


between 1099 and 15,182, with an average value of  $6490 \pm 2715$  sequences per sample. Sequencing reads were deposited in SRA-NCBI (project number PRJNA644889). Reads were clustered into 20,950 amplicon sequence variants (ASVs). As previously reported (169; 171), beta diversity analysis based on both weighted and unweighted Unifrac distances showed a clear separation between ileum and caeca microbial communities, with litter samples clustering in between the two intestinal compartments. In accordance with the available literature (160; 171; 179), caecal microbiota were consistently dominated by Ruminococcaceae and Lachnospiraceae, whereas in ileal samples Lactobacillaceae was the largely dominant family. On the contrary, litter samples showed phylogenetic profiles without a clear dominance, with a high abundance and diversity of families belonging to the Proteobacteria and Actinobacteria phyla. Accordingly, litter samples showed higher biodiversity, both measured by Faith's PD metric ( $3.17 \pm 0.89$ ) and ASV richness ( $68.96 \pm 29.64$ ), with respect to both caecal (Faith's PD index,  $2.37 \pm 0.55$ ; ASV richness,  $61.32 \pm 14.96$ ) and ileal samples (Faith's PD index,  $1.23 \pm 0.57$ ; ASV richness,  $19.18 \pm 9.99$ ). Concerning the Shannon diversity index, litter and caeca samples showed comparable values ( $4.58 \pm 0.73$  and  $4.70 \pm 0.47$ , respectively), both of which were higher than values calculated for ileal samples ( $2.82 \pm 0.62$ ). Beta diversity analysis on available caecal samples (Figure 4.3a,c) showed that samples taken from group B (supplemented with 50 mg/kg vitamin B2) followed a different longitudinal trajectory in terms of microbiota structure with respect to groups A and C. This trend is particularly evident when weighted Unifrac distances are used to plot the whole sample set (Figure 4.3a), whereas the PCoA obtained using unweighted UniFrac distances shows more overlap among the different groups (Figure 4.3c). This indicates that differences in the microbiota of broilers in group B resided in abundant bacterial species, instead of subdominant ones. On the contrary, the beta diversity analysis of ileal samples did not show a clear separation across groups A, B and C (Figure 4.3b,d).

### Caeca, Ileum and Litter Microbiota Composition

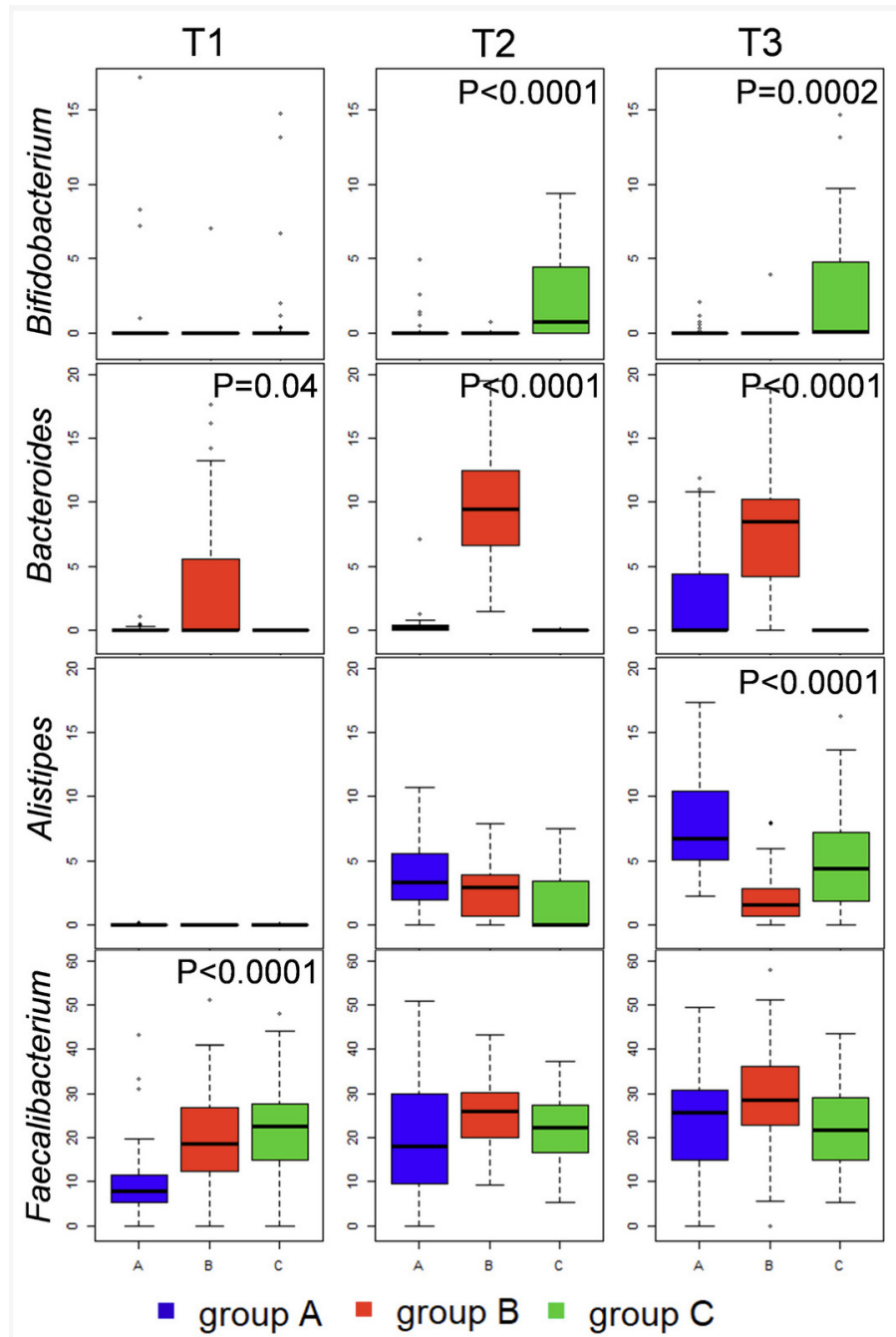
The compositional analysis at a family level highlighted that supplementation of vitamin B2 (50 mg/kg) (group B) promoted the progressive increase of the Bacteroidaceae family in the caeca, whereas in groups A and C the Bacteroidetes phylum was mostly composed by bacteria belonging to the Rikenellaceae family. This observation was statistically confirmed both at family and genus levels (Figure 4.4). Indeed, the family Bacteroidaceae and the genus *Bacteroides* showed significantly higher abundances in group B at all available time points. On the contrary, the family Rikenellaceae and, in particular, its genus *Alistipes* showed higher abundances in groups A and C at T3. Differently, the average family level profiles obtained for the caeca of broilers in group C (supplemented with the highest amount of vitamin B2, 100 mg/kg) showed a progressive increase (from T1 to T3) in the abundance of Bifidobacteriaceae, which was not detected in groups A and B. The significance of this difference was confirmed also at genus levels (genus *Bifidobac-*



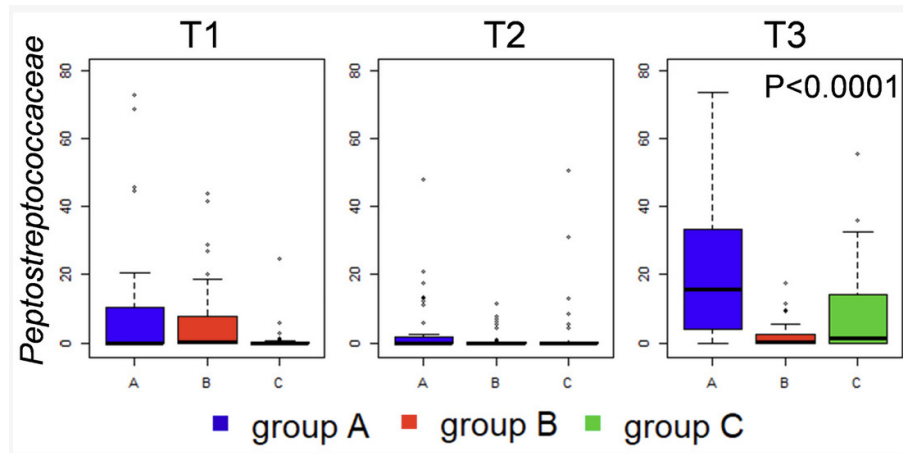


**Figure 4.3:** Principal coordinates analyses (PCoA) based on weighted (a,b) and unweighted (c,d) UniFrac distances of caecal (a,c) and ileal (b,d) microbiota profiles in broilers in groups A (shades of blue), B (shades of red) and C (shades of green). Samples are depicted as dots for caeca and triangles for ileum, filled in different shades of color, from light (earlier samples, day 15, T1) to dark (later samples, day 42, T3), according to the color legend (provided at the bottom). First and second coordination axes are reported in each plot. Percentages of variation in the datasets explained by each axis are reported. MDPI, 2020

terium, Figure 4.4) at T2 and T3. The genus level analysis of caeca profiles also showed that vitamin B2 supplementation, in both groups B and C, accelerated the increase in Ruminococcaceae relative abundance, which was significantly higher in group B and C with respect to the control group A at T1, reflecting an analogous increase in the Ruminococcaceae genus *Faecalibacterium* (Figure 4.4). At the following time points (T2 and T3) the relative abundance of Ruminococcaceae and/or *Faecalibacterium* in the three groups was not significantly different. Ileal microbiota composition was less affected by vitamin B2 supplementation. Compositional analysis only showed that, at T3, both vitamin B2-supplemented diets (groups B and C) significantly inhibited an increase in the abundance of the Peptostreptococcaceae family, which was evident in the control diet (group A) (4.5 4).



**Figure 4.4:** Relative abundance distributions of bacterial genera in the caecal microbiota of broilers. Box and whiskers distributions of relative abundances (%) in all samples at the three time points (from left to right) are depicted for those genera showing significant differences between the three groups (A, blue; B, red; C, green) in at least one time point. Bejamini–Hocherg-corrected  $p$  values obtained from Kruskal–Wallis test are reported when statistical significance was reached ( $p < 0.05$ ). MDPI, 2020

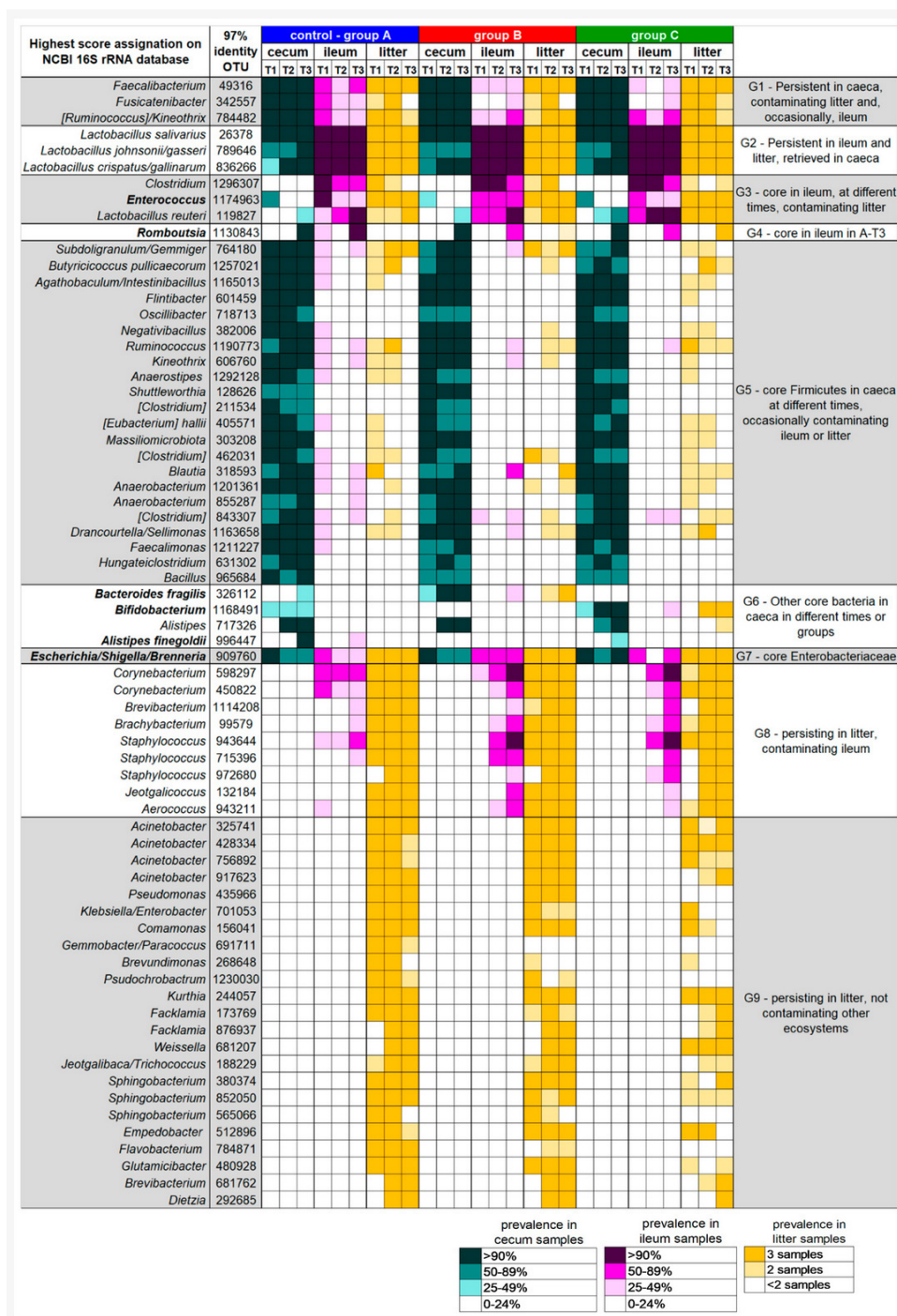


**Figure 4.5:** Relative abundance distributions of *Peptostreptococcaceae* family in the ileum microbiota of broilers. Box and whiskers distributions of relative abundances (%) in all samples, at the three time points (from left to right) are depicted (blue, group A; red, group B; green, group C). Benjamini–Hochberg corrected *p* values obtained from Kruskal–Wallis test are reported when statistical significance was reached ( $p < 0.05$ ). MDPI, 2020

### Ecological Perspective of Broiler Caeca, Ileum and Litter Microbiota

A subsequent re-analysis of the sequences using a different OTU (operational taxonomic unit) picking strategy (UCLUST algorithm with 97% similarity threshold) was performed to facilitate the interpretation of the ecological behavior of the most prevalent and persistent bacterial groups across the three analyzed ecosystems, as well as to evaluate the impact of vitamin B2 supplementation at an ecological level. Indeed, the 97%-similarity threshold allowed us to obtain groups of sequences, possibly ascribable to small group of species, that could play specific ecological roles within and across the caeca, ileum, and litter microbial ecosystems. Among the obtained 97%-similarity OTUs, we filtered those detected with a relative abundance  $> 0.1\%$  in  $> 90\%$  of samples in at least one time point, thus defining ecosystem-specific "core microbiota". For these "core 97%-similarity OTUs" the prevalence, i.e., the percentage of samples in which each OTU was detected at a relative abundance  $> 0.1\%$ , was calculated for all available samples and plotted using a color code in the heatmap in Figure 4.6. The observation of the prevalence of each OTU across the different samples allowed for the clustering of the core OTUs into nine groups. Group G1 comprised those core bacteria of the caeca that were persistent along the longitudinal sampling, including OTUs assigned to the well-known health-promoting *Faecalibacterium*; vitamin B2 supplementation did not affect the prevalence of these OTUs in the caeca, but seemed to have an impact on their prevalence in the ileum. Group G2 included OTUs assigned to *Lactobacillus* species that were persistently part of the ileum core microbiota, confirming the available literature (171), but are also frequently retrieved from the litter and caeca. Group G3 comprised *Clostridium*, *Enterococcus* and *Lactobacillus* OTUs that were part of the core ileal microbiota only

in one or two time points, and only occasionally retrieved from the caeca; *Enterococcus* prevalence in the ileum seemed to be affected by vitamin B2 supplementation. Group G4 included only one 97%-similarity group of sequences, belonging to the Peptostreptococcaceae family and assigned to the genus *Ramboutsia*, that was part of the core ileal and caeca microbiota at T3 in the control group, but less prevalent in groups B and C (62% in both); this taxon could be responsible for the significant decrease in the ileal relative abundance of the Peptostreptococcaceae family associated with vitamin B2 supplementation (4.5). Group G5 included Firmicutes members that were part of the caecal core microbiota in at least one time point, occasionally contaminating ileum and litter samples, and that seemed to maintain such ecological behavior independently of vitamin B2 supplementation; several of these OTUs were assigned to genera known for their butyrate-production capability, such as *Subdoligranulum*, *Butyricoccus*, *Agathobaculum*, *Kineothrix*, and *Anaerostipes* (168; 180; 181), or their acetate-production capability, like *Blautia* and *Faecalimonas* (182). On the contrary, group G6 included non-Firmicutes OTUs of the caeca microbiota, of which the prevalence was deeply affected by vitamin B2 supplementation. An OTU assigned to the species *Bacteroides fragilis* was part of the caeca's core microbiota only at T2 and T3 in broilers supplemented with 50 mg/kg vitamin B2 (group B), reflecting the group B-specific significant increase in abundance of *Bacteroides* (Figure 4.4). Concerning the genus *Alistipes*, of the Rikenellaceae family, an OTU putatively assigned by the BLAST algorithm to the species *Alistipes finegoldii* was probably responsible for the significant difference in the relative abundance of *Alistipes* across groups at the last time point (Figure 4.4), with this genus being part of the caeca core microbiota only in the control group at T3, and being absent in group B. Furthermore, confirming relative abundance data (Figure 4.4), a *Bifidobacterium*-assigned OTU was found in part of the caecal core microbiota only in group C broilers, at T2 and T3. The G7 group included only one OTU belonging to the Enterobacteriaceae family (possibly assigned to the *Escherichia/Shigella* group) that consistently colonized the litter, but also the caeca and ileum. The last two groups of OTUs, G8 and G9, included bacteria prevalently colonizing the litter. OTUs included in group G8 are occasionally found in ileum samples, which they could possibly reach through litter ingestion by the broilers. It is possible to notice that in broilers receiving the highest amount of vitamin B2 (group C) the persistence and prevalence of G9 OTUs was lower, possibly explaining the slight separation observed during the beta diversity analysis using unweighted Unifrac distances.



**Figure 4.6:** Prevalence of core operational taxonomic units (OTUs) at 97% similarity, in broilers' caeca and ileum and litter from groups A, B and C, at the three time points (T1, T2, and T3). Operational taxonomic units (OTUs) at 97% similarity were obtained by using the qiime1 pipeline. OTUs detected with a relative abundance > 0.1% in > 90% of samples in at least 1 time point are shown, together with the identification of the highest score alignment against the NCBI 16S rRNA database obtained by using BLAST nucleotide algorithm. Identification is at the level of species only when > 99% similarity was reached, whereas more than one possible genera are reported when equal scores were obtained. Shades of sea-green, purple, and gold are used to indicate the degree of prevalence of the OTUs in all available sets of samples, according to the provided color legend (bottom). OTUs were grouped according to their ecological behavior across the three analyzed ecosystems (caeca, ileum and litter), obtaining groups G1 to G9, as depicted in the right column. MPDI, 2020

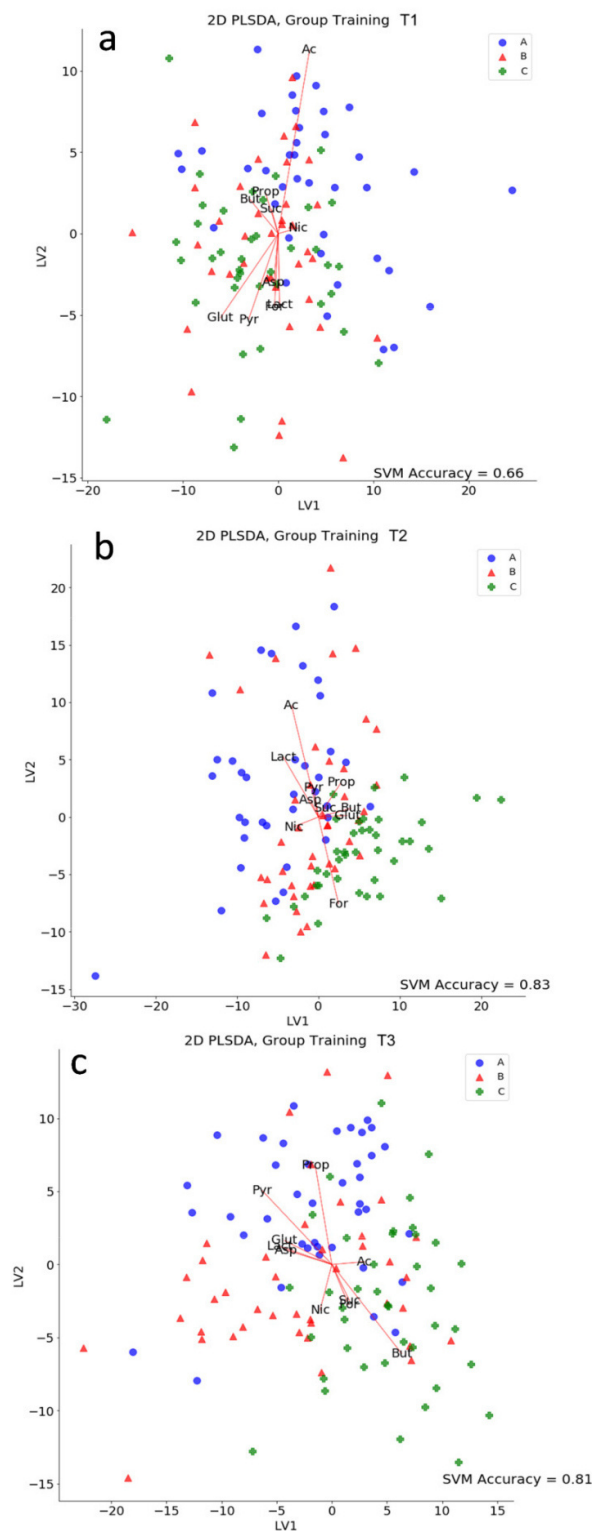


### Caeca Metabolome Analysis

A total of 357 spectra were used to train a PLS-DA for spectral dimensionality reduction. This served as a projection of the whole NMR spectrum to a lower dimensional metabolic space, with the aim of enhancing visualization and finding possible clusters. Preliminary unsupervised analyses showed that animal growth (time) was the most influential factor in metabolic changes, thus being the biggest cause of sample separation in multivariate analysis. PLS-DA was then carried out for all samples at each time point, with the task of discriminating between treatments (Figure 4.7). The best separation between groups, evaluated using SVC accuracy of prediction, was obtained at T2 (Figure 4.7b). At this time point, animals seemed to show the highest metabolic response to treatment. This points out the time window in which treatment effects are the most detectable from a metabolic perspective. PLS-DA was then carried out for all samples at each time point, with the task of discriminating between treatments (Figure 4.7). The best separation between groups, evaluated using SVC accuracy of prediction, was obtained at T2 (Figure 4.7b). At this time point, animals seemed to show the highest metabolic response to treatment. This points out the time window in which treatment effects are the most detectable from a metabolic perspective.

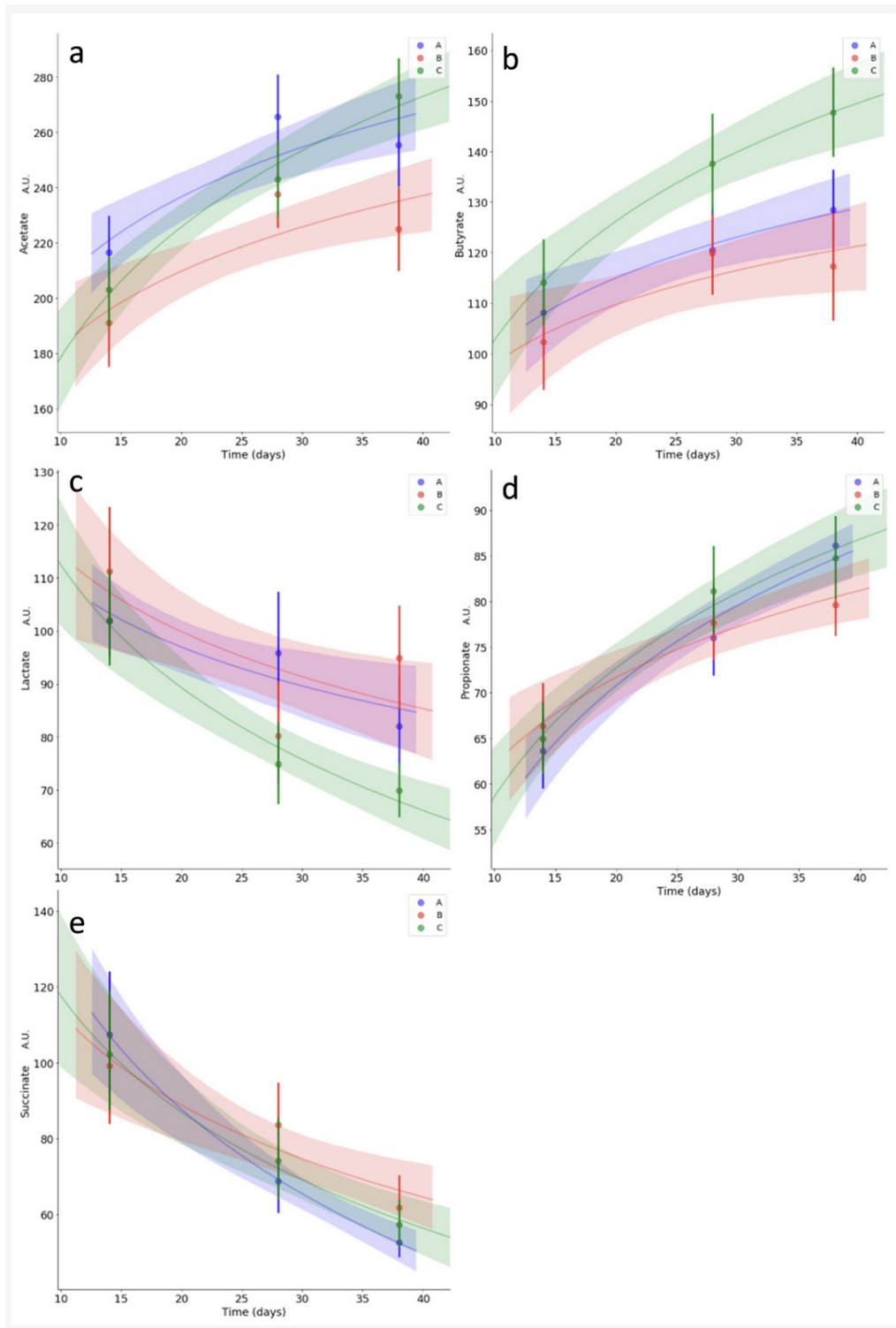
### Kinetics of relevant metabolites

Kinetics studies were carried out on two categories of metabolites of interest: short chain fatty acids and energy metabolism-related metabolites. Nominally, acetate, propionate, lactate, succinate and butyrate were selected for the first category (Figure 4.8), and aspartate, glutamate, nicotinate, formate and pyruvate were selected for the second (Figure 4.9). Treatment B had a significant dampening effect on acetate starting from day 28, whereas group A and C trends remained similar (Figure 4.8a). The butyrate trend was an overall increase over time, with a statistically significant increase for treatment group C starting from day 28 (Figure 4.8b). Lactate had an overall decreasing trend, with group C decreasing significantly faster (Figure 4.8c). Pyruvate showed an overall increasing trend over time, with a statistically significant late dampening effect given by treatment C (Figure 4.9e). Aspartate, formate, nicotinate, glutamate, propionate and succinate showed no statistically meaningful differences in trends for the treatment groups (Figure 4.8d,e; Figure 4.9a–d). Aspartate showed overall high variability, along with nicotinate. Formate and propionate increased over time with a similar trend for all the treatment groups, whereas succinate decreased over time with a similar trend for all the treatment groups.

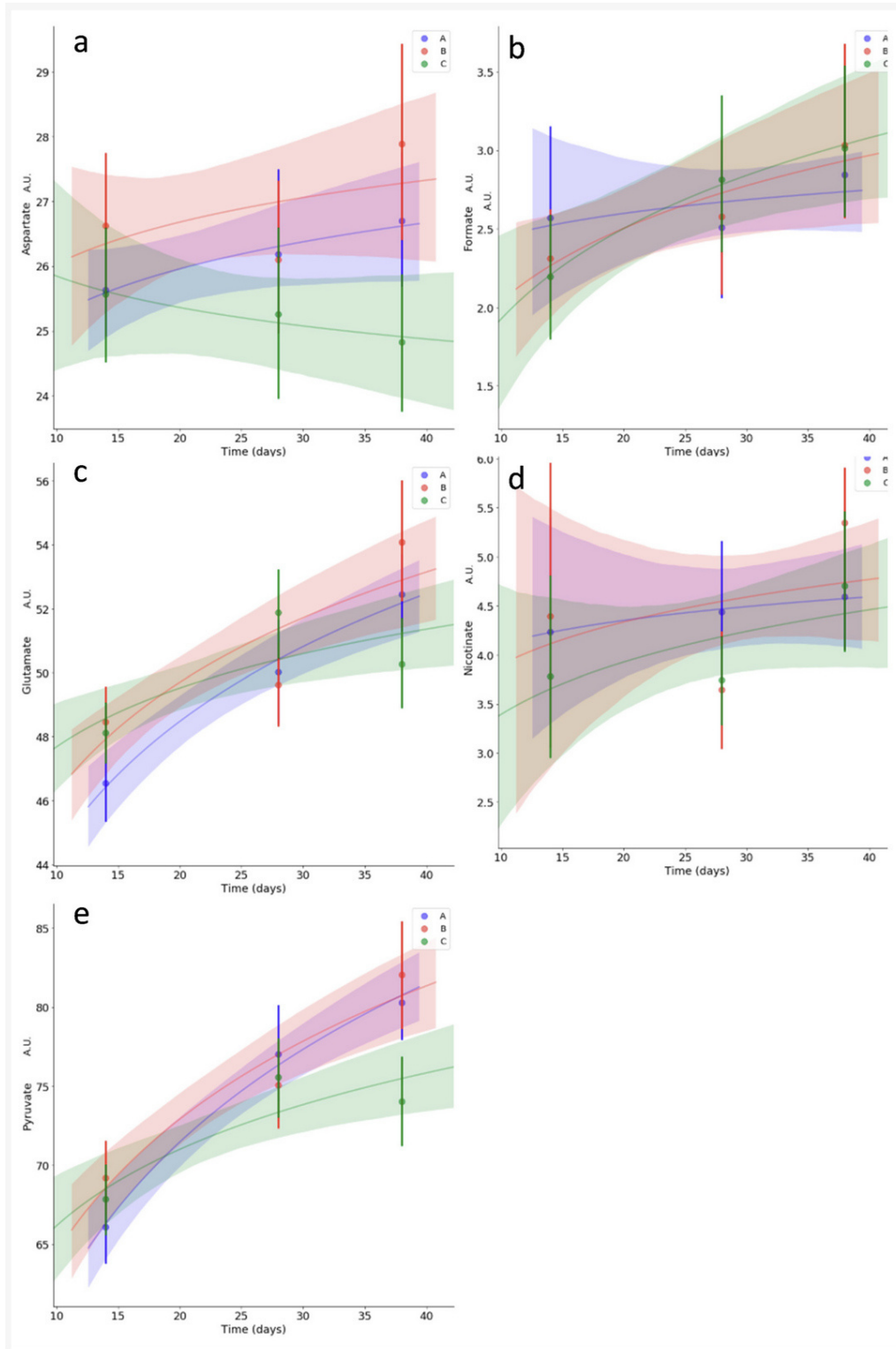


**Figure 4.7:** Partial least square score plots at time points T1 (a), T2 (b), and T3 (c). The support vector machine classifier accuracy score is reported at each time point in order to highlight at which time the effects of the treatment are more detectable in the metabolic space (T2). Red arrows mark the directions of maximum expression in the metabolic space for each metabolite of interest. Points scattered along a particular direction are expected to be characterized by an abundance of the related metabolite. Ac: acetate, Pyr: pyruvate, Asp: aspartate, Lact: lactate, Nic: nicotinate, For: formate, Glut: glutamate, But: butyrate, Suc: succinate, Prop: propionate. MDPI, 2020





**Figure 4.8:** Kinetics for caecal concentration of acetate (a), butyrate (b), lactate (c), propionate (d) and succinate (e). Average values for treatment groups A, B and C are reported in blue, red and green, respectively. Translucent bands represent each fit's 95% confidence interval. Zones of the fits with non-overlapping bands correspond to statistically meaningful differences in trends between groups. Significant differences in trend are seen for acetate, butyrate and lactate. Metabolites are reported using normalized signal arbitrary units (A.U.), proportional to concentration. MDPI, 2020



**Figure 4.9:** Kinetics for caecal concentration of energy metabolism related metabolites: aspartate (a), formate (b), glutamate (c), nicotinate (d) and pyruvate (e). Average values for treatment groups A, B and C are reported in blue, red and green, respectively. Translucent bands represent each fit's 95% confidence interval. Zones of the fits with non-overlapping bands correspond to statistically meaningful differences in trends between groups. Only pyruvate started to show a different late trend for group C with respect to the other treatments. Metabolites are reported using normalized signal arbitrary units (A.U.), proportional to concentration. MDPI, 2020

#### 4.2.4 Discussion and conclusions

The intestinal microbiota of homoeothermic animals constitutes a complex ecosystem composed of a large variety of microorganisms. It plays an important role in maintaining the host's normal gut functions and health, and its imbalance, or dysbiosis, can produce negative effects on gut physiology (183). Since the ban of antibiotics as growth promoters in the European Union, alternative strategies to improve broilers' immunological and metabolic fitness are of great interest. Those strategies involve manipulation of the host-microbiota relationship, through administration of dietary components as well as pro/prebiotics (184; 185). Although not providing a direct substrate for microbial fermentation, riboflavin was reported to influence the gastrointestinal redox state, ultimately modulating the composition of the intestinal microbiota towards an advantageous configuration (186). In the present study, the effects of supplementation of different dosages of vitamin B2 were studied at a model scale. Vitamin B2 supplementation did not affect the ecosystem specificity of the microbial communities, since sample type (caeca, ileum, and litter) remained the main driver of bacterial composition, as previously noted (169; 171). However, the treatment was able to exert a specific effect on both caeca and ileum microbiota components, affecting different bacterial groups and influencing the caecal concentration of different metabolites, depending on the vitamin dosage. Confirming previous reports on the effect of vitamin supplementation on broiler caecal microbiota (187), both vitamin B2 dosages (i.e., 50 and 100 mg/kg) induced an increase of the well-known health-promoting bacteria belonging to the genus *Faecalibacterium* (169) during the first two weeks of the broiler's productive cycle (T1). Moreover, both vitamin B2 dosages also reduced the progressive increase in Rikenellaceae that was observed, through T1 to T2 to T3, in the control group. Indeed, our data showed that OTUs assigned to the species *Alisipites finegoldii* appeared at T3 in the caecal core microbiota of more than 90% of broilers in the control group, whereas this did not happen in broilers in group B and C. This bacterial species had previously been associated with a low food conversion rate (FCR) in broilers (188), whereas *Faecalibacterium* was reported to be positively correlated with FCR, as well as other productivity parameter (171; 188; 189). The highest concentration of vitamin B2 (group C) induced an increase in the abundance of a well-known health-promoting group of lactic acid producers whose genetic makeup lacks enzymes needed for the biosynthesis of this vitamin (*Bifidobacterium*) (190). Interestingly, metabolomics analysis highlighted a progressive decrease of lactate in group C, in favor of butyrate accumulation. This could be explained by the fact that lactate is not usually accumulated in the gut environment, but is consumed as a result of metabolic cross-feeding between lactate-producing and lactate-utilizing bacteria, some of which can use it as a precursor for butyrate synthesis (168). Indeed, the highest dosage of vitamin B2 (group C) seemed to be the one promoting a microbial co-metabolism, leading to a final increased concentration of butyrate, although no significant increase in the abundance of well-known butyrate producers was detected at later time points. On the contrary, the intermediate

concentration of the vitamin (group B, 50 mg/kg) significantly increased the *Bacteroides* abundance in the caeca along the whole productive cycle, with the appearance of an OTU assigned to *Bacteroides fragilis* in the core caecal microbiota. As previously reported (191), the *Bacteroides* increase was to the detriment of the family Rikenellaceae (*Alis-tipes*), also member of the phylum Bacteroidetes. According to a review of the literature, Bacteroidaceae and Rikenellaceae abundances in broilers' guts is strongly influenced by dietary supplements and ingredients (191; 192; 193; 194; 195; 196), and the species *B. fragilis* was already indicated as responding to changes in dietary regimen in broilers (197). Most importantly, an increase in the caecal abundance of *Bacteroides*, and/or the species *B. fragilis*, had already been associated with body weight gain and improved performance (171; 198). The observed changes in microbiota taxonomy in group B were not mirrored at the metabolomics level. Indeed, propionate, which is a common terminal fermentation product of Bacteroidetes, did not increase in the caeca content. This apparent inconsistency may be explained by a subsequent conversion of propionate at higher rates than its increased production from Bacteroidetes, resulting in a null effect on the steady-state concentration of such metabolites.

Looking at proxy metabolites of the energetic metabolism, the only one showing a statistically significant difference with respect to the treatment groups was pyruvate, starting from the late stage of the fitted model for group C. Pyruvate's increasing trend in caeca samples was lower in group C with respect to the other two groups. This may be due to the fact that the energy production progressively decreases with age in all organisms, mainly due to the decline in the function of mitochondria (199). In chickens, such disrupted homeostasis may lead to an increased excretion of involved metabolites and their consequent increasing appearance in the excretory apparatus. Conversely, high doses of vitamin B2 might positively affect the age-related impairment of energy metabolism by slowing it down. Another possible explanation could be related to an increased transformation of pyruvate into butyrate, via Acetyl-CoA intermediate production, operated by members of the Ruminococcaceae family (200). Concerning the ileal microbiota, both diets administered to broilers in groups B and C had a marginal impact on microbial composition. However, it was possible to appreciate that vitamin supplementation counteracted the physiological increase in Peptostreptococcaceae, in particular the increase of an OTU putatively assigned to the genus *Ramboutsia*, a slow-growing taxon known to be detected in the later developmental stage of the ileal microbiota assembly (201; 202). The ecological significance of this taxon still has to be explored.

In conclusion, the supplementation of 50 and 100 mg/kg of vitamin B2 was effective in modulating the composition of caeca microbiota, with a marginal impact also on ileal community structure. In particular, the supplementation of vitamin B2 at 50 mg/kg significantly increased the *Bacteroides* abundance since day 14 up to the end of the rearing cycle. Moreover, the highest dosage of vitamin B2 (100 mg/kg) significantly increased the abundance of *Bifidobacterium* starting from day 28 up to 42 days. This microbiota

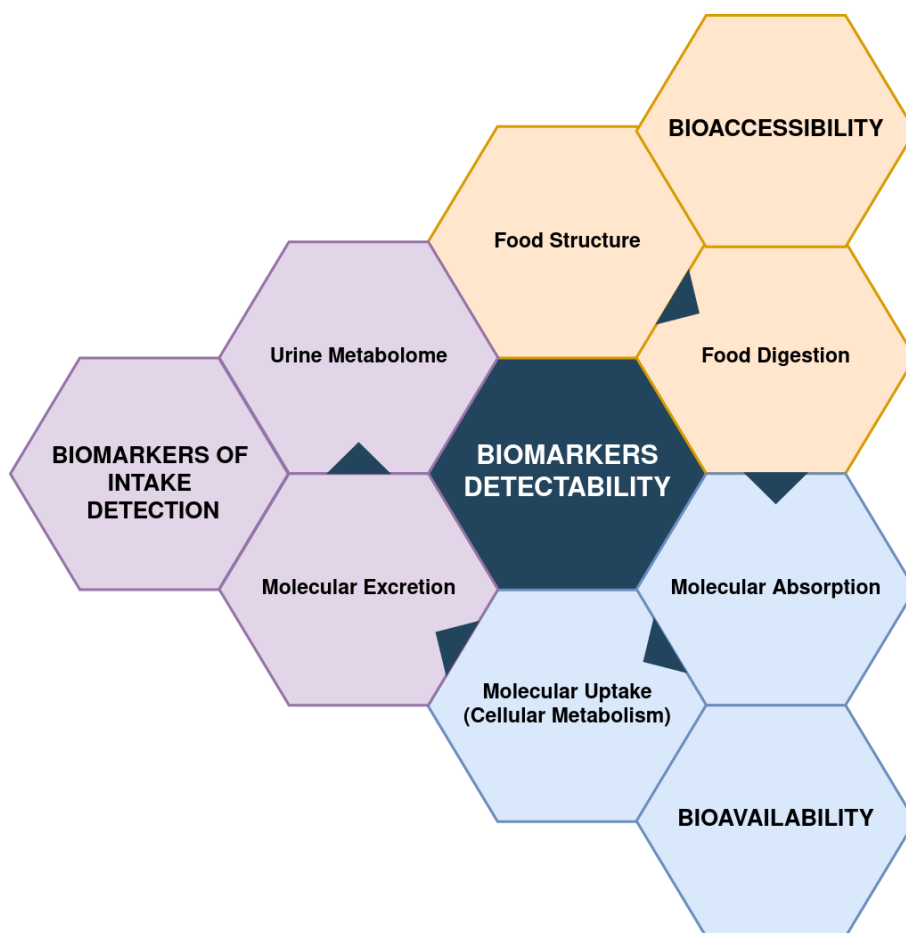
modulation resulted in the boosted production of butyrate, which plays an important role in protection against pathogens in poultry (203). Furthermore, butyrate is involved in several intestinal functions, being an energy source stimulating epithelial cell proliferation and differentiation, other than exerting an antimicrobial effect by promoting the production of peptides and stimulating the production of tight junction proteins (204). Therefore, the proposed nutritional integration could positively affect the host's fitness in reacting to pathogenic infections through a butyrate-mediated improvement of epithelial integrity in the caeca and positive stimulation of the immune system.

### 4.3 Multi-Compartmental Model of Complex Compounds

When modelling complex compounds such as real life food, biomolecules release and their interaction with the human organism have to be modeled at two main levels linked to their composition: the release of the molecules from the matrix and how they interact with the ensemble of functions of the organism that become active from the moment the molecules become bioavailable. The intertwined nature of the relationships intercurring between these two levels, ultimately affects biomarkers of intake detection from a kinetic point of view. In this section, we present and discuss a multi-compartmental simulation reconstructed from urine metabolomes, to evaluate the impact of real life food intake in different individuals. The results hereby reported are part of a paper in preparation, using data from the FOOTBALL project (<http://football.org>). These early stage results were presented by the author at the 6th edition of the International Conference in FoodOmics, Cesena, Italy, 2020, in an award-winning oral presentation.

#### 4.3.1 Interindividual Variability in Bioavailability

Figure 4.10 summarizes the complexity involved in studies relying upon biomarkers detectability, which simultaneously depends on many elements that are in turn linked by relationships of sequentiality and causality. The interaction between food and the human organism begins with bioaccessibility, describing how nutrients can be accessed from their matrix. Food structure and its digestion are the first factors affecting detectability of biomarkers, introducing the timescale of the overall kinetic description. When biomolecules are released and become accessible, competing functions of the organism start the processes that ultimately lead to bioavailable substances. Regarding nutrition, bioavailability can be defined as the fraction of ingested dose that is absorbed, and consequently used or stored. The bioavailability of nutrients is thus heavily related to individuals, being a function of their physiological and nutritional status. Co-ingestion, mucosal mass and individual cellular metabolism needs are all factors tied to the status of the absorbing subject (205). Lastly, mechanisms of molecular excretion and related kinetics constitute the tiles completing the complex mosaic of the urine metabolome. In a hypothetical experiment where the only observable related to metabolite concentration



**Figure 4.10:** Overview of the downstream of elements influencing biomarkers detectability in urine spectra. Dark arrows denote a relationship of sequentiality.

is the urine metabolome, a way to take multi-compartmental kinetic and interindividual complexity into account is to be found.

The GutSelf review (206) provided an extensive overview of intervariability factors of digestion of dietary compounds in the GIT (gastro intestinal tract). According to Walther et al., the analysis of interrelated factors tied to GIT functionality (oral processing, intestinal digestion and absorption) is strictly tied to the specificity of each class of nutrients (protein, fat, carbohydrates, vitamins, and minerals). These can be classed as intrinsic (e.g., genetic polymorphisms) or extrinsic (e.g., diet), molecular (e.g., pepsin activity), or morphological (e.g., BMI) and of genetic (e.g., amylase polymorphism) origin. The elements of GIT functionality that are affected by individual variability, as published evidence suggests, are: chewing, nutrient sensing, saliva composition, nutrient digestibility, composition of the intestinal peptidome, enzymatic activity, genetic polymorphisms and the gut microbiota. These levels of complexity, contributing to the overall variability characterizing the bioavailability of nutrients interacting with the human organism, require a large amount of data output and models capable of mirroring the multifactorial analysis



necessary for a personalized food-human interaction characterization.

### 4.3.2 The Bateman equations

From a modeling perspective, a simple yet effective way of representing interdependent kinetics, describing events linked in a sequential way starting from given initial conditions, is the integration of a chain of differential equations. A system of differential equations with such characteristics, commonly used in pharmacokinetics studies, is the Bateman equations system (207):

$$\begin{aligned} \frac{dN_1(t)}{dt} &= -\lambda_1 N_1(t) \\ \frac{dN_2(t)}{dt} &= -\lambda_2 N_2(t) + \lambda_1 N_1(t) \\ &\vdots \\ \frac{dN_i(t)}{dt} &= -\lambda_i N_i(t) + \lambda_{i-1} N_{i-1}(t) \end{aligned} \tag{4.2}$$


Where each  $N_i(t)$  represents an absorption-release process in a given compartment. Theoretically, with an observable related to metabolite concentrations such as NMR spectra acquired at different time points, kinetics parameters for all compartments can be estimated. For each compartment, the scalar  $\lambda_i$  represents the rate at which biomolecules are transferred to the next compartment, while  $\lambda_{i-1}$  is the rate at which biomolecules are pumped in from the previous. Furthermore, one can interpret  $N_1$ , the initial concentration of the first exponential decay of the system, as the point of the bioavailability curve at which absorption processes start, expecting different results for  $N_1(t_0)$  for different matrices of administered foods and different individuals. The core idea of the framework presented in this study, is to extract a time dependent latent component of NMR spectra of urine from different individuals, using spectra acquired at different time points for different administered food. This latent component, containing information about metabolite concentrations at different times, will serve as the observable to be numerically integrated to solve a Bateman system. By solving a 3-compartment Bateman system, one can theoretically reconstruct blood/transport kinetics and bioavailability effect from urine observation. The distribution of the parameters estimated while solving such a system, could provide useful insights in characterizing how different foods are digested and how different individuals digest the same type of food.

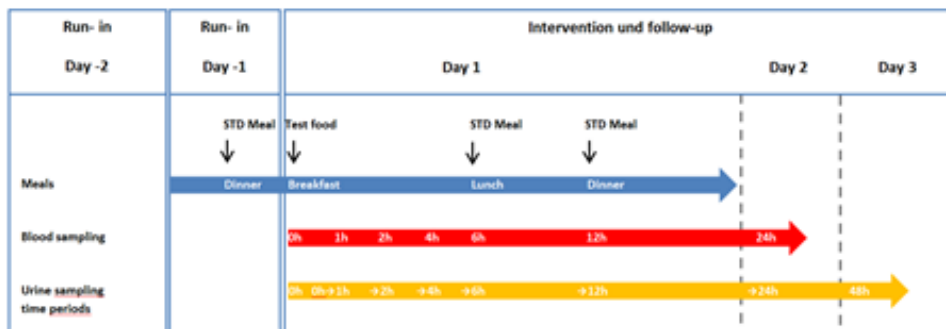
### 4.3.3 Dataset and modelling

The FOOTBALL study (trial samples courtesy of *Pieter Giesbertz, Yu-Mi Lee, Beante Brandl, Thomas Skurk, ZIEL-Institute for Food and Health, Technical University of Mu-*



## 3 Intervention trials:

**Chicken breast**  **Wheat bread**  **Egg** 



2

**Figure 4.11:** Experimental designs of acute dietary interventions

*nich*) provides observations for three different types of dietary acute interventions over a 48 hours timespan (fig.4.11)

- 766 NMR Spectra (urine)
- 31 individuals for 3 different interventions, randomized crossover
- 8 time points (0,1,2,4,6,12,24,48 hrs) of acute dietary intervention monitoring
- **Chicken Breast Intervention (CH):** Volunteers randomly received on three different occasions as breakfast either 100g or 200g chicken breast (both with rice) or rice alone as control food. Control food consists of 125g rice, 30g margarine, and 1,5g salt.
- **Fibers Intervention (TST):** Volunteers randomly received on three different occasions as breakfast either 5g inulin or 2,5g beta-glucan (solved in water, both with toast) or toast alone as control food. Control food consists of 75g toast without crust.
- **Egg intervention (EGG):** Vegan or vegetarian volunteers who have eaten vegan food for 2 weeks received for breakfast either 2 boiled eggs or rice as control food.

### NMR Sample Preparation and Analysis

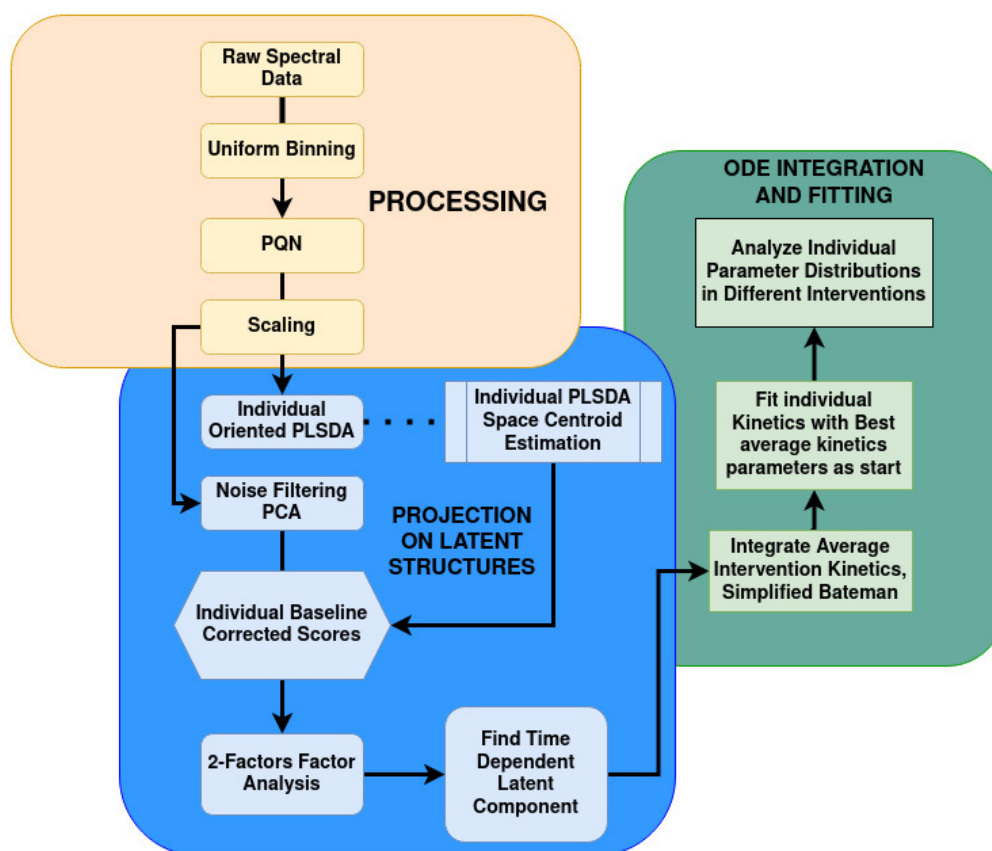
All urine samples have been first collected and stored at  $-80^{\circ}\text{C}$  and then prepared for the NMR analysis according to Linda H. Munger et al. 630  $\mu\text{L}$  of urine sample were first centrifuged to remove debris and then 540  $\mu\text{L}$  of supernatant were placed in a clean microfuge tube containing 60  $\mu\text{L}$  of D<sub>2</sub>O-based phosphate buffer (1.5MKH<sub>2</sub>PO<sub>4</sub> in 100% D<sub>2</sub>O, pH7.4), also containing 2mM sodium azide (NaN<sub>3</sub>) as an antibacterial agent and 10mM TSP (= 3 – (trimethyl – silyl)propionicacid – d<sub>4</sub>, Aldrich 269913) as an internal standard. A total of 590  $\mu\text{L}$  of the mixture was transferred into 5mm NMR tubes. All the <sup>1</sup>H NMR spectra were recorded at 298K and acquired using a Bruker 600MHz spectrometer (Bruker BioSpin, Karlsruhe, Germany) operating at 600.13MHz proton Larmor frequency and equipped with an autosampler with 60 holders. For each urine sample, a one-dimensional (1D) NMR spectrum was acquired with water peak suppression applying the NOESYGPPR1D sequence (a standard pulse sequence included in the Bruker library). Each spectrum was acquired using 32K data points over a 7211.54Hz spectral width (12ppm) and summing up 128 transients. A 90 degrees pulse of 12.5  $\mu\text{s}$  was set up. A delay of 5s between transients, extending the acquisition time of 2.27s, was chosen to provide a recycle time 5 times longer than the longitudinal relaxation time of the protons under investigation, expected to be not longer than 1.4s. The data were Fourier transformed and phase corrected, baseline corrections were automatically applied using TopSpin version 3.0 (Bruker BioSpin, Karlsruhe, Germany).

### Spectral Analysis

Free induction decays (FID) were multiplied by an exponential function equivalent to a 1.0Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated using TopSpin 3.0 (Bruker BioSpin, Karlsruhe, Germany). Spectra were aligned calibrating the TSP peak at 0.00ppm and the spectral regions including only noise (e.g., the spectrum edges below 0.5 and above 10ppm), as well as the data points which are strongly affected by the residual water (between 4.95 and 4.7ppm) and the urea signals (5.45 – 6.1ppm), were removed before data analysis. Each 1D reduced spectrum in the range between 0.5 and 9.00 ppm was segmented into 0.02ppm chemical shift bins (100 spectral points). Prior to pattern recognition, the bucketed dataset underwent a normalization using a probabilistic quotient normalization (PQN) as a scaling method. All resonance of interest were assigned on the 1D NMR by comparing their chemical shift and multiplicity with Chenomx software data bank (version 8.1, Edmonton, Canada) and with literature when available.

### Data Analysis Pipeline and Numerical Integration Methods

All data analysis, machine learning, pipeline implementation and numerical integration has been carried out using *Python 3.8*, mixing functions from standard packages and cus-



**Figure 4.12:** Pipeline structure. Each colored box represents a different phase of the analysis

tom scripts. Data handling and database construction was performed using the *Pandas* library. Spectral projection on latent structure was performed using Principal Component Analysis, Factor Analysis and Partial Least Squares Discriminant Analysis functions from the *SciKitLearn* package. Prior to Factor Analysis decomposition of the spectra and after their normalization, samples underwent a noise filtering PCA. To minimize variability tied to individual metabolic phenotypes, which constitutes an important confounding factor (208), the coordinates in a 20-dimensional PLSDA scores space, trained to discriminate between different spectra of the same individual, has been computed. These centroids have been used to perform an individual baseline correction in the PCA scores space. The emerging time dependent latent component was numerically integrated, to compute average kinetics for each intervention group, using custom scripts based on *Scipy* ODEint function and the *scipy.optimize* module. The 95% confidence interval of the fit has been estimated using a resampling bootstrap technique, using realization drawn from a multivariate gaussian distribution of the parameters based on the estimation of the parameter covariance matrix using functions from the *scipy.stats* module. Figure 4.12 provides a complete overview of the pipeline.

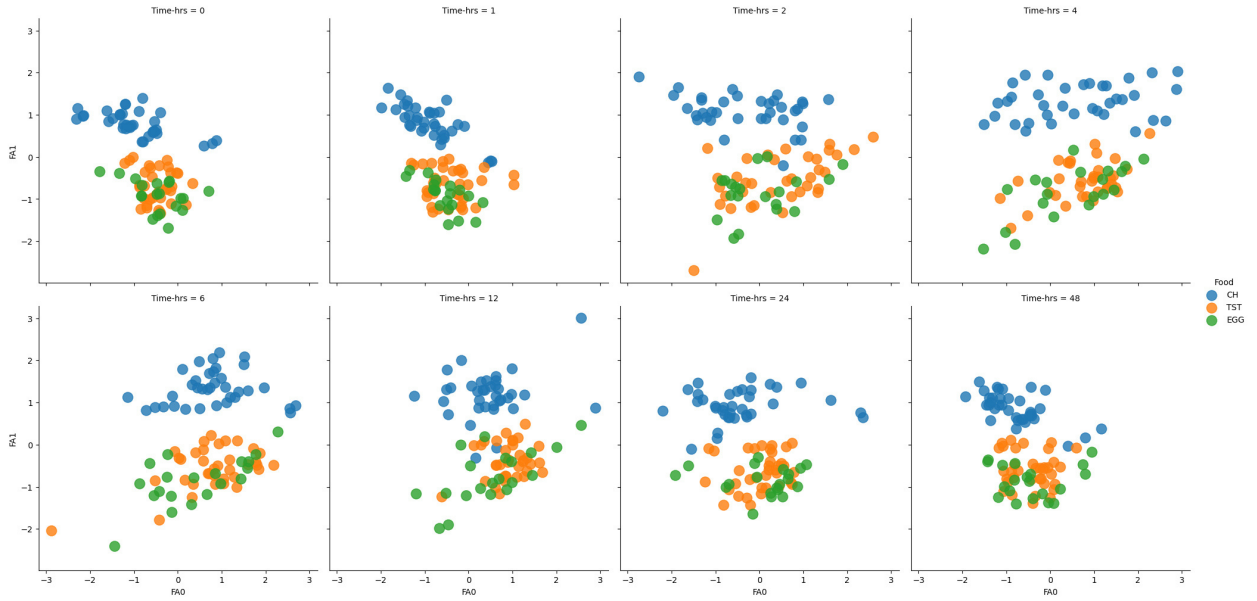
#### 4.3.4 Early-stage results and discussion

The endpoint of the pipeline in 4.12 can be summarized in the following way :

- **Theoretical Benchmarks:** We can predict kinetic parameters for all compartments starting from the one we can observe. Furthermore, if we have a group of individuals eating the same food, we can characterize the variability of their digestive functions by confronting kinetics parameters
- **Requirements:** An observable related to metabolite concentration : spectra acquired at consecutive time points. A frequent time sampling ( $\sim (N_{obs} - 1)/2$ ) required for the estimation of the parameters in all compartments.

#### Extraction of observables with 2-Factor Factor Analysis

From the experimental design and after individual variability correction, the hypothesis on spectral data is that they contain two main sources of variance: the different interventions and the different time-points at which the spectra are collected after the intervention. We thus look for a suitable latent space capable of describing these sources of variance, to assess which metabolites can be considered proxies of food intake and which changes in metabolites concentrations can be used to describe kinetics. This situation, with an available strong a-priori hypothesis about possible factors of variance in spectral data, is particularly suitable for a factor analysis framework, as discussed in 1. The use of factor analysis allows for an unbiased, unsupervised dimensionality reduction, with a powerful advantage in interpretation if the projection is successful. A 2-factor factor analysis of the full dataset, plotted for each sampled time point for visualization purposes, yielded the results in Figure 4.13. The resulting latent components fully reflect the hypothesis on the dataset: FA1 is the latent component describing intervention separation, while FA0 is the latent component describing sampling at different time points. Thus, the analysis of loadings of FA1 should result in a description of metabolites separating the different interventions. Accordingly, a preliminary analysis of FA1 loadings (in progress) indicated creatinine concentration as the strongest factor separating the meat-rich intervention from the other two. FA0 scores show a high variability as a function of time sampling: this makes FA0 the component containing the linear combination of spectral features that retain kinetic information. FA0 scores can be used as the observable to numerically integrate the Bateman equations, characterize average kinetics for each intervention and estimate kinetic parameters to characterize individual with their distributions. The analysis of FA0 loadings (in progress) should result in a description of proxy metabolites whose changes in concentration summarize kinetic aspects.



**Figure 4.13:** Score plot of 2-factors Factor Analysis decomposition of the spectra, after individual baseline correction. FA0 (x-axis) emerges naturally as the time-dependent latent component. FA1 (y-axis) emerges as the component containing separation between meat-rich/meat-less interventions. The two components return an orthogonal representation of these two main factors of variance in the dataset, in agreement with the a-priori hypothesis.

### Trade-off solution: Simplified Bateman Integration

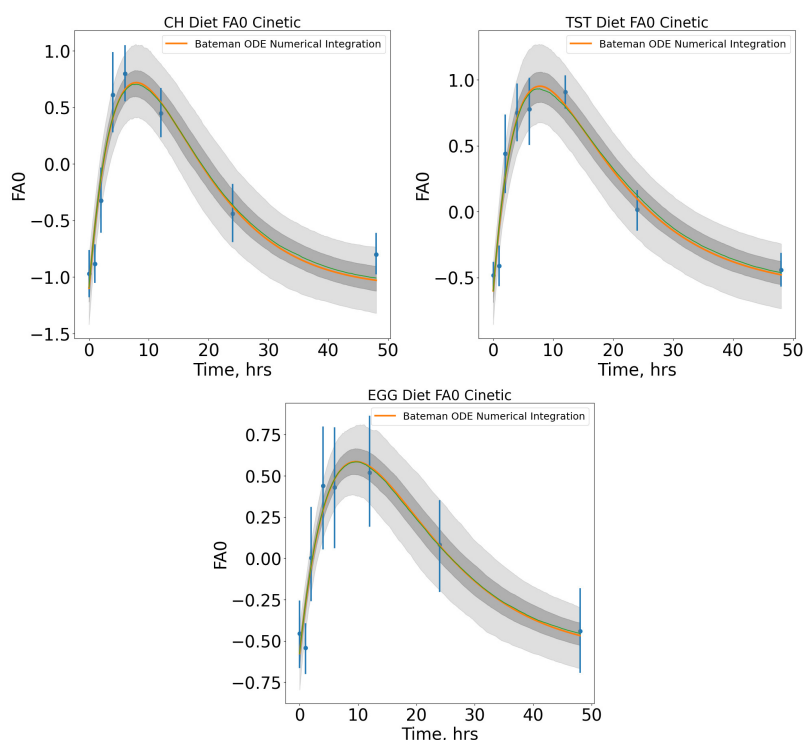
With the current experimental design and available sampled time points, the theoretical complete solution of a Bateman differential equation system could not be achieved. With 8 available observed time points, approximately  $\sim 3$  parameters could be integrated numerically so that their covariance matrix is not singular. This in turn means that the stability of a 3 parameters-only model could be estimated to evaluate the robustness and efficiency of the approach and characterize individual kinetics. From an interpretative point of view, this translates to the fact that the sampling is not sufficient to predict intertwined parameters of blood/transport and interaction with food structure (bioaccessibility-bioavailability). However, a trade-off solution should be enough to evaluate the kinetic effect of different food composition for different individuals. To simplify the model and reach numerical stability, we "compress" the information of individual variability of digestive functions and bioavailability of different foods into a single exponential decay. The Bateman system to solve becomes:

$$\begin{aligned} \frac{dN_1(t)}{dt} &= -c_1 \lambda_2 N_1(t) \\ \frac{dN_2(t)}{dt} &= -\lambda_2 N_2(t) + c_1 \lambda_2 N_1(t) \end{aligned} \quad (4.3)$$

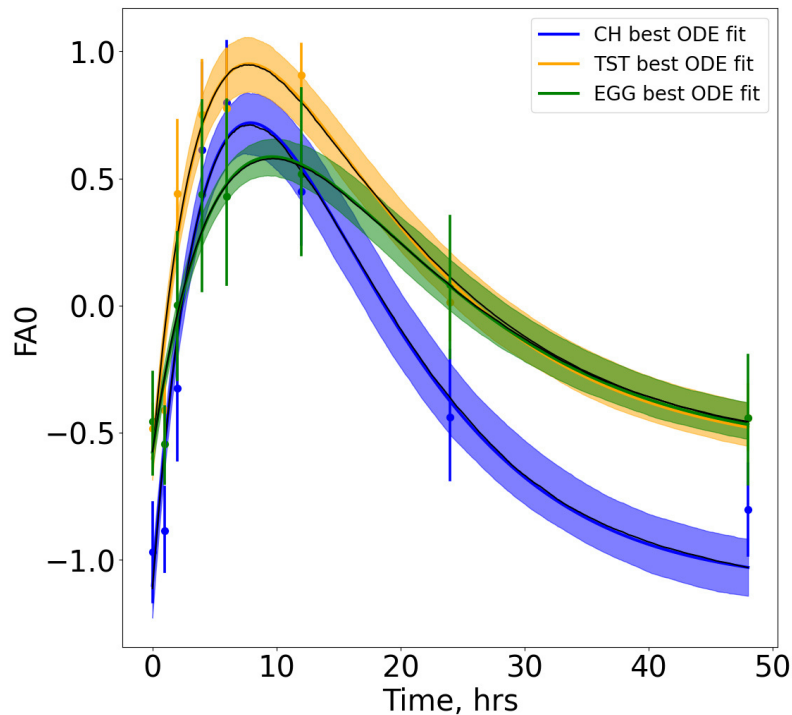
With this form of the Bateman equation, the parameters to be estimated by model are:

- $N_1(t_0)$ : the initial concentration of bio-molecules before excretion starts, containing the downstream information of bioavailability and blood transport/molecular uptake
- $\lambda_2$ : the exponential decay constant of the observed compartment (kinetic of excretion observed by urine spectra)
- $c_1$ : the constant to estimate  $\lambda_1 = c_1\lambda_2$ , which the exponential decay constant of the top compartment kinetic, expressed as function of  $\lambda_2$  (estimated within the observed compartment kinetic) to enhance numerical integration stability.

In such a way, we reduced parameters estimation to a suitable number. **Note:** to enhance numerical precision of the optimizer and numerical integrator, the initial guesses of parameter array is transformed to  $(e^{N_1(t_0)}, e^{\lambda_2}, e^{c_1})$ . This is a known computational trick to ensure the correct sign for each parameter values (so that numerical integration doesn't break) and to enhance float numbers representation precision. We then fit individual kinetics with the average optimal parameters estimation as a starting point for the numerical optimizer. Average fitted kinetics are reported in Figures 4.14, 4.15



**Figure 4.14:** Result of numerical integration and average kinetics fitting of the 3 acute interventions. Light grey shade area marks the 95% confidence interval of the model. Dark grey shade area marks the 75% confidence interval of the model. EGG model has the highest variability due to a lower sample size compared to the other two interventions. ODE = ordinary differential equations.



**Figure 4.15:** Superimposed plot of average kinetics. The light-colored shades in each plot mark the 75% confidence interval. Differences for kinetics related to different interventions are evident in non-overlapping zones of the fit.

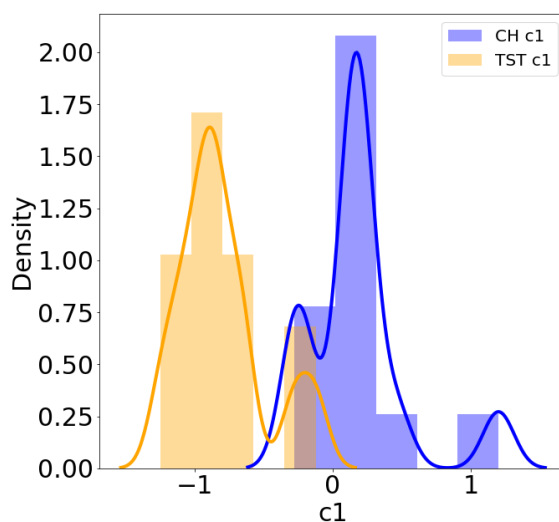
### Individual characterization through parameters distributions

The parameters computed to fit average kinetics, are assumed to be the best guess for the stating parameter in the computation of individual kinetics for each intervention (as described in Figure 4.12). Within the current assumptions for the model, the characterization of how the intake of different foods impacts different individuals can be thought as follow:

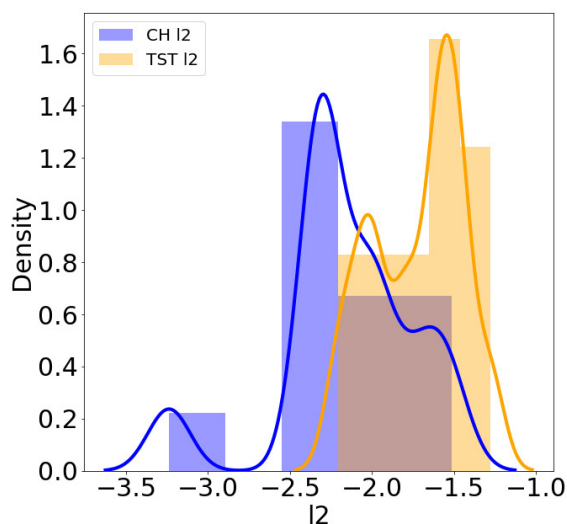
- The morphology (number and sharpness of maxima) of the distribution of the parameter  $c_1$  can highlight group of individuals with different digestive phenotypes, or enterotypes, in terms of how the bio-molecules are assimilated and transported after they become bioavailable.
- The morphology (number and sharpness of maxima) of the distribution of the parameter  $\lambda_2$  can highlight group of individuals with different digestive phenotypes, or enterotypes, in terms of how bio-molecules and their byproducts are excreted through urine
- The comparison of  $c_1$  and  $\lambda_2$  distributions can highlight if the approach is able to discriminate how different foods are assimilated and excreted.



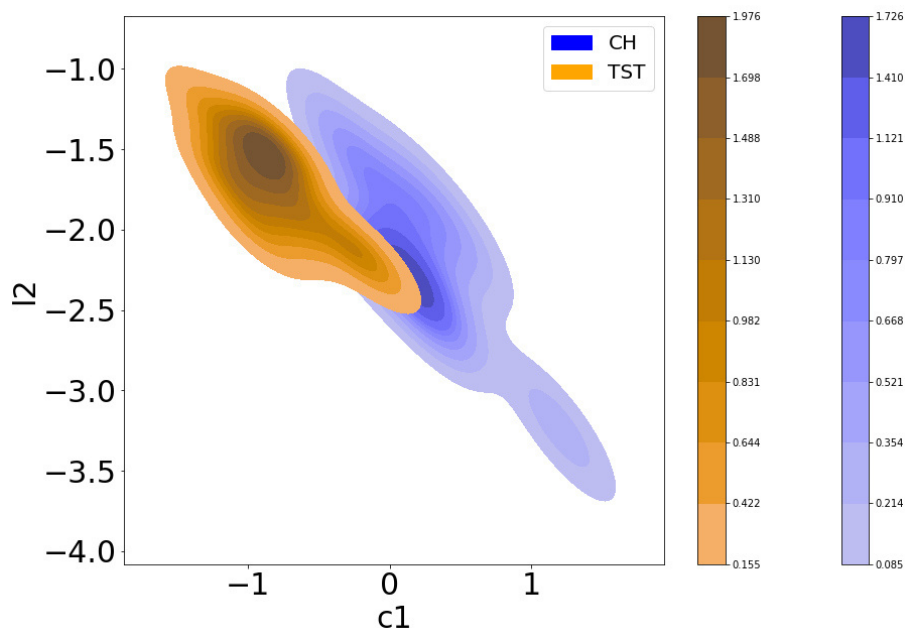
The results are reported in Figures 4.16, 4.17, 4.18. Of note, the fitting of individual kinetics for the EGG intervention resulted in distributions with a very high second central moment (variance in more general terms). This is due to the low number of people that completed the EGG intervention, compared to the other two interventions (160 spectra in total over 8 time points, versus 290 spectra in total over 8 time points for the other two). The effect of this low sample size is also reflected by the high standard error of the average fit in Figure 4.14; for this reason, parameters distributions results for the EGG intervention are not reported and discussed. The parameters computed by the models, as reflected by their distribution, offer insightful information about how different food are assimilated by different people. The moments of the distribution of parameters of different foods are significantly different, as confirmed by the Kruskal-Wallis test results reported in Figures 4.16, 4.17. This means that the trade off-solution of the model is still able to discriminate the impact on physiological functions of the GIT of complex real-life foods with different composition, at both compartments of the model. Furthermore, the different maxima emerging in distributions mark the presence of groups of individuals with different enterotypes. Figure 4.18 offers a perspective of the crosstalk between the two compartments of the model, in terms of different phenotypes of digestive functions studied at two intertwined level from a single observed compartment (urine metabolome). Specifically, the CH intervention resulted in a joint distribution with a local maximum in extreme values of the parameters, highlighting a greater heterogeneity in the response to meat-based products. This is maybe due to the overall high number of elements (such as processing, fermentation, tenderness..) contained in meat-based products that can impact proteolysis, which is in turn tied to interindividual intestinal peptidome composition, protease activity and genetic polymorphisms (206). The TST intervention joint distribution has a sharp maximum around a very different range of parameters values with respect to the CH intervention, with a nearer local maximum. This denotes a characteristic timescale for the kinetics of this intervention, mainly focused on carbohydrates and fibers, which also finds an overall homogeneous response from individuals with the respect to the CH intervention. The underlying cause of the emerging stratifications, that can shed light upon different enterotypes, is still under investigation using clinical data (such as age, BMI, sex, clinical parameters, lifestyle) from patients that underwent the trial. An etiologic interpretation of parameters values, giving rise to the different timescales of the phenomena in the compartments of the systems is also in progress. Furthermore, loadings analysis of FA0 and FA1 are being performed, to find patterns of metabolites describing different aspects of kinetics and food intake impact.



**Figure 4.16:** Distribution of  $c_1$  parameters computed for individual kinetics. The continuous line is a kernel density estimation of the probability density function. The presence of distinct maxima in the same distribution, reflects the presence of group of individuals with different enterotypes reacting to the same food intake, in terms of bioavailability, uptake and transport. The separation between the two distribution remarks the ability of the model to characterize kinetics of different food intake from urine (Kruskal test  $p$ -value =  $2.92 \times 10^{-5}$ ).



**Figure 4.17:** Distribution of  $\lambda_2$  parameters computed for individual kinetics. The continuous line is a kernel density estimation of the probability density function. The presence of distinct maxima in the same distribution, reflects the presence of group of individuals with different enterotypes reacting to the same food intake, in terms of excretion and byproducts permanence in urine. The separation between the two distribution remarks the ability of the model to characterize kinetics of different food intake from urine (Kruskal test  $p$ -value =  $4.42 \times 10^{-3}$ ).



**Figure 4.18:** Joint KDE plot of  $c_1$  and  $\lambda_2$  parameters computed for individual kinetics. Darker spots in the plot correspond to high density zones of the multivariate distribution (maxima). The presence of different maxima reflects the presence of individuals with different enterotypes, considering the two compartments of the model simultaneously.

#### 4.4 Chapter Conclusions

In this chapter, we introduced a framework that emphasizes the importance of time-resolved experiments and models to evaluate the impact of food intake and diet within a holistic paradigm. The foodomic approach points toward the directions needed for an evaluation based on molecular level descriptions, which can deepen the understanding of the impact of complex food compounds by considering composition. The kinetic description of molecular states is required to study the impact of different foods and biomolecules, as their effect on the organism stretches over different time scales. This specific aspect is what makes an evaluation of exposure to diet based only on epidemiological data incomplete. We proposed and discussed two different frameworks to show how the organism can be studied, to assess the effect of compounds of different compositions at different timescales. The first is a prolonged exposure to a bio-active compound, that can be studied by the crosstalk of metabolomic and microbiomic changes during the entire lifespan of an organism. The second is a characterization of individual response to short-term acute exposures to different diets, through the reconstruction of metabolic kinetics from urine metabolome. Food composition is tied to the bioavailability of their contents, which in turn affects how physiological functions of digestion are activated in different individuals. Within the right frameworks, it is possible to take this complex aspect of food-organism interaction into account and produce predictive models to stratify

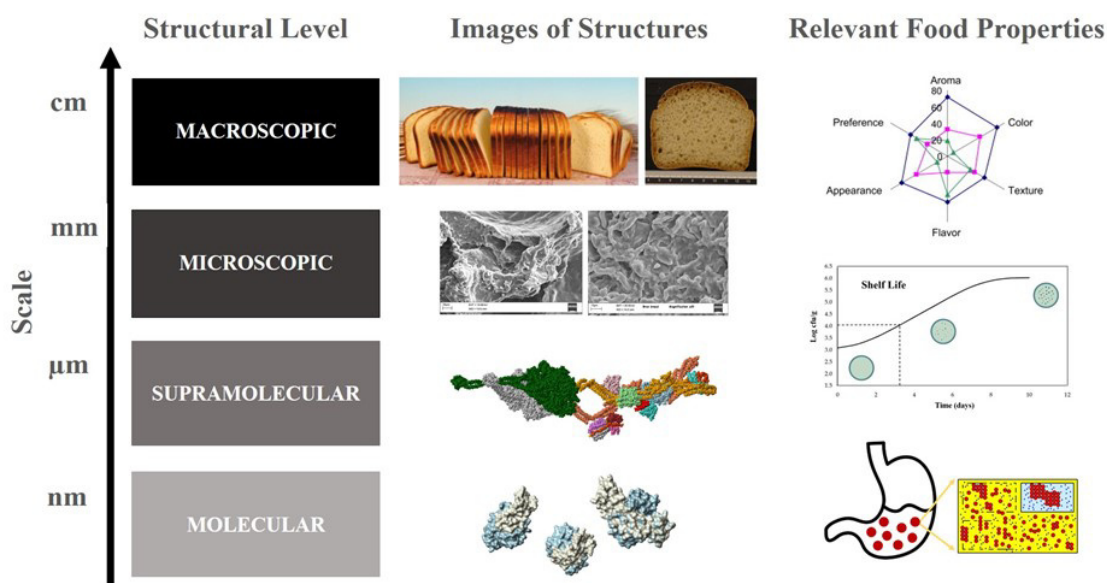
populations. However, bioavailability is ultimately tied to bioaccessibility, which cannot be modeled by considering only composition. Bioaccessibility, or the fraction of a given food compound released from the food matrix, is obviously linked to the nature of the food matrix itself. To reach the modelling of food as a complex ensemble of molecules, encased in heterogeneous phases, that interacts with the human organism perturbing molecular states, a way to describe different food matrices is needed. In the next chapter, we provide an overview on how artificial intelligence can be exploited to define food matrices by modelling their structures.

## Beyond Composition: the Role of Structure in Physiological Interactions

This chapter is based on the work by Mengucci, Capozzi et al., 2021, which is, at the moment of the writing of this thesis, under review for the European Federation of Food Science and Technology (EFFoST) special issue of *Trends in Food Science and Technology*, Elsevier (accepted and published during the revision of this thesis).

### 5.1 Food Structure, Function and Artificial Intelligence

Although the description of food has traditionally been based on analytical chemical composition, many of the important properties of food are determined by structural elements. This limitation in the descriptive capacity of a food is also reflected in many mathematical models that currently aim to predict the sensory, functional, and nutritional properties, including for example digestibility. For this reason, the contribution of the structure of a food is often overlooked including when studying the effect of diet on health. In fact, the nutritionist tends to consult compositional databases when the correctness of a diet must be evaluated, having no indication on how to use any structural data even when available. Nevertheless, before collecting structural information, it would be necessary to establish how to use them to build predictive models for nutritional functions that depend on it. Understanding how the ingredients and each unit operation of food processes make up the structure of the foods and how this structure changes during its life or on eating will play a main role in the development and management of the food science and industry. For this reason, a tailored collection of scientific work described in the literature has been examined to pave the way for a future approach using matrix structural data to predict food functions, also exploiting artificial intelligence (AI).



**Figure 5.1:** Food matrix is defined by structures at different length scales consisting of elements spanning nanometres to millimetres and above. Many of the important properties of foods are determined by structural elements at micro-scale. Molecules such as carbohydrates, proteins, and lipids, indeed form supramolecular clusters that behave as pseudo-molecules of higher molecular weight. Linking organised structural elements to food properties through imaging may be feasible by means of artificial intelligence applications.

### 5.1.1 What is the structure of a food?

Most foods are complex, heterogeneous materials composed of structural elements or domains (co-) existing as solids, liquids and/or gases, where length scales span nanometres to millimetres (209). Many of the important properties of foods are determined by structural elements of micro-scale and above, such as bubbles, drops, strings and particles (210). Food products consist largely of carbohydrates, proteins, and lipids, forming clusters that behave as pseudo-molecules of higher molecular weight than the individual constituent molecules (210). These interactions are primarily hydrogen-bonding interactions between the hydroxyl groups or Van der Waals interactions between nonpolar molecules, but also ionic or covalent bonds, such as disulphide or isopeptide, may be very important. The supramolecular organization of foods gives rise to their structure. Complex food structures are formed, not because of the abundance of elemental components, but because of the multiple interactions that proteins, lipids and polysaccharides undergo at different conditions in an aqueous medium. In natural and processed foods, the structure (or matrix) of a food is defined as the organization of its constituent molecules at multiple spatial length scales (209). At one extreme, a food product is macroscopic, and at the other extreme, it is composed of molecules and atoms characterized by molecular length scales (210). The matrix of a food is in fact scale-sensitive, i.e., interactions may take place at several scales in the same food as shown in Figure 5.1

For example, the matrix in a bakery product responsible for the textural properties of the porous crumb are the protein-starch walls surrounding the air cells (211), and the relevant scale is on the order of a few hundred microns (212). Starch granules undergoing gelatinization may be regarded as inclusions in the continuous gluten matrix at a scale of approximately  $10\ \mu\text{m}$  (213). At the nanoscale, gelatinized starch granules are the matrix onto which  $\alpha$ -amylases exert their action during digestion to release glucose molecules (214). By and large, foods are systems of dispersed phases, such as mesoscale particulate structures (colloids) derived from natural food products constructed by self-assembly (e.g., granules, micelles, globules, and fibres) or are created artificially via food processing (215). Next to these mesoscale structures, food contains smaller molecular species, like salts, sugars, polyols and phospholipids, which moderate the properties of the continuous or dispersed phases, or their interfaces. The structure of a given food depends, however, enormously on the product, its constituents and which of the many length scales are dominant in establishing the product properties (210). For an emulsion-based food such as mayonnaise, it is the droplet size of around  $1\ \mu\text{m}$  which is the relevant length scale, whereas for dairy products it is typically the size of a casein micelle ( $\sim 50\text{-}100\ \text{nm}$ ) (216) and the size of the individual casein subunits ( $\sim 2\ \text{nm}$ ) that matter. The relevant length scale of food powders is typically between  $10$  and  $500\ \mu\text{m}$ , and the structure of starch is described at length scales between the macromolecular ( $\sim 1\ \text{nm}$ ) and the size of the starch granules ( $\sim 1\ \text{mm}$ ). Even length scales substantially smaller than  $1\ \text{nm}$  matter in foods, as diffusion and the interaction of water with the food matrix occur at these distances. Food structure is important at all dimensional scales for texture, sensory properties, shelf life and stability and can alter the kinetics and extent of food digestion (209; 217). It plays a vital role in how food interacts with the gastrointestinal tract (GIT) (e.g., bodily fluids and receptors) and the resulting release and uptake of nutrients (209) and post-prandial outcomes (218). In addition, the breakdown of the food matrix is a major controlling factor for the perception of texture and flavour in the mouth (219).

### 5.1.2 How to measure food structure

Several techniques can be applied to measure the structure of food materials either directly (optical and confocal microscopy, tomography, scanning and electron microscopy) or indirectly from measurements of the mechanical response or spectroscopy (5.1). Some challenging techniques such as Differential Scanning Calorimetry (DSC) (220), Thermogravimetric analysis (TGA) (221), Nuclear Magnetic Resonance (NMR) spectroscopy and relaxometry (222), Near-Infrared Reflectance spectroscopy (NIR) (223), Attenuated Total Reflectance (ATR) spectroscopy (224) and FT-Raman spectroscopy provide quantitative parameters that are related to the interactions among molecules, thus making measurable physical-chemical properties that depend on the supramolecular structure of the food matter. However, imaging techniques are essentially dedicated to the investigation of the real 3D structure (225). Static Bragg-type diffraction of neutrons and X-rays has been



applied to either fluid or viscous food systems to reveal the structure in the 10–100 nm length scale range (210). Insight into lipid polymorphism, liquid crystallinity, protein folding, etc. can typically be gained by using these techniques. Because most common food properties are, however, directly related to the  $\mu\text{m}$  length scale, light scattering techniques are primarily exploited. The application of the dynamic light scattering (DLS) experiment to foods yields information on the diffusion coefficient of the scattering objects (209). Tomographic techniques such as magnetic resonance imaging (MRI) and X-ray tomography are extremely powerful since they allow a full 3D reconstruction of the sample structure but tend to be limited in resolution and/or slow in acquisition times. Optical or Light Microscopy (LM) suffers from a similar limitation in resolution, in this case due to the wavelength of visible light, even though structures of the order of  $1\ \mu\text{m}$  can still be imaged using confocal microscopy. A further limitation of optical techniques is that the food sample should be sufficiently transparent. Conversely, a major advantage of optical microscopy is that dynamic processes on time scales larger than about 10 ms can easily be followed (210). In the imaging of samples using transmission electron microscopy (TEM), special staining, embedding and cutting techniques are indispensable, whereas the use of scanning electron microscopy (SEM) is much more straightforward (226). An interesting development is the progress in so called environmental scanning electron microscopy (ESEM), which allows the analysis of samples at a desired relative humidity and thus avoids artifacts due to the dehydration of foodstuffs (210). Different methods for image acquisition (light microscopy, transmission electron microscopy and scanning electron microscopy) are generally coupled to digital analysis to quantitatively define, with structural parameters, food at different structural levels. This provides a measurement of different aggregation descriptors. The gel network can be characterized by structural parameters such as pore size, strand dimensions and how these are distributed in the volume. In the case of particulate gels, the diameter size of the pore is large, up to hundreds of microns, compared to the size of the particle, around microns (227). At low magnifications LM is used to estimate the size of the large pores. At higher magnifications TEM estimates the size of the particles forming the strands of networks. The pore size is more easily measured by digital image analysis than by evaluating the difference in aggregation of particles in the network. In SEM the fracture plane is visualized, and the fracture will follow the weakest structure, i.e., large pores. Thus, SEM micrographs tend to show larger pores. and smaller pores could be embedded in clusters or conglomerates. Stereology is a tool for measuring complex biopolymer gels, where no assumptions of the shape can be made. A stereological approach was used to classify the mode of aggregation by a group of experienced microscopists evaluating SEM-micrographs, to quantify pore size, particle size and amount of threads within the pores in volume weighted mean volumes (227). Five structural descriptors were quantified, namely porosity (number of pores), clusters (many particles attached to each other like bunches of grapes), conglomerates (as if the particles were joined together in non-linear, irregular, inhomogeneous

Scale Length	Methods	Physical State/ Structural Elements	Information
>1 cm	-Texture Analyser -Sensory Panel	Liquid, Gel, Solid, Porous Solid	-properties of network at large deformation -size and shape of macrostructural elements -sensorial attributes
1 mm- 1 cm	-Texture Analyzer -Microscopy	-Liquid-aqueous matrix -Liquid emulsion-matrix -Gels -Porous matrix -Viscoelastic Matrix	-microstructure
1-500 $\mu\text{m}$	-Confocal Microscopy -Light Microscopy -Rheology	micelles, droplets, air cells, crystals, fibres, granules	-size and shape of structures -properties of network at small deformation -ingredient interaction
10-500 nm	-Light Scattering -Electron Microscopy	micelles, droplets, air cells, crystals, fibres, granules	-aggregation, density arrangement -size of structures
<10 nm	-Raman -Chromatography -Thermal Analysis -SDS Page -NIR	carbohydrates, proteins, lipids, water etc...	-molecular structure -proportion of elementary parts -unfolding vs. native -denaturation/transition temperature

**Table 5.1:** Principal methods for structural analyses at characteristic length scales in foods, appearance of food matrix and structural elements

order), strings of beads (as if the particles were attached to each other in a linear order forming strings of beads) and hairiness (as if small threads were attached to the surface of the particles and their outline is indistinct). The three-dimensional gel network is responsible for bulk properties such as diffusion and rheological properties, sensory quality and liquid holding capacity (227).

### 5.1.3 Properties affected by food structure: sensory, stability, digestibility and bioaccessibility

The dimensions/size and shape/form of the particles, strands and pores create the different textural properties of the food products and expert panellists can detect differences between very small particles  $< 1\mu\text{m}^3$  in volume (228). In fact, texture is a multi-parameter attribute, that derives from the molecular, microscopic or macroscopic structure of a food and is detected by several senses, the most important ones being the senses of touch and pressure (229). Food structure, food texture, nutrients digestibility and consumer product preferences and choices are intrinsically linked 5.1. Texture influences people's acceptance of food and may be more important than the flavour in some products (230). The sensory perception during food consumption depends not only on the concentrations of odour- and taste-active compounds but also on the texture of food matrix (231). Multivariate techniques are used to create models to describe groups of the sensory descriptors by some of the microstructural parameters (232). Correlations between the microstructure and sensory descriptors have been found: grainy appearance, gritty texture, creamy texture and tendency to fall apart have a logarithmic dependence on the particle size, and size of small and large pores (228). The soft and springy textures are influenced by combinations of microstructural parameters, where the formation of strands into strings of beads or in clusters and conglomerates seems to play an important role. Conversely, the sticky texture is negatively correlated to the proportion of threads within the pores (228). Stability can be fully grasped only if food molecular dynamics and structure are taken into consideration, i.e., an appropriate understanding of the behaviour of food products requires knowledge of its composition, structure and molecular dynamics, through the three-dimensional arrangement of the various structural elements and their interactions (232). In addition to water, other structural elements can be identified in foods at a supramolecular structure level, such as oil droplets, gas cells, fat crystals, strands, granules, micelles and interfaces. These structural elements, composed of proteins, carbohydrates and lipids (in various combinations and proportions), can exist in different states (glassy/rubbery/crystalline/liquid and solubilised) even at uniform temperatures and water activity. This structural heterogeneity will necessarily affect the molecular dynamics in the system and consequently the macroscopic food quality attributes and their behaviour along storage. Physically separating the reactants in microstructural locations can control the biochemical activity by avoiding the reactants to be in contact. It is a matter of fact that the gastrointestinal fate of lipids depends on their level, type, and structural organization in foods (232). Matrices could be formed by controlled gelation of single or mixed biopolymer systems around lipid droplets, by dehydration of oil-in-water emulsions containing biopolymers or other wall materials, or by thermal treatment or extrusion of starch matrices containing lipid droplets. Several studies have recently investigated the impact of the food matrix on the digestibility of lipids using either *in vitro* or *in vivo* digestion models (233). When oil droplets are dispersed in a solid-like food matrix (e.g., cheese

or strained-type yogurt), the structure of the surrounding food matrix becomes the dominant factor controlling digestion. For instance, the size of lipid droplets dispersed of oil-in-water emulsions and nano emulsions can affect, during digestion, oil-soluble vitamins (vitamins A, D, E and K) bioavailability in fortified foods (234); increasing oil droplet size reduces the bioaccessibility by inhibiting lipid digestion and reducing micelle solubilisation(235). The knowledge advances provided by these studies are setting the foundation for modulating fat digestion through food structure design, as exhaustively reviewed by (209). In this sense, food structure design can be a tool to develop foods that enable to control the body district as well as the extent and rate of release of food lipids along the digestion process. During digestion, the 3D network structure within a food matrix can obstruct the diffusion of enzymes towards the surface of dispersed oil droplets. That is the reason why bile salts are produced by the intestinal tract and released during food digestion to create an emulsion where the digestive enzymes can act onto the food lipids. Compared to interfacial films, the solid like-food matrix is potentially capable of providing enhanced protection against lipolysis (209). Evidence is increasing that a structured food with a high protein content may show slower lipid digestion (236). An investigation on near forty food types, based on the harmonized INFOGEST digestion method (136), found that those with medium and low lipid content showed a limited lipolysis extent when the content of protein or starch was high (237). In protein-rich foods such as cheese, the disintegration of the protein network occurs mainly in gastric and intestinal steps, thus facilitating the subsequent release of fat aggregates from the degraded matrix (238). These results underline the importance of microstructure and the digestive environment on the release of cheese components. The *in vitro* digestion rate of lipids and starch was also reduced due to the intact vegetal cell walls (239). The intact cell wall structure and protein matrices are impervious to amylase and can prevent or slow down enzyme diffusion to substrate. In general, the intactness of cell walls is related to particle size, which is dependent on mastication habits and processing conditions, for example, milling and heating (240). The hydrolysis of intracellular starch and protein in the essentially intact cells was 2–3%, whereas this increased to 40–45%, when the cells were mechanically broken and digested, suggesting a barrier effect of intact cell walls to digestive enzyme access to starch and proteins substrate (241). In support to this hypothesis, it has been shown that solubilisation of pectin cell wall, induced by thermal treatment of bean, exerted higher degrees of cell wall permeability so that starch hydrolysis increased proportionally to the cell damage (242). The morphology and the particle size of starch granules from different plants is also considered an important factor affecting their digestion, as smaller granules have greater enzymatic susceptibility regardless of botanical origin, due to their larger specific surface area (243). Moreover starch granules vary in the level of porosity and can have openings (pores) on the surface of the granule (244). During processing, starch granules swell and lose their crystallinity and molecular organization in a process commonly known as gelatinization. *In vitro* studies have

demonstrated that the rate of enzyme breakdown of gelatinized starch is much higher than that of native starch; native wheat starches are degraded by only 10–15%, but after partial gelatinization the rate of enzymatic degradation increased three-fold (245). Therefore, gelatinization may strongly influence the rate at which starch is digested and elicits the glycaemic response. Starch–protein interaction in white flours might account for a decrease in *in vivo* glycaemic response as well as for a reduction in *in vitro* digestibility, so that the removal of gluten from wheat flour induces a high GI value in 11 kinds of gluten-free bread. In addition to acting as an enzyme barrier, proteins also affect the properties of starch (gelatinization, retrogradation, etc.) which is then less digestible (246). If proteins are present in a structured matrix or a clot-like structure is formed in the gastric environment, gastric juice needs to penetrate this structured matrix to digest the protein. A 2–10 reduction factor for the diffusion coefficient of pepsin has been measured in a structured matrix as compared to water. The diffusion of pepsin is one of the limiting factors in the digestion rate of a structured food matrix (247). Different egg-white gel structures, with a similar protein composition, induced different proteolysis kinetics and provoked the release of different specific peptides (248). Proteins can form supramolecular assemblies also because of thermal treatment. The formation of aggregates may hide peptide bonds from proteases compared to denatured but isolated molecules. The effect of cooking on the digestibility of meat proteins is a good example of such complex relationships. Meat digestibility of regular-cooked beef was higher (95% digested) than that of ‘well-done’ cooked beef (90% digested). Meat analogues are a class of food products that imitate the sensory attributes of meat products but are produced from protein from more sustainable sources, e.g., plant protein isolates, that are subjected to extrusion or shear-cell technology. In these products, the presence of other food ingredients or components, such as lipids and polyphenols, may affect protein digestibility. These effects are still poorly understood for the lack of knowledge of the matrices and by the absence of predictive models. Therefore, in the design of novel foods the effects of components on protein digestibility should be carefully considered in the optimization of the processing parameters (247). The process-induced modifications, in primis the Maillard reaction, could also play a role in modulating the food digestibility and the bioavailability of protein amino acids, by altering the chemical structural of protein networks and in turn the food microstructure: this is the case of bread, dairy and meat products. Not secondarily, these modifications can also affect the food allergenicity, through the interactions of protein-bound advanced glycation end-products (AGE) with immune cells receptors, as evidenced for egg, dairy and peanut allergens (249; 250; 251).

#### **5.1.4 Structure and functional food design**

The main objective of the food industry is to produce products with specific properties and characteristics which have a positive consumer impact. In recent years, the food industry, aware of resource scarcity, is looking for nutritional alternatives, including functional

foods, that promote optimal health and help reduce the risk of disease and “tailored”. Tailoring is a process whereby the provision of information, advice and support is individualized to the user (252). Mimic foods to be substituted, include also new functional ingredients in formulation. The attempt to design new foods starting from more sustainable or more nutrient-rich ingredients, with optimal characteristics for target population groups with specific needs, has always clashed with the need to make these new foods at least as palatable, if not preferable, to traditional ones. The limit is often in the obtainment of a desirable structure. In fact, unlike some homogeneous foods, such as drinks, extracts or oils, most foods are heterogeneous multiphase mixtures, having nutritional and sensory characteristics that strongly depend on the placement with which the different phases are distributed in space, while forming the food matrix. For this reason, the food technologists make use of structure-targeted toolboxes to mimic successful matrices or invent new ones with even more performing characteristics. This is usually carried out empirically in lab scale plants but, to avoid prolonged and expensive physical research trials, the structure of the food could be preliminarily built in-silico also in the design phase. This effective approach could be realized using conceptual toolboxes (simulating unit operations, order of sequential steps, formulations) assisted by mathematical prediction models. The purpose of designing the most suitable structures is then fulfilled, through combinations of formulations and processes, to achieve the desired outcomes, like the optimized durability, palatability, bioaccessibility and bioavailability of nutrients. This way, food design considers not only composition, but also structure affecting chemical stability, texture and dynamics of digestion and absorption of a food or its components. In this perspective, tailored foods provide not only the necessary nutrients but also new functions, linked to the matrix structure, targeted for specific populations groups such as the elderly, babies, athletes, allergic peoples, vegans or for special diets such as low salt, sugars and fats, or lactose- and gluten-free, and to increase the quantity of proteins, vitamins, dietary fibres, and bioactive phytochemicals. Mimic foods to be substituted, include also new functional ingredients in formulation. The attempt to design new foods starting from more sustainable or more nutrient-rich ingredients, with optimal characteristics for target population groups with specific needs, has always clashed with the need to make these new foods at least as palatable, if not preferable, to traditional ones. The limit is often in the obtainment of a desirable structure. In fact, unlike some homogeneous foods, such as drinks, extracts or oils, most foods are heterogeneous multiphase mixtures, having nutritional and sensory characteristics that strongly depend on the placement with which the different phases are distributed in space, while forming the food matrix. For this reason, the food technologists make use of structure-targeted toolboxes to mimic successful matrices or invent new ones with even more performing characteristics. This is usually carried out empirically in lab scale plants but, to avoid prolonged and expensive physical research trials, the structure of the food could be preliminarily built in-silico also in the design phase. This effective approach could be realized using conceptual toolboxes (simulating unit operations, order



of sequential steps, formulations) assisted by mathematical prediction models. The purpose of designing the most suitable structures is then fulfilled, through combinations of formulations and processes, to achieve the desired outcomes, like the optimized durability, palatability, bioaccessibility and bioavailability of nutrients. This way, food design considers not only composition, but also structure affecting chemical stability, texture and dynamics of digestion and absorption of a food or its components. In this perspective, tailored foods provide not only the necessary nutrients but also new functions, linked to the matrix structure, targeted for specific populations groups such as the elderly, babies, athletes, allergic peoples, vegans or for special diets such as low salt, sugars and fats, or lactose- and gluten-free, and to increase the quantity of proteins, vitamins, dietary fibres, and bioactive phytochemicals. Designer-made supramolecular food materials may form the basis for personalized, health-promoting diets of the future (253). As already described in the previous section 5.1, foods are made by colloids toolboxes provided by nature, to which food technologists have added ‘artificial’ colloids, e.g., gas bubbles, oil droplets, ice crystals, fat crystals, and protein aggregates, created by external forces (e.g., extrusion, compression, electric fields) or heating applied by food processing equipment (215). With these ‘artificial’ colloids, foods adhere to the length scales dictated by our tasting senses, which are sensitive enough to detect structures of millimetre down to micrometre size (215). In this sense, a palatable food must be designed by finely modulating these structures to enhance their nutritional function as well. The structure of all foods can be imagined as the result of combinations of structural elements provided by nature or imparted during processing and preparation. Food structure design is the dedicated conception and fabrication of foods in such a way as to attain specific structures, functions or properties (209). Knowledge on how foods and beverages interact with the digestive system, where they transform into supramolecular structures, can in fact have a direct impact on the rational design of such advanced materials for functional food delivery applications. For example, delivering a complete diet with a content of hydrophobic, amphiphilic, and hydrophilic nutrients, which is personalized to the needs of the consumers, could be beneficial for clinical and infant nutrition (236). Otherwise, as confirmed by recent studies on the use in pasta formulation of alternative flour from different sources, such as potato and pigeon pea flour (254) or flours from legumes such as chickpea (255) or bean (256), pasta nutritional profile is usually improved, leading to an increase in protein, ash, fibre contents, and antioxidant compounds together with a decrease in the starch content and of in vitro starch digestibility. What is missing in these approaches, solely accounting for the nutritional profile, based on the composition of the ingredients, is the input related to the target structural characteristics at different scale lengths. Although structure has been shown to have an equally important impact on nutritional quality, a novel food is designed with great care for its composition, stability and acceptability but, often, its structural optimization for nutrient accessibility is omitted in the preliminary conceptualisation phase and studied only ex post. Ultimately, the food structure design



has the potential to be personalized to digestive conditions and dietary nutrient requirements of the consumer or patient. From a nutritional perspective, the ability to control food digestion is extremely important to design food with desired characteristics: the key to control such process is to modulate the accessibility of digestive enzymes to their substrate. Recently, considerable interest has also arisen in the application of by-products of food processing with specific properties in food structure design, such as agar or locust bean gum substitutes.

### 5.1.5 Predictive models and structure design: how do we feed AI?

As described in the previous section, stability, palatability, bioaccessibility and bioavailability of nutrients are the target properties of food optimization. These properties must be expressed using numerical descriptors, such as concentrations of degradation biomarkers, food sensory scores, preferably assessed by instrumental devices (electronic nose or tongue), post-prandial nutrients level in blood. Chemical and instrumental sensory analyses provide objective parameters intrinsic to the food, that are independent from the individual interaction with it. Conversely, parameters related to the digestive functions are strongly linked to the subjects' variability. For this reason, experiments simulating different individual physiological and pathological conditions are necessary, even when characterizing the target properties of a single food. Whereas *in vivo* experiments give a global indication of food nutrients digestibility in its full biological context, and *in vitro* experiments provide more insight into the different chemical and physical mechanisms, the mathematical, or *in silico* modelling can connect these two domains (247). The hydrolysis kinetics of the main macronutrients (proteins, starch, and lipids) are modelled to predict the concentration and their degree of hydrolysis in one or more compartments of the digestive system, or to predict the transport of the food through the digestive system. The most popular approaches assume the digestive tract as a series of bioreactors that can be described by mass balances, written as a set of differential equations (257). In recent years, models that also consider the food matrix together with the reaction and diffusion phenomena have been developed. Modelling of the swelling of protein gels by using the Flory-Rehner theory has been combined with the Gibbs-Donnan theory to include the distribution of ions between the gastric juice and the protein matrix, to gain a better understanding of the phenomena that are essential in the digestion of the food matrix (258). Up to now, the role of modelling has been that of linking and explaining *in vivo* and *in vitro* experiments. However, a further step is required to use modelling for food properties prediction as a function of food structure. Suitable numerical descriptors of structure are required as inputs for AI systems, to predict properties that can define food in a functional way. In the next section, available emerging approaches and those foreseen for the next future are described, emphasizing how structure descriptors have been employed to predict sensory properties and stability toward chemical transformations.

### **Describing the structure with Imaging**

The most straightforward way one can think of to parametrize food structure is through descriptors extracted from imaging. Given the number of existing imaging techniques (microscopy, spectral and hyperspectral imaging, nuclear magnetic resonance imaging, ultrasound, microwave, etc.), many different aspects of food structure can be characterized and digitalized. Furthermore, each imaging technique has its own array of analytics and descriptors, capable of grasping and describing physical quantities tied to the physical nature of the specific imaging technique. All these heterogeneous descriptors, together with general texture analysis and computer vision descriptors, that can be obtained from images under certain conditions, constitutes interesting inputs for artificial intelligence (machine and deep learning) frameworks. As a matter of fact, the role of artificial intelligence in describing food structure from images, is that of finding complex relationships between heterogeneous features describing different aspects of the structure and the different structure-dependent properties of a food. Furthermore, researcher in the field of deep learning, will rightfully argue that in the next future, a general characterization of structure directly from images without a-priori features and descriptors knowledge or assumptions could be possible. From an operative point of view, this means feeding a neural network, as complex as needed, each pixel (or voxel in 3D) of an image as an input and let the network learn how to build the best features to describe the problem (in this case, predict food properties from structure description). To reach this goal, huge quantities of suitable training data are however required to avoid some known problems of deep learning architectures, such as overparameterization and overfitting. While some imaging techniques are inherently suitable for the high-throughput standardized data production (such as magnetic resonance imaging) required by deep learning architectures to achieve good prediction and generalization, other imaging techniques (such as electronic microscopy) suffer from a series of issues that make them less suitable for automation and high-throughput data production. Overall, we are quite far from the data production required to have a huge amount of labelled training data, especially regarding certain imaging techniques. In the next section, a high-throughput imaging technique (MRI) and a high-resolution imaging technique (electronic microscopy) are compared in terms of descriptors and suitability for automation. This is done to outline possible directions to facilitate an efficient use of artificial intelligence at this stage of structure description.

### **On the suitability of data production and imaging parameters for AI: a comparison**

To grasp the meaning of what has been said in the previous section about data production and generality of descriptors, it may be useful to focus on a comparison between electronic microscopy (high-resolution, non-high-throughput) and magnetic resonance imaging (low resolution, high-throughput). Table 5.2 sums up the main categories of descriptors that can be extracted from images coming from these two different techniques,

followed by a synopsis highlighting the upsides and downsides of each technique as far as automation and generalization are concerned. While MRI has many upsides when it comes to data production, generalization, automation of analysis and feature extraction for classification, a trade off exists in terms of spatial resolution. On the other hand, advocating the importance of high-resolution aspects in terms of food structure description implies the necessity of high-resolution imaging techniques. Electronic microscopy can fill in the role provided it becomes suitable for high-throughput data production and data-driven modelling. At present, microscopic image production is not optimized for automatic extraction of general features and descriptions, which are at the core of frameworks using integrated data and automated workflows based on machine learning. The first issue comes from image acquisitions inherently suffering from parameter dependency. Lighting conditions and magnification which are obviously related to experimental purposes, tend to shift microscopic imaging production toward less generalized datasets. Moreover, most canonical morphological and structural descriptors that are quantified from this type of imaging, while being directly related to physical and easily interpreted quantities, require specific assumptions (i.e., presence/absence of pores, spheres, shapes, fibers etc.). Characterizing portions of images with ad-hoc assumptions is ill-suited for automation and generalized parameter extraction. On the other hand, the power in terms of spatial resolution of electronic microscopy cannot be overlooked when trying to characterize food structure. The solution may lie in shifting microscopy data production toward a more pipeline-oriented way. The creation of a consensus for data harmonization of microscopic images in the field, could lead to parameter and feature extraction based upon low level and more general operators, analogous to the ones used for MR images. This shift of paradigm in data production and descriptor extraction, may contribute to boost modelling by facilitating the linking of the many levels of complexity characterizing real life foods, using general parameters. A shift in data production is also needed to pave the way for efficient deep learning approaches.

### Structure images and sensory quality

Some scientific research, considered as an original reference works for these aspects, have laid the foundations for the way a set of fundamental or derived parameters  $X$ , defining the food structure, can be linked to a functional property  $Y$  through a mathematical function (259). For instance, the microstructural parameters may be presented as the estimated model parameters  $A$  and  $B$  necessary to solve a correlating equation, e.g.,  $Y = A + B \ln(X)$ , where  $Y$  is a sensory vector descriptor,  $X$  the model matrix for microstructural parameter. The exemplary work by Langton et al. (259), carried out on whey protein gels, defined nine quantified microstructural parameters constituting the  $X$  vector feeding the model: four parameters were the output of the digital image analysis (i.e., pore size at x20 magnitude; pore size at x40 magnitude; particle size; amount of threads), and five parameters were mode of aggregation as perceived by the test panel and already explained at the

	Sem	MRI
Descriptors	<ul style="list-style-type: none"> <li>-Particle size and morphology</li> <li>-Pore size and morphology</li> <li>-Size distribution and morphology</li> <li>-Shape orientation and diameters distributions</li> </ul>	<ul style="list-style-type: none"> <li>-First order gray level statistics</li> <li>-Roughness of textures</li> <li>-Degree of linearity</li> <li>-Co-occurrence matrix statistics</li> <li>-Structural or morphological features of ROIs</li> <li>-Transform features</li> </ul>
Pros & Cons	<ul style="list-style-type: none"> <li>-Not immediately suitable for high throughput production</li> <li>-No data armonization standard</li> <li>-Wide application fields</li> <li>-Canonical descriptors immediately linked to physical quantities</li> <li>-Requires specific assumptions for image analysis</li> <li>-Very high resolution</li> </ul>	<ul style="list-style-type: none"> <li>-Inherently suitable for high-throughput data production</li> <li>-Data harmonizations standards are widely supported in many biomedical fields</li> <li>-Descriptors come from low-level, general texture analysis and morphological studies alike</li> <li>-Low resolution</li> <li>-Does not requires specific assumptions for image analysis</li> </ul>

**Table 5.2:** Main descriptors and (dis)advantages for electronic microscopy and magnetic resonance imaging

end of section 3 (Porosity; Clusters; Conglomerates; String of beads; Hairiness). Principal component analysis (PCA) of the textural sensory data identifies two groups: (i) grainy appearance, gritty, creamy and falling apart; and (ii) soft, springy, surface moisture and sticky. To find trends in groups of variables (microstructural and sensory variables), PCA on the whole data set was performed. The PCA had the purpose of creating, for each orthogonal component, linear combinations of variables characterized by a high degree of co-variance, thus evidencing their interdependence, by collecting them in different groups. One group of variables, defined by the large and small star volume of pores, the star volume of particles, porosity, clusters, gritty, falling apart and creamy (and acid) was found to take part in the systematic variation. Two groups of microstructural parameters and sensory descriptors were found: one group depending on the dimensions of the overall network and the other depending on the shape of the strands and filling of the pores. This kind of data analysis made model building a realizable approach.

### Structure Images, water dynamics and chemical transformations

Food systems behaviour is strongly dependent on water. Besides water content in a food material, it is important to understand the water state and dynamics for a proper comprehension of properties and stability of food structure. Understanding changes in location and mobility of water represents a significant step in food stability knowledge, since water “availability” within the matrix profoundly influences the chemical, physical and microbiological quality of foods. Water mobility/dynamics can be described as a manifestation how “freely” water molecules can participate in reactions or how easily water molecules diffuse to participate in reactions occurring in different sites (260). Nuclear magnetic resonance is a powerful technique to investigate water dynamics and physical structures of foods, through analysis of nuclear magnetisation relaxation times, because it provides information on molecular dynamics of different components in dense complex systems. The application of this technique may be very useful in predicting food systems physicochemical changes, namely texture, viscosity or water migration (260). Finding correlations amongst parameters based on time domain (TD)-NMR T2 decays, describing water dynamics, and texture-derived features based on SEM images is a challenging issue, when the aim is the quantitative characterization and parametrization of porous food matrices and the transformation that food undergoes due to processing (such as cooking). A comprehensive pipeline for parameter extraction, describing the porous food at different cooking time, must be set accurately. TD-NMR raw data are preferable to classical exponential fitting parameters, for building a general model accounting for the water status, as different phenomena participate in the modulation of the relaxation times of the water population in the compartmentalized porous matrix. For this reason, when matrix effects are investigated with TD-NMR, a probabilistic PCA with Radial Base Function (RBF) kernel may constitute the solution to find a latent space explaining differences in data tied to different matrices (pasta type) and cooking times. The RBF kernel can take the non-linearity of decays into account, projecting data into a suitable latent space, as shown in section 5.2. The next section outlines the necessity to take another level of complexity into account when trying to predict bioavailability and bioaccessibility: the physiological interaction with the human organism.

#### 5.1.6 Structure and in-silico simulators

Gathering an almost infinite set of model foods covering each possible category is a difficult goal to achieve. For this reason, with an available exemplary set of model foods, the next step could be the creation of in-silico models, derived from the mathematical combination of basic models, to simulate each existing real food. As previously stressed, in silico simulations of food as complex particle based soft matter, are strictly bound to the various length scales in the structure and occurring phenomena. As such, different properties must be simultaneously investigated at different scales, from mesoscale to

nanoscale. While mesoscale properties (i.e. for emulsions and fat droplets) can be investigated using coarse-grained particle-based simulations (261), at finer length scales quantum-mechanical effects might occur. While hybrid multiscale models, capable of joining coarse and fine level descriptions, are already available (262), making predictive multiscale simulation approaches seemingly viable, the true complexity of food as a system is still unaddressed. A complete review of available simulation tools, with a breakdown of all the levels of complexity that must be addressed while trying to predict food properties and functionalities from its structure and molecular-level interactions, is provided by Barroso da Silva, et al. (1). Amongst other issues, a predictive model relying solely upon multiscale simulation, can suffer from high computational complexity. Simulating systems consisting of extremely high number of particles, for which free-energy properties and kinetic properties must be computed for several time-steps, can easily lead to unrealistic computational time, even for specialized high-end hardware. However, machine and deep learning can prove useful in decoupling multiscale descriptions from approaches based exclusively on simulation. Quantitative structure-activity relationship (QSAR) based approaches are, in example, very useful in predicting bio-chemical properties of compounds, including biological activity (263). These approaches are based on linking sets of molecular descriptors to a given response variable; essentially the goal is to find a solution to a supervised learning problem by coming up with an optimal set of user-defined molecular descriptors and a suitable model to link them to the outcomes (response variable). A recent development of such a framework involves the use of deep learning architectures, using recurrent and convolutional neural networks (264). The use of such neural networks allows for a generalization of the learning problem, by eliminating the necessity of an a priori definition of the molecular descriptors, at the cost of a very high pool of training data. Approaches of these types, when the interpretability of the network-extracted descriptors is ensured, can minimize the bias introduced by the users when choosing the descriptors and the difficulty of interpreting descriptors that are not directly related to chemical structures. Results from framework of these types, can furthermore be linked with outcomes from physiological experiments (i.e., experiments involving digestibility or involving health effects of certain compounds). In this way, the molecular scale and the macroscale of physiological effects are encased in a multiscale data-driven description. In a similar and more general fashion, many levels and scale of complexity can be linked through machine and deep learning, by finding ways of extracting general descriptors to be related to a response variable. Given the sheer complexity of food, data-driven description of the various levels of complexity of food structure and food-human interactions seems to be a promising way of predicting properties and health effects. In the next section, an example of how to extract joint general descriptors from different scales of complexity (water-matrix interaction and morphology) of a real-life food, that can be ultimately related to outcomes from physiological experiments, is presented. The example, set up by the authors, shows how to use SEM images in a more general



way, by extracting texture analysis descriptors, when the acquisition experimental design is suitable. An example of how to correlate such structure descriptors to properties such as water mobility, using raw data and machine learning, is also proposed.

## 5.2 Case Study: Modelling with Texture Analysis and Raw Data

In this section, we explore the capabilities of a framework based on general image-derived descriptors of pasta structure and their correlation with descriptors of properties affected by structure. Pasta is a good case study for its interesting structural characteristics; it has a compact and dense microstructure, which: i) limits water absorption and thus starch swelling and gelatinization, during cooking; ii) entraps the starch granules reducing the accessibility of  $\alpha$ -amylase and (iii) releases  $\alpha$ -amylase inhibitors during cooking that can immobilize the enzyme into the gluten network. Microstructural changes of starch and proteins during cooking depend on water availability, and the kinetics of solvation of each biopolymer have a major role on the final texture of cooked pasta (265). Thus, pasta is a promising real-life food case study to link structural descriptions with water mobility properties, how they are affected by chemical transformation (cooking) and eventually how structural changes and water availability can impact digestibility.

### 5.2.1 Cooking and water-matrix interaction

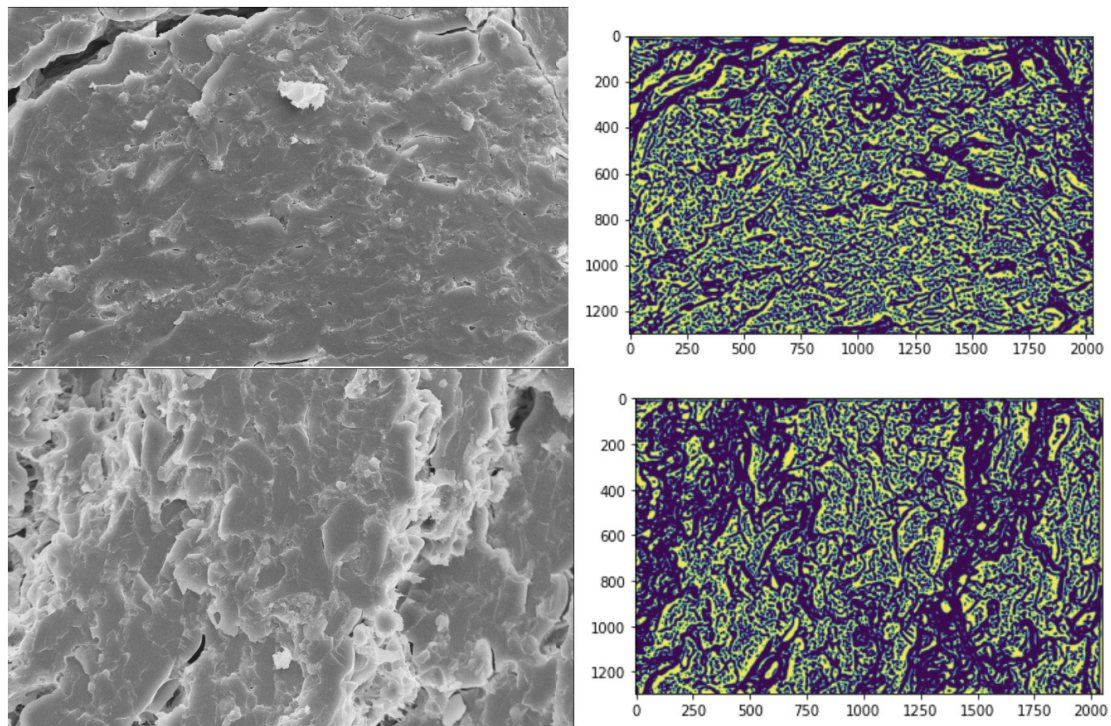
To date, the structure of cooked pasta has been analysed at various microscopic and mesoscopic levels by means of different methods, such as MRI. In fact it can be used to evaluate water distribution and mobility in dry pasta, and in pasta at various cooking time (266). Even these studies revealed that water penetration, distribution, and mobility during cooking were highly dependent on the degree of protein reticulation, which in turns is greatly affected by process conditions and food formulation (267) MRI represents a non-invasive method that spatially resolves the amount and dynamics of water and macromolecules-protons. For this reason, (266) used MRI to make a real time assessment of the effect of starch-gluten ratio on water distribution in dry spaghetti during cooking. Therefore, investigating such properties can help to understand how pasta components (water, gluten, starch, fibre, etc.) interact with each other defining its structure, quality, acceptability, and stability. In this respect, Gallo, et al. (268) investigated the impact of pasta composition (semolina and durum whole-wheat semolina) on water mobility in spaghetti before and after cooking by low-resolution  $^1\text{H}$  NMR experiments. In detail T1 and T2 proton relaxation times as indicators of the molecular water mobility, have been determined (269). The uncooked spaghetti had T1 and T2 values much lower than the cooked ones suggesting a very low water mobility in the dry pasta. With increasing cooking time, it was observed a significant increase of both T1 and T2 relaxation times, either for semolina or whole wheat spaghetti, suggesting that molecular water mobility within the pasta structure increases as protein coagulation and starch gelatinization proceed. According to



Bosmans et al. (270), this behaviour could be explained in term of three phenomena: i) water uptake in pasta structure; ii) starch gelatinization with the subsequent destruction of the original structure; iii) gluten polymerization accompanied by water expulsion from the gluten network. By comparing the behaviour of the two samples, one observes that the presence of fibre led to a reduction in water mobility, since they can keep a substantial excess of water during the cooking process (271). The intermediate zone was characterized by swollen starch granules embedded in a coagulated but dense protein network; the presence of fibre resulted in an irregular structure in which there were a small number of still intact and therefore non-gelatinized starch granules. As reported by Manthey and Schorno (272), in whole-wheat pasta bran particles cause a dilution of the gluten proteins, interfering with proper gluten development. This results in a highly porous structure in which starch granules are more accessible to water molecules. Starch granules in the surface region were fully gelatinized and thus completely disintegrated in amylose and amylopectin. In the intermediate zone, starch granules were highly hydrated increasing in size. Concerning the analysis of surface roughness, laser microscopy stressed an irregular surface structure for dry pasta (due to the presence of intact starch granules) which became more homogeneous after 1 min of cooking, due to the starch gelatinization.

### **5.2.2 Toward the automatization of water-matrix interactions and structure characterization**

Joining measurement of NMR T1 and T2 proton relaxation time with SEM images, seems a promising way of intertwining water mobility related phenomena with morphological variations, thus including structure into food characterization. Parameters extracted with these techniques, can furthermore be modelled using machine and deep learning architectures. However, both methodologies require a fair amount of expertise in acquisition and processing of the data, making standardization and automation of modelling pipelines challenging. Extracting parameters and quantities from SEM images, is especially challenging as it requires the use of dedicated software (e.g., when measuring particle size) to extract the distributions of nanostructures and microstructures in an image. Accurate particle size distributions can be difficult to obtain, as they require images with highly detectable particles and morphologies to build a suitable statistic. Furthermore, the observable size of particles and structures depends drastically on the viewing angle, while measures such as porosity and surface roughness are affected by lighting and zooming. A complete list of issues and standardization of measures for SEM image analysis is provided by the ISO (International Organization for Standardization) (<https://www.iso.org/obp/ui/fr/iso:std:iso:19749:ed-1:v1:en>). On the other hand, NMR relaxometry, while being a high-throughput technique with relatively low acquisition times and high reproducibility, requires expertise in sample preparation and acquisition sequence engineering. Furthermore, studying T1 and T2 distributions with inversion software such as the UPEN algorithm (273) requires a deep understanding of the physical

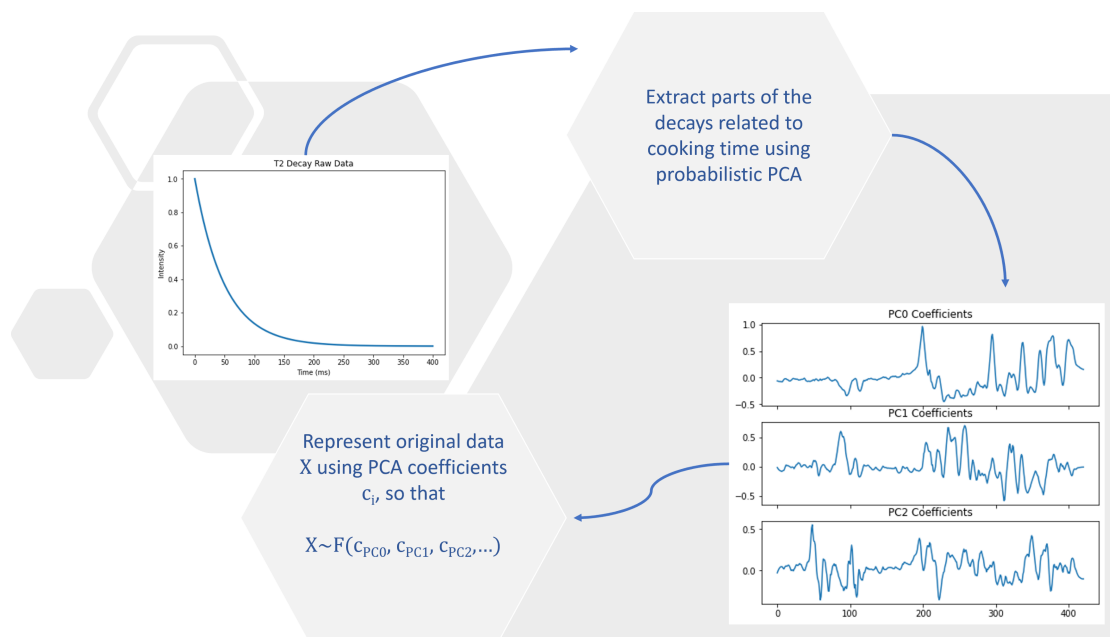


**Figure 5.2:** SEM image processing example of different cooking times (*top: 1 min, bottom: 10 mins*) for the same zone of the same pasta type.

and mathematical nature of the inversion problem, making this kind of analyses extremely variable and elaboration parameters dependent.

### 5.2.3 Is learning from raw data and general descriptors promising?

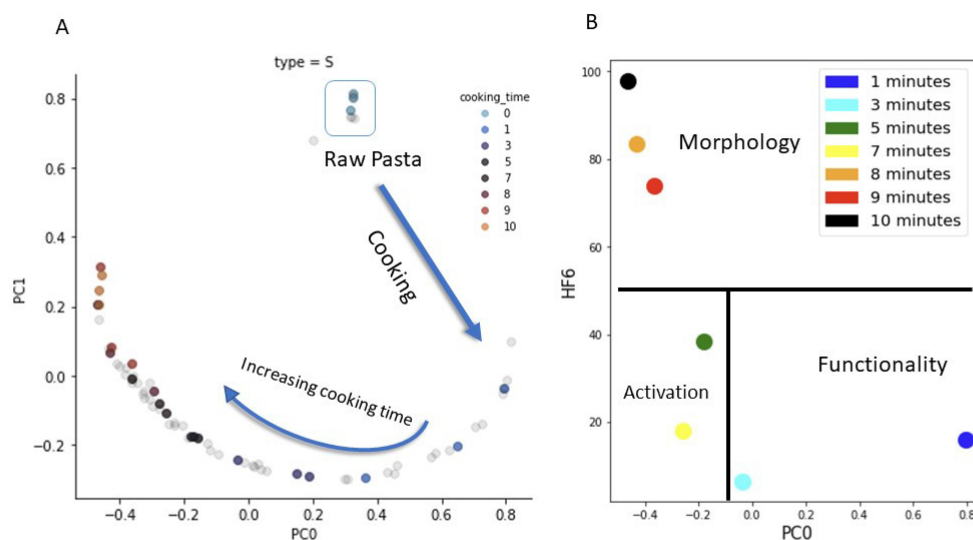
A possible way to bypass some of these issues and make automatization and learning easier, moving toward a more general framework, is to analyse raw TD-NMR decays and study SEM images by extracting general texture analysis features and learning latent components in the data, instead of specific measurements and physical quantities. In this qualitative example, a way to correlate water mobility phenomena and morphology related features using machine learning is proposed. SEM images of different zones of semola spaghetti, acquired at different cooking time points, are processed and segmented using various filtering techniques and morphological operators. A set of minimum image acquisition parameter can be chosen (i.e., zoom, lighting, well defined morphological regions of the pasta to acquire), to minimize variability in the final dataset related to possible acquisition biases. SEM images have been processed and segmented using various filtering techniques and morphological operators. A gaussian filter has been applied to smooth contrast related to lighting and zooming, thresholding has been performed using Bernsen local thresholding, in order to emphasize the formation of pores and nuclei



**Figure 5.3:** The process behind the decomposition of T2 decays raw data into a lower dimensional space. Each time point of each decay is interpreted as variable and fed to a probabilistic PCA with an RBF kernel. Data are transformed according to coefficients which are dependent on the kernel parameters, optimized through machine learning. An example of resulting latent space is showed in the following Figure 5.4a, where each T2 decay, measured for each different cooking time, is represented as a point in a two-dimensional space.

during cooking. An erosion morphological operator with a  $4 \times 4$  structuring element has been applied in order to remove artifacts and spurious pixel zones (Figure 5.2).

The 13 Haralick descriptors (274) are computed from the images of the complete cooking profile of the pasta. These general descriptors, widely used in texture analysis and computer vision, are moments computed from the segmented image cooccurrence matrix. These moments are intended to describe the characteristics of the patterns of the textures of the image, in term of the probability of occurrence of grey levels. As such, they serve as general morphological descriptors, whose relationships with descriptors extracted from TD-NMR can be estimated. These descriptors can be studied as a function of time-dependent latent components extracted from TD-NMR raw decays, with a process summarized in Figure 5.3, to find links with water mobility related phenomena. Typical raw decays of pasta at different cooking time points, are shown as projection into a latent variable space using a probabilistic KPCA (Kernel Principal Component Analysis). Using an RBF kernel in a self-optimizing learning pipeline, each decay curve is projected into a lower dimensional space with the aim of detecting differences tied to phenomena occurring during cooking (Figure 5.4a). Some of the Haralick descriptors appear to have strong linear and non-linear correlations with the time dependent latent variables extracted from TD-NMR raw decays (Figure 5.4b). Moreover, correlations seem to be



**Figure 5.4:** **A.** Resulting lower dimensional latent space, computed according to Figure 2. In this space, each T2 decay measured at different cooking times (indicated by the colour gradient) is represented as a point. The points are the projection in the 2-d latent features space, learned by the kernel, of each T2 decay. In this space, differences tied to effects of cooking on water mobility are the most detectable. **B.** Scatter plot of PC0 vs HF6 (Sum Average, computed from SEM images of the central zones). A qualitative interpretation of the relationship between these two variables can be given as follows: in the functionality phase, water mobility is mainly related with starch gelatinization phenomena, resulting in little morphological changes. After an activation phase, where the rupture of structures in the food matrix begin to arise, the morphological changes detected in images start a strictly monotonous trend related to cooking time (morphology phase).

different from zone to zone, highlighting the expected behaviour of TD-NMR to discriminate information about different characteristics of water populations at different cooking times in different pasta zones.

Some texture analysis descriptors, such as texture Sum Average (HF6, y axis of Figure 5.4b) which is tied to “homogeneity” of the texture, describing the central zone images, show an exclusively monotonous relationship with cooking time and PC scores (both PC0 and PC1) after a certain cooking time (Morphologic Phase, Figure 5.4b). Looking at the KPC space, this phenomenon corresponds to a steep variation in PC1 score and a low variation in PC0 scores. On the contrary, below this time (light blue to dark blue points, up to 5 mins, Figure 5.4a), steep variations along PC0 and slow variations of PC1 scores are encountered, until PC1 score variation minimum is reached (around 5-7 mins, dark points, Figure 5.4a). After this, variation on PC1 scores starts to rise again (Activation, Figure 5.4b) while PC0 scores variation starts to reach its minimum. Above this threshold

of cooking time, both PC1 and many HF descriptors, such as HF6 in Figure 5.4b, start a trend with a strict monotonous dependence with time. This time range may represent the threshold for which changes in the texture of the matrix start to be exclusively dependent on cooking time, maybe due to the irreversible rupture of structures in the food matrix and the consequent variation of the timescale of water exchanges. Looking at Figure 5.4, one can argue that the description of the morphological changes emerging from these preliminary results, is in agreement with findings from Manthey and Schorno (272). If in the early moments of cooking starch gelatinization prevails, the resulting SEM images tend to show more homogeneous surfaces, with little differences from a morphological point of view. However, with raising cooking time the observed increase in the inhomogeneity of pasta surface and the changes in water mobility become a monotonous function of cooking time, as the partial detachment of solid materials such as starch and starch-attached proteins probably becomes the prevalent phenomena. Haralick descriptors for SEM images, together with self-learned latent components extracted from TD-NMR raw decays, are capable of picking up this sort of threshold behaviour and successfully merging description of the morphology and water-matrix interaction. Learning latent features and parameters from raw NMR data and images processed to a bare minimum, studying and understanding the correlation amongst the extracted descriptors can help building digital twins of food with an included structural characterization of the matrix. In the example, water mobility and morphology are investigated with a general data-driven framework, using machine learning and canonical texture analysis to find suitable features and descriptors. The main advantages of this approach are the generality and the lack of assumptions needed for the description of structural elements from images. Using raw data (such as T2 decays in the example) and letting AI methods learn the best way to represent them is optimal when dealing with many heterogeneous datasets, in terms of automation and feature discovery. Moreover, bypassing the necessity of assumptions when describing structure from images, becomes an advantage when parametrizing real-life foods in which matrix structures can be extremely heterogenous along the different length scales. Consequently, different types of images and raw data from experiments regarding digestibility, stability and bioaccessibility can be explored to shed light on their relationship with structural properties, even with complex real-life food.

### 5.3 Chapter Conclusions

Understanding how formulations of ingredients and unitary operations of food processes make up the structure of food and how this structure changes during its shelf - life or eating will play an important role in the development and management of food science and industry. Much of the information that defines the structure of a food is currently neglected when entering the domain of nutrition, as the structural dimension is too complicated to be quantitatively measured and related to sensorial properties, stability, digestibility and



bioaccessibility of nutrients. Not even the momentum given by the considerable progress achieved in the design of functional foods has so far been sufficient to assign the correct importance to the structural nature of food. Certainly, the complexity of the information is such as to hinder the creation of predictive-based models based on analysis of a limited amount of available data. For this reason, it is certainly conceivable a considerable impulse determined using artificial intelligence capable of handling certain quantities of heterogeneous data. It would be useful to be able to predict the sensory quality and stability of food designed to become carriers of healthy nutrients through images that show their supramolecular structure. It would be also desirable for these same foods designed *in silico*, to predict the duration as a function of the dynamic state of the water capable of modulating the chemical transformations underlying physiological or anomalous phenomena, also to include the aspect of sustainability in the conception phase. A model food such as pasta, widely consumed all over the world, object of studies for possible functionalization as a vehicle for bioactive substances useful for health, can serve as a case study to build a pipeline of an automated approach. The endpoint of such a pipeline is a direct extraction of information on rheological and sensory properties starting from images of the structure and from raw data of the dynamic state of the water. The main advantages of such a framework are : i) an efficient automatization of parameter extraction useful for building suitable inputs for AI architectures, which require high-throughput data for proper training ii) a more efficient and general way of extracting parameters especially from imaging; using general parameters for image analysis instead of measured technique-dependent parameters or measured quantities that requires ad-hoc assumptions on structures (i.e. presence/absence of pores, fibres etc.), can prove more useful given the high heterogeneity of structural elements at different length scales iii) a more efficient way of linking different levels of complexity of structure description and properties to be predicted, through the use of general parameters and features learned directly from data with machine learning; this step is crucial to avoid oversimplification generated by canonical interpretative models. However, extending this framework to all the aspects of food modelling for properties prediction, poses quite a few challenges. The first one is a required shift of paradigm of imaging data production. Certain techniques (such as SEM) suffer from a lack of a consensus of acquisition standards, hindering data harmonization which is essential for high-throughput input production. Another major challenge is the complexity of modelling and parametrizing properties such as bioaccessibility and bioavailability. These properties not only require a comprehensive parametrization of the structure to be predicted but are also linked to the interaction with digestive functions. The interaction with the human organism, especially with GIT functions, adds a whole new level of complexity that must be addressed. The compartments of the GIT and their functions are interlinked and impacted by food structure, while also being subjected to interindividual variability. Hybrid approaches linking structure at molecular level and physiological outcomes, based on deep learning architectures, are however gaining pop-

ularity (5.1.6) due to their computational performances. The ultimate goal of AI oriented frameworks is to be able to make more limited use of expensive and time-consuming experiments on physically prepared foods, by using digital twins of foods designed in the laboratory. This, in turn, could lead to a more efficient data production for studies of physiological outcomes of functional foods.



## Conclusions

The advent of the omic sciences era has opened the way for a whole new paradigm to study, model and solve complex biochemical and systems. Omic high-throughput data production grants the access to unprecedented levels of information about the human organism, through snapshots of molecular states at various resolutions and perspectives. Within the right frameworks, the human organism can be described as the ensemble of the crosstalk of its molecular states, observed and characterized through various omic techniques. Moreover, the interactions of the organism with the environment (pathological states, diet, exposition to various factors), can be studied as the downstream of changes in molecular patterns, induced as a perturbation of the molecular ground state of an individual. Within this picture, the role of machine learning is that of linking different levels of molecular descriptors, so that complexity-related characteristics of the human organism, such as emergent properties, non-linearity, self-organization, feedback and transition of its functions, can be modeled intertwining nanoscale (enzymes) and macroscale (organs, systems, individuals, social behavior) and everything in between. In this thesis, the leitmotiv of defining the perturbations of a biochemical complex system, such as the human organism, through patterns of its molecular states defined at different levels of resolution, is proposed through the discussion of several frameworks and models developed by the author, with  $^1\text{H}$  NMR metabolomic spectral data at their core. In the first part of this work, we proposed a way to integrate systemic and intracellular metabolomics with genomic to create an enzyme-network level simulation of perturbations of a pathological state (acute myeloid leukemia) through machine learning. This allowed for important etiologic conclusions about the disease, a reliable method for molecular fingerprinting of the pathological condition and the individuation of possible therapeutic targets. A framework to study the dynamics of perturbed molecular states in the metabolic space is also proposed, using breast cancer clinical trials, to emphasize the importance of information added by studying molecular fingerprints not only as a single snapshot frozen in time, but also by their changes over a time span. This frameworks and models lay the foundation to explore the

topic of the second part of this thesis: the links between health and nutrition, which can be considered the most common long-term perturbation to which an individual is exposed. From epidemiological data to molecular characterization of food structure, we discuss all the possible elements that can influence the molecular state of the human organism when exposed to diet, in order to link different levels of complexity and take a step forward toward a truly holistic model. First, we propose a framework to treat epidemiological data on nutrition as a pattern of macronutrients, showing the advantages of considering population stratifications and the multivariate effects of different macronutrients through a self-optimizing penalized regression model. Then, we add the missing levels of complexity to determine the effect of food intake: chemical composition and structure. To study the impact of dietary intervention based on complexes with different composition, we proposed an integrated microbiomic-metabolomic framework to study the effect of a single bioactive compound in modulating the cross talk between microbiota and energetic metabolism. Then, we developed and discussed a model to characterize individual aspects of digestion kinetics of different real life foods, in a free living population. We showed how, within the right experimental design, kinetic phenomena of different compartments of the gastrointestinal tract can be reconstructed from a single observed biofluid. The result is a model based on the Bateman equations system to simulate individual kinetics, using parameters extracted from unsupervised decomposition of spectral data, that contain the convolution of all information about the metabolic state. Eventually, we discussed the importance of also considering food structure and how it can be described in a general way using machine learning applied to imaging and raw data. A case study on the topic, using relaxometry raw data and SEM images is proposed to emphasize the efficiency of a general structure description, that is also suitable for data integration. Overall, the endpoint of this work is to discuss and enhance the possibilities offered by heuristic modelling, in the era of omic data production, in finding links between health and perturbations of the molecular state, by considering the human organism in all its complexity. Frameworks, models and pipelines developed by the author, are provided as tools to embed the entire omic data workflow into the description of biochemical systems. To do this, the tools developed are proposed as an ensemble of: i) ad-hoc data preprocessing (depending on acquisition technique) or raw data reduction, ii) feature integration, selection and engineering for multi-omic crosstalk, iii) automated model selection for robust data analysis, predictive models and classification, iv) ad-hoc numerical methods for modelling and systems integration. The general breakdown of an omic workflow into these points, is a key aspect to move toward a true holistic description of the complexity underlying links between health and nutrition. The sheer amount of data alone, while opening the way for powerful machine learning techniques to be used and considered reliable, is often not enough for true etiologic conclusions. Thus, the methods developed in this work are intended to emphasize important aspects of complexity disentanglement in an omic workflow used for biochemical description: i) knowledge of experimental techniques for

data processing and unbiased integration of different omic data, ii) raw data reduction for the extraction of general parameters when exploring novel associations between different experimental techniques, iii) interpretability of features and parameters used to reduce the dimensionality of a given problem, iv) interpretability of machine learning algorithms used for classification and prediction, v) efficiency and generality of numerical methods used to model kinetics. Throughout the research presented in this thesis, it is shown that an approach within the aforementioned premises, not only provides great flexibility in its applications due to the generality of many of its aspects, but is also capable of helping etiologic findings by linking different description levels and paradigms of a biochemical system.



## Experimental Designs and Methods Additional Info

### A.1 Simonetti, Mengucci et al., 2021, Springer Nature, Materials and Methods

#### NMR spectra acquisition and processing

A 500  $\mu\text{L}$  aliquot of serum sample was placed in a clean microfuge tube containing 130  $\mu\text{L}$  of D<sub>2</sub>O-based phosphate buffer pH 7.4, 70 mM sodium azide (NaN<sub>3</sub>), 20 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS-d<sub>6</sub>) as chemical shift reference standard and 20 mM 2-chloro pyrimidine-5-carboxylic acid (2CLPYR5CA) as reference standard. The mixture was homogenized by vortexing and 590  $\mu\text{L}$  were transferred into 5x 178 mm (7") 5 mm outer diameter NMR tubes (for Bruker Match holder). <sup>1</sup>H-NMR spectra were recorded from serum and urine samples at 298 K with an AVANCE spectrometer (Bruker BioSpin, Fällanden, Switzerland) operating at a proton frequency of 600.13 MHz, equipped with an autosampler with 60 holders. <sup>1</sup>H-NMR spectra were acquired by applying a standard spin echo Carr-Purcell-Meiboom-Gill (CPMG; cpmgpr1d.comp; Bruker BioSpin, Fällanden, Switzerland) pulse sequence with 256 scans (NS), 32768 data points (TD), a spectral width (SW) of 11.9705 ppm, an acquisition time (AQ) of 2.28 s, and a saturation time of 0.3 milliseconds (D20). A relaxation delay (D1) of 4 s was needed to reduce the signals arising from macromolecules. The data were Fourier transformed and phase and baseline corrections were automatically applied (TopSpin 3.0, Bruker BioSpin). Signals were assigned by comparing their chemical shift and multiplicity with Chenomx software data bank 8.1. 324 serum and 378 urine spectra passed quality control procedures (145 and 139 from controls, 179 and 139 from AML, respectively). Uniform bucketing was applied, resulting in 421 spectral features for each subject. Median Control Specter was used as reference for probabilistic quotient normalization (PQN) and processing. During the quality control procedures, duplicate spectra (urine or serum samples from two different collection days) were compared, by taking into account diet,

drugs, physical exercise data. Spectra showing unmatched peaks resulting from potential confounding factors (e.g. drugs) were excluded from downstream analysis. Signals and corresponding metabolites were matched using Chenomx NMR suite 8.1.

### **Sample Mass spectrometry (MS)-based metabolomics and data analysis**

MS-based metabolomics was performed using an ultra-performance liquid chromatography (Waters ACQUITY, Waters, Milford, MA, USA) and a Q-Exactive high resolution/accurate mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution (Metabolon, Morrisville, NC, USA). Raw data were extracted, peak-identified and QC processed. Compounds were identified by comparison with library entries of purified standards or recurrent unknown entities. Peaks were quantified using area-under-the-curve. Metabolite levels were normalized to DNA content.

### **WES-Whole Exome Sequencing**

WES was performed on 100 AML cases, 17 belonging to a published dataset (275) and 83 new cases. Libraries were prepared from matched tumor and germline DNA (saliva or complete remission samples, Nextera Rapid Capture Expanded and TruSeq Rapid Exome kits, Illumina, San Diego, CA, USA) according to manufacturer's protocol, and 75/125-bp paired-end sequences were generated (Illumina NextSeq550/HiSeq2500, Illumina). Sequencing data are available in the European Genome-Phenome Archive (EGAS00001005422).

### **Constraint-based metabolic network analysis**

We translated gene expression alterations into constraints reducing the feasible space of a metabolic network model (adapted from Shlomi et al. (276)). The impact of a set of these constraints on the feasible space of the metabolic network was evaluated by calculating the minimum and maximum reaction rates (flux variability analysis, FVA), and the instantaneous capability of the network to produce/consume a certain metabolite. Details are reported in the Supplementary Methods. Codes used in constraint-based metabolic network analysis are available in <https://github.com/cladelpino/GenePerturbations>.

### **Constraint-based models**

We selected a hematopoietic model derived from Recon2 via the integration of proteomic data for bone marrow hematopoietic cells from the Human Protein Atlas. A genetic perturbation was characterized by a gene and a direction of perturbation (up/down). All reactions involving any downregulated genes in their regulatory rules were set to zero

flux, while for those involving upregulated genes, the minimum flux was set to a non-zero value, which we will informally call "cut depth". This type of specification can lead to situations where no mass balanced flux distribution can satisfy them. In this implementation, when a new perturbation set is specified, its maximal perturbation sets (which have a non-zero maximum cut depth value) are found, along with the minimal sets of perturbations, which are incompatible. The effect, if any, on the metabolic network is then calculated by maximizing or minimizing the reaction rates (flux variability analysis, FVA) for the reactions. For the metabolites, the corresponding mass balance is relaxed and the corresponding row of the stoichiometric matrix is used as the objective vector. This is analogous to adding a sink or source reaction and maximizing their flux.

## Statistics

Associations in contingency tables were performed by the Monte Carlo (B=1000,000) simulated Fisher's exact test. Continuous variables were compared with Mann–Whitney, Kolmogorov–Smirnov, Kruskal–Wallis test, or Welch's t-test. All tests were performed using either python v3.6.5 (packages scipy v1.3.2, statsmodels v0.10.1) or Rv3.6.3. When appropriate, p values were adjusted for multiple comparisons using the Bonferroni or Benjamini–Hochberg method. To investigate the distribution of sera profile according to blast percentage, samples were divided in three classes (bone marrow: 20–49%, 50–74%,  $\geq 75\%$  blasts, peripheral blood:  $< 30\%$ , 30–69%,  $\geq 70\%$  blasts, according to tertiles). In the drug response analysis, the average area-under-the curve values of the two cohorts were compared. NMR peaks, signal integrals (related to metabolite concentration) and intracellular metabolite levels among three groups were compared by Kruskal–Wallis test. For intracellular metabolite levels Welch t-test was also used as post-hoc test. Random Forest analysis was used to estimate the accuracy of individual classification in each group based on metabolomic data. Metabolic pathway analysis was performed using Metaboanalyst (<http://www.metaboanalyst.ca>) with KEGG annotation. A threshold of five standard deviations from the mean of the control population was used for the identification of outliers.

## Metabolic network reconstruction

The Reaction-Reaction network was generated using the Metabolite-Reactions (1581x2274) stoichiometric matrix mapped in the bone marrow-Recon model. Two reactions are linked if a metabolite produced in one is consumed in the other, resulting in a directed network (source:production, target:consumption). Moreover, if a perturbed metabolite is produced in a reaction and consumed in the other, those reactions are marked as perturbed. In this way, a subgraph of altered reaction can easily be extracted and analyzed. Network building and analyses were performed using Python 3.6 NetworkX package, while visualization and graphic processing were obtained using the Cytoscape 3.7.2 framework. Genes



involved in the identified reactions were retrieved from Recon3D and their interconnection was evaluated by protein-protein interaction analysis on STRING (<https://string-db.org>).

## **A.2 Biagi, Mengucci et al. 2020, MDPI, Materials and Methods**

### **DNA Extraction from Caeca, Ileum and Litter Samples**

A DNeasy PowerSoil kit (Qiagen, Hilden, Germany) was used for DNA extraction from caeca contents following the manufacturer instructions. The protocol used for caeca content was applied to ileal contents with the following modifications to increase DNA yield: (i.) whenever possible, 300 mg of ileal content were used for the DNA extraction, instead of the suggested 200–250 mg; (ii.) elution of the DNA from the Qiagen column was carried out in two steps, using 50  $\mu$ L each time and incubating the columns for 15 min at 4 °C before each centrifugation. As for litter samples, since the starting material was drier than the intestinal content, the buffer present in the bead tube was not enough to hydrate the 250 mg of litter; thus, 100  $\mu$ L of sterile physiological solution was added to the samples. The protocol was then carried on as for the caeca samples.

### **16S rRNA Gene PCR Amplification and Sequencing**

All DNA samples (extracted from caecal, ileal and litter samples) were treated using the same amplification and sequencing protocols. The V3–V4 hypervariable region of the 16S rRNA gene was PCR-amplified using 341F and 785R primers with Illumina overhang adapter sequences. Amplicon purification was performed by using AMPure XP Beads magnetic beads (Beckman Coulter, Brea, CA, USA). For the indexed library preparation, the Nextera XT DNA Library Prep Kit (Illumina, San Diego, CA, USA) was used. A further magnetic bead purification step was performed, and libraries were quantified using the Qubit 3.0 fluorimeter (Invitrogen), then pooled at 4 nM. The library pool was denatured with NaOH 0.2 N and diluted to 6 pM. Sequencing was performed on Illumina MiSeq platform using a 2  $\times$  250 bp paired-end protocol, according to the manufacturer's instructions (Illumina). Three Illumina sequencing runs were necessary in order to sequence all samples with the appropriate sequencing depth. Care was taken in mixing caeca and ileum samples, as well as samples from the different groups (A, B and C) across the different sequencing runs.

### **Bioinformatics and Statistics in Microbiota Analysis**

Raw sequences were processed using a pipeline combining PANDAseq and QIIME 2 (<https://qiime2.org>). High-quality reads were filtered and binned into amplicon sequence variants (ASVs) through an open-reference strategy performed with dada2. The com-

mand "qiime dada2 denoise-single" with QIIME 2 version 2019.10 was used with default parameters, with the exception of length filtering (that is already performed by the PANDAseq pipeline). The method used for chimera seq was "pooled". Taxonomy was assigned using the vsearch classifier and the SILVA database for reference. Alpha diversity was measured using Faith's phylogenetic distance (PD) index, number of observed ASVs and the Shannon diversity index. Statistics was performed using R Studio software version 1.0.136 running on R software 3.1.3 (<https://www.r-project.org/>), implemented with the libraries vegan, made4 and PMCMR. Beta diversity was estimated by computing weighted and unweighted UniFrac distances and was visualized by principal coordinates analyses (PCoAs). Bacterial phylogenetic groups showing a minimum relative abundance of 0.5% in at least the 1% of the samples (for each type of sample) were kept for further analysis and graphical visualization. Compositional differences among groups of samples were tested using the Kruskal–Wallis test. P values were corrected for multiple comparisons using the Benjamini–Hochberg method. In addition, bioinformatics analyses were repeated using the QIIME1 pipeline and operational taxonomic unit (OTU) clustering was performed using a 97% similarity threshold and the UCLUST algorithm. This re-analysis allowed for the definition of group of sequences (97%-similarity OTUs) at an intermediate level between genera and species, for which the ecological behavior across the three considered ecosystems (caeca, ileum and litter) was explored as follows. Core 97%-similarity OTUs were identified as those detected with a relative abundance > 0.1% in > 90% of samples in at least 1 time point, as previously reported (171). Prevalence of the same 97%-similarity OTUs was calculated for all type of samples, at the 3 time points in the 3 groups of broilers (A, B and C), as the percentage of samples in which a given OTU was detected at a relative abundance > 0.1%. The highest score alignment against NCBI 16S rRNA database was obtained by using the BLAST algorithm (<https://blast.ncbi.nlm.nih.gov/>); identification was limited at the genus level for the majority of the core OTUs, whereas identification at the level of species was considered only when > 99% identity was reached.

### Sample Preparation for NMR Analysis and spectra acquisition

Samples were prepared for NMR analysis by vortex mixing for 5 min stool with 1 mL of deionized water, followed by centrifugation for 10 min at 14,000 rpm at 4 °C. Approximately 540 mL of supernatant was added to 100 µL of a D2O 1.5 M phosphate buffer solution containing 0.1% TSP (3-(trimethylsilyl) propionic acid-d4) and 2 mM NaN<sub>3</sub>, set at pH 7.40. Before analysis, samples were centrifuged for 10 min again and then 590 µL were transferred into an NMR tube. Proton NMR (1H-NMR) spectra were recorded at 298 K with an AVANCE III spectrometer (Bruker, Milan, Italy) operating at a frequency of 600.13 MHz. The hydrogen deuterium oxide (HOD) residual signal was suppressed by presaturation, whereas broad signals from slowly tumbling molecules were removed by including a Carr–Purcell–Meiboom–Gill filter with a free induction decay sequence.

The filter was made by a train of 400 echoes separated by 800  $\mu\text{s}$ , for a total time of 328 ms. Each spectrum was acquired by summing up 256 transients using 32 K data-points over a 7211.54-Hz spectrum (for an acquisition time of 2.27 s). The recycle delay was set to 8 s, keeping into consideration the longitudinal relaxation time of the protons under investigation. Each spectrum was processed with Top Spin 3.0 (Bruker) by using an automatic command `apk0.noe`, which performs in one shot the baseline and phase correction, and by applying a line-broadening factor of 1 Hz. The peaks were assigned by comparing their chemical shift and multiplicity with the literature and by using Chenomx NMR suit 8.1 software.

# Bibliography

- [1] F. L. Barroso da Silva, P. Carloni, D. Cheung, G. Cottone, S. Donnini, E. A. Foegeding, M. Gulzar, J. C. Jacquier, V. Lobaskin, D. MacKernan, Z. Mohammad Hosseini Naveh, R. Radhakrishnan, and E. E. Santiso, “Understanding and controlling food protein structure and function in foods: Perspectives from experiments and computer simulations,” *Annual Review of Food Science and Technology*, vol. 11, no. 1, pp. 365–387, 2020. PMID: 31951485.
- [2] S. Le Feunteun, A. R. Mackie, and D. Dupont, “In silico trials of food digestion and absorption: how far are we?,” *Current Opinion in Food Science*, vol. 31, pp. 121–125, 2020. Food Chemistry and Biochemistry • Food Bioprocessing.
- [3] T. E. Moxon, P. Nimmegeers, D. Telen, P. J. Fryer, J. Van Impe, and S. Bakalis, “Effect of chyme viscosity and nutrient feedback mechanism on gastric emptying,” *Chemical Engineering Science*, vol. 171, pp. 318–330, 2017.
- [4] B. Walther, A. M. Lett, A. Bordoni, L. Tomás-Cobos, J. A. Nieto, D. Dupont, F. Danesi, D. R. Shahar, A. Echaniz, R. Re, A. S. Fernandez, A. Deglaire, D. Gille, A. Schmid, and G. Vergères, “Gutself: Interindividual variability in the processing of dietary compounds by the human gastrointestinal tract,” *Molecular Nutrition & Food Research*, vol. 63, no. 21, p. 1900677, 2019.
- [5] Y. Bar-Yam, “General features of complex systems,” *Encyclopedia of Life Support Systems (EOLSS)*, UNESCO, EOLSS Publishers, Oxford, UK, vol. 1, 2002.
- [6] D. R. Hartree, “The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods,” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 89–110, Cambridge university press, 1928.
- [7] G. S. Omenn, S. J. Nass, C. M. Micheel, *et al.*, “Evolution of translational omics: lessons learned and the path forward,” 2012.
- [8] J. C. Lindon, J. K. Nicholson, E. Holmes, and J. R. Everett, “Metabonomics: metabolic processes studied by nmr spectroscopy of biofluids,” *Concepts in Magnetic Resonance: An Educational Journal*, vol. 12, no. 5, pp. 289–320, 2000.
- [9] D. B. Kell, “Metabolomics and systems biology: making sense of the soup,” *Current opinion in microbiology*, vol. 7, no. 3, pp. 296–307, 2004.
- [10] B. H. ter Kuile and H. V. Westerhoff, “Transcriptome meets metabolome: hierar-

- chical and metabolic regulation of the glycolytic pathway,” *FEBS letters*, vol. 500, no. 3, pp. 169–171, 2001.
- [11] H. Pearson, “Meet the human metabolome,” *Nature*, vol. 446, no. 7131, p. 8, 2007.
- [12] B. Worley and R. Powers, “Multivariate analysis in metabolomics,” *Current Metabolomics*, vol. 1, no. 1, pp. 92–107, 2013.
- [13] S. K. Bharti and R. Roy, “Quantitative 1h nmr spectroscopy,” *TrAC Trends in Analytical Chemistry*, vol. 35, pp. 5–26, 2012.
- [14] A.-H. Emwas, R. Roy, R. T. McKay, L. Tenori, E. Saccenti, G. Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, *et al.*, “Nmr spectroscopy for metabolomics research,” *Metabolites*, vol. 9, no. 7, p. 123, 2019.
- [15] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, *et al.*, “Hmdb 4.0: the human metabolome database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D608–D617, 2018.
- [16] D. L. Bandalos, *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.
- [17] D. Ruiz-Perez, H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan, “So you think you can pls-da?,” *BMC bioinformatics*, vol. 21, no. 1, pp. 1–10, 2020.
- [18] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.
- [19] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.
- [21] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [22] D. Müllner, “fastcluster: Fast hierarchical, agglomerative clustering routines for r and python,” *Journal of Statistical Software*, vol. 53, no. 1, pp. 1–18, 2013.
- [23] G. Simonetti, C. Mengucci, A. Padella, E. Fonzi, G. Picone, C. Delpino, J. Nanni, R. De Tommaso, E. Franchini, C. Papayannidis, *et al.*, “Integrated genomic-metabolic classification of acute myeloid leukemia defines a subgroup with npm1 and cohesin/dna damage mutations,” *Leukemia*, vol. 35, no. 10, pp. 2813–2826, 2021.
- [24] S. Gudmundsson and I. Thiele, “Computationally efficient flux variability analysis,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–3, 2010.
- [25] J. W. Tyner, C. E. Tognon, D. Bottomly, B. Wilmot, S. E. Kurtz, S. L. Savage, N. Long, A. R. Schultz, E. Traer, M. Abel, *et al.*, “Functional genomic landscape of acute myeloid leukaemia,” *Nature*, vol. 562, no. 7728, pp. 526–531, 2018.

- [26] C. G. A. R. Network, “Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia,” *New England Journal of Medicine*, vol. 368, no. 22, pp. 2059–2074, 2013.
- [27] E. Papaemmanuil, M. Gerstung, L. Bullinger, V. I. Gaidzik, P. Paschka, N. D. Roberts, N. E. Potter, M. Heuser, F. Thol, N. Bolli, *et al.*, “Genomic classification and prognosis in acute myeloid leukemia,” *New England Journal of Medicine*, vol. 374, no. 23, pp. 2209–2221, 2016.
- [28] I. Tzoulaki, R. Castagne, C. L. Boulange, I. Karaman, E. Chekmeneva, E. Evangelou, T. M. Ebbels, M. R. Kaluarachchi, M. Chadeau-Hyam, D. Mosen, *et al.*, “Serum metabolic signatures of coronary and carotid atherosclerosis and subsequent cardiovascular disease,” *European heart journal*, vol. 40, no. 34, pp. 2883–2896, 2019.
- [29] G. Barbara, E. Scaiola, M. R. Barbaro, E. Biagi, L. Laghi, C. Cremon, G. Marasco, A. Colecchia, G. Picone, N. Salfi, *et al.*, “Gut microbiota, metabolome and immune signatures in patients with uncomplicated diverticular disease,” *Gut*, vol. 66, no. 7, pp. 1252–1261, 2017.
- [30] J. Liu, S. Semiz, S. J. van der Lee, A. van der Spek, A. Verhoeven, J. B. van Klinken, E. Sijbrands, A. C. Harms, T. Hankemeier, K. W. van Dijk, *et al.*, “Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study,” *Metabolomics*, vol. 13, no. 9, pp. 1–11, 2017.
- [31] A. Hasim, M. Ali, B. Mamtimin, J.-Q. Ma, Q.-Z. Li, and A. Abudula, “Metabonomic signature analysis of cervical carcinoma and precancerous lesions in women by 1h nmr spectroscopy,” *Experimental and therapeutic medicine*, vol. 3, no. 6, pp. 945–951, 2012.
- [32] L. Puchades-Carrasco and A. Pineda-Lucena, “Metabolomics applications in precision medicine: an oncological perspective,” *Current topics in medicinal chemistry*, vol. 17, no. 24, pp. 2740–2751, 2017.
- [33] P. S. Ward, J. Patel, D. R. Wise, O. Abdel-Wahab, B. D. Bennett, H. A. Collier, J. R. Cross, V. R. Fantin, C. V. Hedvat, A. E. Perl, *et al.*, “The common feature of leukemia-associated *idh1* and *idh2* mutations is a neomorphic enzyme activity converting  $\alpha$ -ketoglutarate to 2-hydroxyglutarate,” *Cancer cell*, vol. 17, no. 3, pp. 225–234, 2010.
- [34] A. T. Fathi, H. Sadrzadeh, D. R. Borger, K. K. Ballen, P. C. Amrein, E. C. Attar, J. Foster, M. Burke, H. U. Lopez, C. R. Matulis, *et al.*, “Prospective serial evaluation of 2-hydroxyglutarate, during treatment of newly diagnosed acute myeloid leukemia, to assess disease activity and therapeutic response,” *Blood, The Journal of the American Society of Hematology*, vol. 120, no. 23, pp. 4649–4652, 2012.
- [35] M. E. Figueroa, O. Abdel-Wahab, C. Lu, P. S. Ward, J. Patel, A. Shih, Y. Li, N. Bhagwat, A. Vasanthakumar, H. F. Fernandez, *et al.*, “Leukemic *idh1* and *idh2*

- mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation,” *Cancer cell*, vol. 18, no. 6, pp. 553–567, 2010.
- [36] A. Chaturvedi, M. A. Cruz, N. Jyotsana, A. Sharma, R. Goparaju, A. Schwarzer, K. Görlich, R. Schottmann, E. A. Struys, E. E. Jansen, *et al.*, “Enantiomer-specific and paracrine leukemogenicity of mutant idh metabolite 2-hydroxyglutarate,” *Leukemia*, vol. 30, no. 8, pp. 1708–1715, 2016.
- [37] C. D. DiNardo, K. J. Propert, A. W. Loren, E. Paietta, Z. Sun, R. L. Levine, K. S. Straley, K. Yen, J. P. Patel, S. Agresta, *et al.*, “Serum 2-hydroxyglutarate levels predict isocitrate dehydrogenase mutations and clinical outcome in acute myeloid leukemia,” *Blood, The Journal of the American Society of Hematology*, vol. 121, no. 24, pp. 4917–4924, 2013.
- [38] H. Ye, B. Adane, N. Khan, E. Alexeev, N. Nusbacher, M. Minhajuddin, B. M. Stevens, A. C. Winters, X. Lin, J. M. Ashton, *et al.*, “Subversion of systemic glucose metabolism as a mechanism to support the growth of leukemia cells,” *Cancer Cell*, vol. 34, no. 4, pp. 659–673, 2018.
- [39] M. Škrtić, S. Sriskanthadevan, B. Jhas, M. Gebbia, X. Wang, Z. Wang, R. Hurren, Y. Jitkova, M. Gronda, N. Maclean, *et al.*, “Inhibition of mitochondrial translation as a therapeutic strategy for human acute myeloid leukemia,” *Cancer cell*, vol. 20, no. 5, pp. 674–688, 2011.
- [40] E. D. Lagadinou, A. Sach, K. Callahan, R. M. Rossi, S. J. Neering, M. Minhajuddin, J. M. Ashton, S. Pei, V. Grose, K. M. O’Dwyer, *et al.*, “Bcl-2 inhibition targets oxidative phosphorylation and selectively eradicates quiescent human leukemia stem cells,” *Cell stem cell*, vol. 12, no. 3, pp. 329–341, 2013.
- [41] C. L. Jones, B. M. Stevens, A. D’Alessandro, J. A. Reisz, R. Culp-Hill, T. Nemkov, S. Pei, N. Khan, B. Adane, H. Ye, *et al.*, “Inhibition of amino acid metabolism selectively targets human leukemia stem cells,” *Cancer cell*, vol. 34, no. 5, pp. 724–740, 2018.
- [42] C. L. Jones, B. M. Stevens, A. D’Alessandro, R. Culp-Hill, J. A. Reisz, S. Pei, A. Gustafson, N. Khan, J. DeGregori, D. A. Pollyea, *et al.*, “Cysteine depletion targets leukemia stem cells through inhibition of electron transport complex ii,” *Blood, The Journal of the American Society of Hematology*, vol. 134, no. 4, pp. 389–394, 2019.
- [43] N. Jacque, A. M. Ronchetti, C. Larrue, G. Meunier, R. Birsén, L. Willems, E. Salland, J. Decroocq, T. T. Maciel, M. Lambert, *et al.*, “Targeting glutaminolysis has antileukemic activity in acute myeloid leukemia and synergizes with bcl-2 inhibition,” *Blood, The Journal of the American Society of Hematology*, vol. 126, no. 11, pp. 1346–1356, 2015.
- [44] J. R. Molina, Y. Sun, M. Protopopova, S. Gera, M. Bandi, C. Bristow, T. McAfoos, P. Morlacchi, J. Ackroyd, A.-N. A. Agip, *et al.*, “An inhibitor of oxidative phospho-



- rylation exploits cancer vulnerability,” *Nature medicine*, vol. 24, no. 7, pp. 1036–1046, 2018.
- [45] F. Mussai, S. Egan, J. Higginbotham-Jones, T. Perry, A. Beggs, E. Odintsova, J. Loke, G. Pratt, K. P. U, A. Lo, *et al.*, “Arginine dependence of acute myeloid leukemia blast proliferation: a novel therapeutic target,” *Blood, The Journal of the American Society of Hematology*, vol. 125, no. 15, pp. 2386–2396, 2015.
- [46] P. Gallipoli, G. Giotopoulos, K. Tzelepis, A. S. Costa, S. Vohra, P. Medina-Perez, F. Basheer, L. Marando, L. Di Lisio, J. M. Dias, *et al.*, “Glutaminolysis is a metabolic dependency in flt3itd acute myeloid leukemia unmasked by flt3 tyrosine kinase inhibition,” *Blood, The Journal of the American Society of Hematology*, vol. 131, no. 15, pp. 1639–1653, 2018.
- [47] N. Fenouille, C. F. Bassil, I. Ben-Sahra, L. Benajiba, G. Alexe, A. Ramos, Y. Pikman, A. S. Conway, M. R. Burgess, Q. Li, *et al.*, “The creatine kinase pathway is a metabolic vulnerability in evi1-positive acute myeloid leukemia,” *Nature medicine*, vol. 23, no. 3, pp. 301–313, 2017.
- [48] A. Barve, A. Vega, P. P. Shah, S. Ghare, L. Casson, M. Wunderlich, L. J. Siskind, and L. J. Beverly, “Perturbation of methionine/s-adenosylmethionine metabolism as a novel vulnerability in mll rearranged leukemia,” *Cells*, vol. 8, no. 11, p. 1322, 2019.
- [49] H. Ju, G. Zhan, A. Huang, Y. Sun, S. Wen, J. Yang, W. Lu, R. Xu, J. Li, Y. Li, *et al.*, “Itid mutation in flt3 tyrosine kinase promotes warburg effect and renders therapeutic sensitivity to glycolytic inhibition,” *Leukemia*, vol. 31, no. 10, pp. 2143–2150, 2017.
- [50] Y. Itahana and K. Itahana, “Emerging roles of p53 family members in glucose metabolism,” *International journal of molecular sciences*, vol. 19, no. 3, p. 776, 2018.
- [51] R. Boidot, F. Végran, A. Meulle, A. Le Breton, C. Dessy, P. Sonveaux, S. Lizard-Nacol, and O. Feron, “Regulation of monocarboxylate transporter mct1 expression by p53 mediates inward and outward lactate fluxes in tumors,” *Cancer research*, vol. 72, no. 4, pp. 939–948, 2012.
- [52] K. Sumita, Y.-H. Lo, K. Takeuchi, M. Senda, S. Kofuji, Y. Ikeda, J. Terakawa, M. Sasaki, H. Yoshino, N. Majd, *et al.*, “The lipid kinase pi5p4k $\beta$  is an intracellular gtp sensor for metabolism and tumorigenesis,” *Molecular cell*, vol. 61, no. 2, pp. 187–198, 2016.
- [53] B. Zand, R. A. Previs, N. M. Zacharias, R. Rupaimoole, T. Mitamura, A. S. Nagaraja, M. Guindani, H. J. Dalton, L. Yang, J. Baddour, *et al.*, “Role of increased n-acetylaspartate levels in cancer,” *JNCI: Journal of the National Cancer Institute*, vol. 108, no. 6, 2016.
- [54] S. G. Musharraf, A. J. Siddiqui, T. Shamsi, and A. Naz, “Serum metabolomics

- of acute lymphoblastic leukaemia and acute myeloid leukaemia for probing biomarker molecules,” *Hematological oncology*, vol. 35, no. 4, pp. 769–777, 2017.
- [55] I. S. Grønningsæter, H. K. Fredly, B. T. Gjertsen, K. J. Hatfield, and Ø. Bruserud, “Systemic metabolomic profiling of acute myeloid leukemia patients before and during disease-stabilizing treatment based on all-trans retinoic acid, valproic acid, and low-dose chemotherapy,” *Cells*, vol. 8, no. 10, p. 1229, 2019.
- [56] W.-L. Chen, J.-H. Wang, A.-H. Zhao, X. Xu, Y.-H. Wang, T.-L. Chen, J.-M. Li, J.-Q. Mi, Y.-M. Zhu, Y.-F. Liu, *et al.*, “A distinct glucose metabolism signature of acute myeloid leukemia with prognostic value,” *Blood, The Journal of the American Society of Hematology*, vol. 124, no. 10, pp. 1645–1654, 2014.
- [57] Y. Wang, L. Zhang, W.-L. Chen, J.-H. Wang, N. Li, J.-M. Li, J.-Q. Mi, W.-N. Zhang, Y. Li, S.-F. Wu, *et al.*, “Rapid diagnosis and prognosis of de novo acute myeloid leukemia by serum metabolomic analysis,” *Journal of proteome research*, vol. 12, no. 10, pp. 4393–4401, 2013.
- [58] W. Wojtowicz, A. Chachaj, A. Olczak, A. Ząbek, E. Piątkowska, J. Rybka, A. Butrym, M. Biedroń, G. Mazur, T. Wróbel, *et al.*, “Serum nmr metabolomics to differentiate haematologic malignancies,” *Oncotarget*, vol. 9, no. 36, p. 24414, 2018.
- [59] H. Bhanot, M. M. Reddy, A. Nonami, E. L. Weisberg, D. Bonal, P. T. Kirschmeier, S. Salgia, K. Podar, I. Galinsky, T. K. Chowdary, *et al.*, “Pathological glycogenesis through glycogen synthase 1 and suppression of excessive amp kinase activity in myeloid leukemia cells,” *Leukemia*, vol. 29, no. 7, pp. 1555–1563, 2015.
- [60] B. Stockard, T. Garrett, J. Guingab-Cagmat, S. Meshinchi, and J. Lamba, “Distinct metabolic features differentiating flt3-itd aml from flt3-wt childhood acute myeloid leukemia,” *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.
- [61] D. A. Pollyea, B. M. Stevens, C. L. Jones, A. Winters, S. Pei, M. Minhajuddin, A. D’Alessandro, R. Culp-Hill, K. A. Riemondy, A. E. Gillen, *et al.*, “Venetoclax with azacitidine disrupts energy metabolism and targets leukemia stem cells in patients with acute myeloid leukemia,” *Nature medicine*, vol. 24, no. 12, pp. 1859–1866, 2018.
- [62] P. Puchalska and P. A. Crawford, “Multi-dimensional roles of ketone bodies in fuel metabolism, signaling, and therapeutics,” *Cell metabolism*, vol. 25, no. 2, pp. 262–284, 2017.
- [63] R. A. Casero, T. M. Stewart, and A. E. Pegg, “Polyamine metabolism and cancer: treatments, challenges and opportunities,” *Nature Reviews Cancer*, vol. 18, no. 11, pp. 681–695, 2018.
- [64] E. F. Mason, F. C. Kuo, R. P. Hasserjian, A. C. Seegmiller, and O. Pozdnyakova, “A distinct immunophenotype identifies a subset of npml1-mutated aml with tet2 or idh1/2 mutations and improved outcome,” *American journal of hematology*,

- vol. 93, no. 4, pp. 504–510, 2018.
- [65] S. Cuartero, F. D. Weiss, G. Dharmalingam, Y. Guo, E. Ing-Simmons, S. Masella, I. Robles-Rebollo, X. Xiao, Y.-F. Wang, I. Barozzi, *et al.*, “Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation,” *Nature immunology*, vol. 19, no. 9, pp. 932–941, 2018.
- [66] K. Stegmaier, S. M. Corsello, K. N. Ross, J. S. Wong, D. J. DeAngelo, and T. R. Golub, “Gefitinib induces myeloid differentiation of acute myeloid leukemia,” *Blood*, vol. 106, no. 8, pp. 2841–2848, 2005.
- [67] A. Kentsis, C. Reed, K. L. Rice, T. Sanda, S. J. Rodig, E. Tholouli, A. Christie, P. J. Valk, R. Delwel, V. Ngo, *et al.*, “Autocrine activation of the met receptor tyrosine kinase in acute myeloid leukemia,” *Nature medicine*, vol. 18, no. 7, pp. 1118–1122, 2012.
- [68] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, “Soft similarity and soft cosine measure: Similarity of features in vector space model,” *Computación y Sistemas*, vol. 18, no. 3, pp. 491–504, 2014.
- [69] M. J. Kim, W. H. Jung, and J. S. Koo, “Expression of sarcosine metabolism-related proteins in estrogen receptor negative breast cancer according to the androgen receptor and her-2 status,” *International journal of clinical and experimental pathology*, vol. 8, no. 7, p. 7967, 2015.
- [70] H. Cena and P. C. Calder, “Defining a healthy diet: Evidence for the role of contemporary dietary patterns in health and disease,” *Nutrients*, vol. 12, no. 2, 2020.
- [71] D. Katz and S. Meller, “Can we say what diet is best for health?,” *Annual Review of Public Health*, vol. 35, no. 1, pp. 83–103, 2014. PMID: 24641555.
- [72] Y. Y. Lam and E. Ravussin, “Analysis of energy metabolism in humans: A review of methodologies,” *Molecular Metabolism*, vol. 5, no. 11, pp. 1057–1071, 2016.
- [73] A. R. Kristal, U. Peters, and J. D. Potter, “Is it time to abandon the food frequency questionnaire?,” *Cancer Epidemiology and Prevention Biomarkers*, vol. 14, no. 12, pp. 2826–2828, 2005.
- [74] T. Byers, “Food Frequency Dietary Assessment: How Bad Is Good Enough?,” *American Journal of Epidemiology*, vol. 154, pp. 1087–1088, 12 2001.
- [75] F. Danesi, C. Mengucci, S. Vita, A. Bub, S. Seifert, C. Malpuech-Brugère, R. Richard, C. Orfila, S. Sutulic, L. Ricciardiello, E. Marcato, F. Capozzi, and A. Bordoni, “Unveiling the correlation between inadequate energy/macronutrient intake and clinical alterations in volunteers at risk of metabolic syndrome by a predictive model,” *Nutrients*, vol. 13, no. 4, 2021.
- [76] R. L. Hanson, G. Imperatore, P. H. Bennett, and W. C. Knowler, “Components of the “Metabolic Syndrome” and Incidence of Type 2 Diabetes ,” *Diabetes*, vol. 51, pp. 3120–3127, 10 2002.

- [77] Y. Rochlani, N. V. Pothineni, S. Kovelamudi, and J. L. Mehta, "Metabolic syndrome: pathophysiology, management, and modulation by natural compounds," *Therapeutic Advances in Cardiovascular Disease*, vol. 11, no. 8, pp. 215–225, 2017. PMID: 28639538.
- [78] D. Keane, "Diet and metabolic syndrome: An overview," *Current Vascular Pharmacology*, vol. 11, no. 6, pp. 842–857, 2013.
- [79] D. E. Warburton, C. W. Nicol, and S. S. Bredin, "Health benefits of physical activity: the evidence," *CMAJ*, vol. 174, no. 6, pp. 801–809, 2006.
- [80] M. Rodriguez-Monforte, E. Sánchez, F. Barrio, B. Costa, and G. Flores-Mateo, "Metabolic syndrome and dietary patterns: a systematic review and meta-analysis of observational studies," *European journal of nutrition*, vol. 56, no. 3, pp. 925–947, 2017.
- [81] M. Vajdi, M. A. Farhangi, and L. Nikniaz, "Diet-derived nutrient patterns and components of metabolic syndrome: a cross-sectional community-based study," *BMC Endocrine Disorders*, vol. 20, pp. 1–13, 2020.
- [82] E. K. Calton, A. P. James, P. K. Pannu, and M. J. Soares, "Certain dietary patterns are beneficial for the metabolic syndrome: reviewing the evidence," *Nutrition research*, vol. 34, no. 7, pp. 559–568, 2014.
- [83] I. Drake, E. Sonestedt, U. Ericson, P. Wallström, and M. Orho-Melander, "A western dietary pattern is prospectively associated with cardio-metabolic traits and incidence of the metabolic syndrome," *British Journal of Nutrition*, vol. 119, no. 10, pp. 1168–1176, 2018.
- [84] A. J. Ahola, V. Harjutsalo, L. M. Thorn, R. Freese, C. Forsblom, S. Mäkimattila, and P.-H. Groop, "The association between macronutrient intake and the metabolic syndrome and its components in type 1 diabetes," *British Journal of Nutrition*, vol. 117, no. 3, pp. 450–456, 2017.
- [85] M. R. Skilton, M. Laville, A. E. Cust, P. Moulin, and F. Bonnet, "The association between dietary macronutrient intake and the prevalence of the metabolic syndrome," *British journal of nutrition*, vol. 100, no. 2, pp. 400–407, 2008.
- [86] K. G. Alberti, R. H. Eckel, S. M. Grundy, P. Z. Zimmet, J. I. Cleeman, K. A. Donato, J.-C. Fruchart, W. P. T. James, C. M. Loria, and S. C. Smith Jr, "Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; american heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity," *Circulation*, vol. 120, no. 16, pp. 1640–1645, 2009.
- [87] A. Bub, C. Malpuech-Brugère, C. Orfila, J. Amat, A. Arianna, A. Blot, M. Di Nunzio, M. Holmes, Z. Kertész, L. Marshall, *et al.*, "A dietary intervention of bioactive enriched foods aimed at adults at risk of metabolic syndrome: Protocol and results

- from pathway-27 pilot study,” *Nutrients*, vol. 11, no. 8, p. 1814, 2019.
- [88] S. Sutulic, J. Amat, A. Blot, I. Nemeth, Z. Kertész, L. Marshall, S. Seifert, L. Ricciardiello, C. Malpuech-Brugère, A. Bordoni, *et al.*, “Protocol for pilot studies: Effectiveness of bioactive enriched foods (bef) on markers of metabolic syndrome,” *Pathway-27 Project*, 2019.
- [89] S. A. Bingham, A. A. Welch, A. McTaggart, A. A. Mulligan, S. A. Runswick, R. Luben, S. Oakes, K. T. Khaw, N. Wareham, and N. E. Day, “Nutritional methods in the european prospective investigation of cancer in norfolk,” *Public health nutrition*, vol. 4, no. 3, pp. 847–858, 2001.
- [90] Á. Ambrus, Z. Horváth, Z. Farkas, E. Dorogházi, J. Cseh, S. Petrova, P. Dimitrov, V. Duleva, L. Rangelova, E. Chikova-Iscener, *et al.*, “Pilot study in the view of a pan-european dietary survey—adolescents, adults and elderly,” *EFSA supporting publications*, vol. 10, no. 11, p. 508E, 2013.
- [91] G. Goldberg, A. Black, S. Jebb, T. Cole, P. Murgatroyd, W. Coward, and A. Prentice, “Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. derivation of cut-off limits to identify under-recording.,” *European journal of clinical nutrition*, vol. 45, no. 12, pp. 569–581, 1991.
- [92] A. E. Black, “Critical evaluation of energy intake using the goldberg cut-off for energy intake: basal metabolic rate. a practical guide to its calculation, use and limitations,” *International journal of obesity*, vol. 24, no. 9, pp. 1119–1130, 2000.
- [93] A. E. Black, “Critical evaluation of energy intake using the goldberg cut-off for energy intake: basal metabolic rate. a practical guide to its calculation, use and limitations,” *International journal of obesity*, vol. 24, no. 9, pp. 1119–1130, 2000.
- [94] N. EFSA Panel on Dietetic Products and A. (NDA), “Scientific opinion on dietary reference values for energy,” *EFSA Journal*, vol. 11, no. 1, p. 3005, 2013.
- [95] E. F. S. A. (EFSA), “Dietary reference values for nutrients summary report,” tech. rep., Wiley Online Library, 2017.
- [96] W. J. FAO, “Who expert consultation on fats and fatty acids in human nutrition: report of an expert consultation,” *Rome, Italy: Food and Agriculture Organization*, 2010.
- [97] W. H. Organization, *Diet, nutrition, and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation*, vol. 916. World Health Organization, 2003.
- [98] A. You, “Dietary guidelines for americans,” *US Department of Health and Human Services and US Department of Agriculture*, 2015.
- [99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–

- 2830, 2011.
- [100] M. H. Gruber, *Improving efficiency by shrinkage: the James-Stein and ridge regression estimators*. Routledge, 2017.
- [101] F. Zhang, T. M. Tapera, and J. Gou, “Application of a new dietary pattern analysis method in nutritional epidemiology,” *BMC medical research methodology*, vol. 18, no. 1, pp. 1–10, 2018.
- [102] R. Garcia-Carretero, L. Vigil-Medina, O. Barquero-Perez, I. Mora-Jimenez, C. Soguero-Ruiz, R. Goya-Esteban, and J. Ramos-Lopez, “Logistic lasso and elastic net to characterize vitamin d deficiency in a hypertensive obese population,” *Metabolic syndrome and related disorders*, vol. 18, no. 2, pp. 79–85, 2020.
- [103] W. Daniel, “Spearman rank correlation coefficient. applied nonparametric statistics . 358–365. boston, ma: Pws-kent,” 1990.
- [104] A. D. Pradhan, “Sex differences in the metabolic syndrome: implications for cardiovascular health in women,” *Clinical chemistry*, vol. 60, no. 1, pp. 44–52, 2014.
- [105] Y.-M. Yang, B.-C. Shin, C. Son, and I.-H. Ha, “An analysis of the associations between gender and metabolic syndrome components in korean adults: A national cross-sectional study,” *BMC endocrine disorders*, vol. 19, no. 1, pp. 1–10, 2019.
- [106] E. W. Steyerberg *et al.*, *Clinical prediction models*. Springer, 2019.
- [107] D. S. Moore and S. Kirkland, *The basic practice of statistics*, vol. 2. WH Freeman New York, 2007.
- [108] L. Hooper, A. Abdelhamid, D. Bunn, T. Brown, C. D. Summerbell, and C. M. Skeaff, “Effects of total fat intake on body weight,” *Cochrane database of systematic reviews*, no. 8, 2015.
- [109] J. Aranceta and C. Pérez-Rodrigo, “Recommended dietary reference intakes, nutritional goals and dietary guidelines for fat and fatty acids: a systematic review,” *British Journal of Nutrition*, vol. 107, no. S2, pp. S8–S22, 2012.
- [110] A. E. Field, W. C. Willett, L. Lissner, and G. A. Colditz, “Dietary fat and weight gain among women in the nurses’ health study,” *Obesity*, vol. 15, no. 4, pp. 967–976, 2007.
- [111] A. E. Field, W. C. Willett, L. Lissner, and G. A. Colditz, “Dietary fat and weight gain among women in the nurses’ health study,” *Obesity*, vol. 15, no. 4, pp. 967–976, 2007.
- [112] W. S. Yancy, C.-C. Wang, and M. L. Maciejewski, “Trends in energy and macronutrient intakes by weight status over four decades,” *Public health nutrition*, vol. 17, no. 2, pp. 256–265, 2014.
- [113] K. L. Stanhope, “Sugar consumption, metabolic disease and obesity: The state of the controversy,” *Critical reviews in clinical laboratory sciences*, vol. 53, no. 1, pp. 52–67, 2016.

- [114] B. A. Hannon, S. V. Thompson, C. G. Edwards, S. K. Skinner, G. M. Niemi, N. A. Burd, H. D. Holscher, M. Teran-Garcia, and N. A. Khan, "Dietary fiber is independently related to blood triglycerides among adults with overweight and obesity," *Current developments in nutrition*, vol. 3, no. 2, p. nzy094, 2019.
- [115] B. A. Hannon, S. V. Thompson, C. G. Edwards, S. K. Skinner, G. M. Niemi, N. A. Burd, H. D. Holscher, M. Teran-Garcia, and N. A. Khan, "Dietary fiber is independently related to blood triglycerides among adults with overweight and obesity," *Current developments in nutrition*, vol. 3, no. 2, p. nzy094, 2019.
- [116] V. Wijendran and K. Hayes, "Dietary n-6 and n-3 fatty acid balance and cardiovascular health," *Annu. Rev. Nutr.*, vol. 24, pp. 597–615, 2004.
- [117] A. Abdelhamid, T. Brown, J. Brainard, *et al.*, "Omega 3 fatty acids for the primary and secondary prevention of cardiovascular disease. cochrane database of syst rev. 7: Cd003177," 2018.
- [118] C.-Q. Lai, D. Corella, S. Demissie, L. A. Cupples, X. Adiconis, Y. Zhu, L. D. Parnell, K. L. Tucker, and J. M. Ordovas, "Dietary intake of n-6 fatty acids modulates effect of apolipoprotein a5 gene on plasma fasting triglycerides, remnant lipoprotein concentrations, and lipoprotein particle size: the framingham heart study," *Circulation*, vol. 113, no. 17, pp. 2062–2070, 2006.
- [119] F. M. Sacks and M. Katan, "Randomized clinical trials on the effects of dietary fat and carbohydrate on plasma lipoproteins and cardiovascular disease," *The American journal of medicine*, vol. 113, no. 9, pp. 13–24, 2002.
- [120] P. W. Siri-Tarino, Q. Sun, F. B. Hu, and R. M. Krauss, "Saturated fat, carbohydrate, and cardiovascular disease," *The American journal of clinical nutrition*, vol. 91, no. 3, pp. 502–509, 2010.
- [121] K. McAuley and J. Mann, "Thematic review series: patient-oriented research. nutritional determinants of insulin resistance," *Journal of lipid research*, vol. 47, no. 8, pp. 1668–1676, 2006.
- [122] E. J. Feskens, J. G. Loeber, and D. Kromhout, "Diet and physical activity as determinants of hyperinsulinemia: the zutphen elderly study," *American journal of epidemiology*, vol. 140, no. 4, pp. 350–360, 1994.
- [123] S. O. Ebbesson, M. E. Tejero, J. C. López-Alvarenga, W. S. Harris, L. O. Ebbesson, R. B. Devereux, J. W. MacCluer, C. Wenger, S. Laston, R. R. Fabsitz, *et al.*, "Individual saturated fatty acids are associated with different components of insulin resistance and glucose metabolism: the gocadan study," *International journal of circumpolar health*, vol. 69, no. 4, pp. 344–351, 2010.
- [124] R. D. Freire, M. A. Cardoso, S. G. Gimeno, S. R. Ferreira, and J.-B. D. S. Group, "Dietary fat is associated with metabolic syndrome in japanese brazilians," *Diabetes care*, vol. 28, no. 7, pp. 1779–1785, 2005.
- [125] K. S. Reddy and M. B. Katan, "Diet, nutrition and the prevention of hypertension



- and cardiovascular diseases,” *Public health nutrition*, vol. 7, no. 1a, pp. 167–186, 2004.
- [126] A. Grynberg, “Hypertension prevention: from nutrients to (fortified) foods to dietary patterns. focus on fatty acids,” *Journal of human hypertension*, vol. 19, no. 3, pp. S25–S33, 2005.
- [127] P. Pérez-Martínez, D. P. Mikhailidis, V. G. Athyros, M. Bullo, P. Couture, M. I. Covas, L. de Koning, J. Delgado-Lista, A. Díaz-López, C. A. Drevon, *et al.*, “Lifestyle recommendations for the prevention and management of metabolic syndrome: an international panel recommendation,” *Nutrition reviews*, vol. 75, no. 5, pp. 307–326, 2017.
- [128] K. Sun, J. Liu, and G. Ning, “Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies,” 2012.
- [129] A. E. Ivanescu, P. Li, B. George, A. W. Brown, S. W. Keith, D. Raju, and D. B. Allison, “The importance of prediction model validation and assessment in obesity and nutrition research,” *International journal of obesity*, vol. 40, no. 6, pp. 887–894, 2016.
- [130] C. Mengucci, A. Bordoni, and F. Capozzi, “Understanding the kinetics of nutrients bioaccessibility by modelling foodomics data,” *Current Opinion in Food Science*, vol. 31, pp. 114–120, 2020. Food Chemistry and Biochemistry • Food Bioprocessing.
- [131] J. A. Betts and J. T. Gonzalez, “Personalised nutrition: What makes you so special?,” *Nutrition Bulletin*, vol. 41, no. 4, pp. 353–359, 2016.
- [132] J. De Toro-Martín, B. J. Arsenault, J.-P. Després, and M.-C. Vohl, “Precision nutrition: A review of personalized nutritional approaches for the prevention and management of metabolic syndrome,” *Nutrients*, vol. 9, no. 8, 2017.
- [133] L. Laghi, G. Picone, and F. Capozzi, “Nuclear magnetic resonance for foodomics beyond food analysis,” *TrAC Trends in Analytical Chemistry*, vol. 59, pp. 93–102, 2014.
- [134] A. Scalbert, L. Brennan, C. Manach, C. Andres-Lacueva, L. O. Dragsted, J. Draper, S. M. Rappaport, J. J. van der Hoof, and D. S. Wishart, “The food metabolome: a window over dietary exposure,” *The American Journal of Clinical Nutrition*, vol. 99, pp. 1286–1308, 04 2014.
- [135] A. Bordoni and F. Capozzi, “Foodomics for healthy nutrition,” *Current Opinion in Clinical Nutrition and Metabolic Care*, vol. 17, no. 5, pp. 418–424, 2014.
- [136] A. Brodkorb, L. Egger, M. Alminger, and al., “Infogest static in vitro simulation of gastrointestinal food digestion,” *Nature protocols*, vol. 14, pp. 991–1014, 2019.
- [137] D. Dupont, S. Le Feunteun, S. Marze, and I. Souchon, “Structuring food to control its disintegration in the gastrointestinal tract and optimize nutrient bioavailability,”

- Innovative Food Science and Emerging Technologies*, vol. 46, pp. 83–90, 2018.
- [138] E. Brouwer-Brolsma, L. Brennan, C. Drevon, H. Van Kranen, C. Manach, L. Dragsted, H. Roche, C. Andres-Lacueva, S. Bakker, J. Bouwman, F. Capozzi, S. De Saeger, T. Gundersen, M. Kolehmainen, S. Kulling, R. Landberg, J. Linseisen, F. Mattivi, R. Mensink, C. Scaccini, T. Skurk, I. Tetens, G. Vergeres, D. Wishart, A. Scalbert, and E. Feskens, “Combining traditional dietary assessment methods with novel metabolomics techniques: Present efforts by the food biomarker alliance,” *Proceedings of the Nutrition Society*, vol. 76, no. 4, pp. 619–627, 2017.
- [139] A. Tebani and S. Bekri, “Paving the way to precision nutrition through metabolomics,” *Frontiers in Nutrition*, vol. 6, 2019.
- [140] K. Westerman, A. Reaver, C. Roy, M. Ploch, E. Sharoni, B. Nogal, D. Sinclair, D. Katz, J. Blumberg, and G. Blander, “Longitudinal analysis of biomarker data from a personalized nutrition platform in healthy subjects,” *Scientific Reports*, vol. 8, no. 1, 2018.
- [141] M. Hiolle, V. Lechevalier, J. Floury, N. Boulier-Monthéan, C. Prioul, D. Dupont, and F. Nau, “In vitro digestion of complex foods: How microstructure influences food disintegration and micronutrient bioaccessibility,” *Food Research International*, vol. 128, p. 108817, 2020.
- [142] M. El-Bakry and J. Sheehan, “Analysing cheese microstructure: A review of recent developments,” *Journal of Food Engineering*, vol. 125, pp. 84–96, 2014.
- [143] D. Groß, K. Zick, and G. Guthausen, “Chapter four - recent mri and diffusion studies of food structures,” vol. 90 of *Annual Reports on NMR Spectroscopy*, pp. 145–197, Academic Press, 2017.
- [144] R. Deng, A. E. Janssen, F. J. Vergeldt, H. Van As, C. de Graaf, M. Mars, and P. A. Smeets, “Exploring in vitro gastric digestion of whey protein by time-domain nuclear magnetic resonance and magnetic resonance imaging,” *Food Hydrocolloids*, vol. 99, p. 105348, 2020.
- [145] W. Cheng, D.-W. Sun, H. Pu, and Q. Wei, “Heterospectral two-dimensional correlation analysis with near-infrared hyperspectral imaging for monitoring oxidative damage of pork myofibrils during frozen storage,” *Food Chemistry*, vol. 248, pp. 119–127, 2018.
- [146] L. Schoeman, P. Williams, A. du Plessis, and M. Manley, “X-ray micro-computed tomography ( $\mu$  ct) for non-destructive characterisation of food microstructure,” *Trends in Food Science & Technology*, vol. 47, pp. 10–24, 2016.
- [147] E. Capuano, T. Oliviero, and M. A. van Boekel, “Modeling food matrix effects on chemical reactivity: Challenges and perspectives,” *Critical Reviews in Food Science and Nutrition*, vol. 58, no. 16, pp. 2814–2828, 2018. PMID: 28662371.
- [148] T. Grauwet, L. Vervoort, I. Colle, A. Van Loey, and M. Hendrickx, “From finger-

- printing to kinetics in evaluating food quality changes,” *Trends in Biotechnology*, vol. 32, no. 3, pp. 125–131, 2014.
- [149] Z. Zhang, R. Zhang, and D. McClements, “Establishing the impact of food matrix effects on the bioaccessibility of nutraceuticals and pesticides using a standardized food model,” *Food and Function*, vol. 10, no. 3, pp. 1375–1385, 2019.
- [150] I. Garcia-Perez, J. Posma, E. Chambers, J. Nicholson, J. Mathers, M. Beckmann, J. Draper, E. Holmes, and G. Frost, “An analytical pipeline for quantitative characterization of dietary intake: application to assess grape intake,” *Journal of agriculture food chemistry*, vol. 64, pp. 2423–2431, 2016.
- [151] H. Gibbons, C. J. R. Michielsen, M. Rundle, G. Frost, B. A. McNulty, A. P. Nugent, J. Walton, A. Flynn, M. J. Gibney, and L. Brennan, “Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example,” *Molecular Nutrition & Food Research*, vol. 61, no. 10, p. 1700037, 2017.
- [152] L. H. Münger, A. Trimigno, G. Picone, C. Freiburghaus, G. Pimentel, K. J. Burton, F. P. Pralong, N. Vionnet, F. Capozzi, R. Badertscher, and G. Vergères, “Identification of urinary food intake biomarkers for milk, cheese, and soy-based drink by untargeted gc-ms and nmr in healthy humans,” *Journal of Proteome Research*, vol. 16, no. 9, pp. 3321–3335, 2017. PMID: 28753012.
- [153] W. Cheung, P. Keski-Rahkonen, N. Assi, P. Ferrari, H. Freisling, S. Rinaldi, N. Slimani, R. Zamora-Ros, M. Rundle, G. Frost, H. Gibbons, E. Carr, L. Brennan, A. J. Cross, V. Pala, S. Panico, C. Sacerdote, D. Palli, R. Tumino, T. Kühn, R. Kaaks, H. Boeing, A. Floegel, F. Mancini, M.-C. Boutron-Ruault, L. Baglietto, A. Trichopoulou, A. Naska, P. Orfanos, and A. Scalbert, “A metabolomic study of biomarkers of meat and fish intake,” *The American Journal of Clinical Nutrition*, vol. 105, pp. 600–608, 01 2017.
- [154] L. Dragsted, Q. Gao, A. Scalbert, G. Vergères, M. Kolehmainen, C. Manach, L. Brennan, L. Afman, D. Wishart, C. Andres Lacueva, M. Garcia-Aloy, H. Verhagen, E. Feskens, and G. Praticò, “Validation of biomarkers of food intake-critical assessment of candidate biomarkers,” *Genes and Nutrition*, vol. 13, no. 1, 2018.
- [155] I. Garcia-Perez, J. M. Posma, R. Gibson, E. S. Chambers, T. H. Hansen, H. Vestergaard, T. Hansen, M. Beckmann, O. Pedersen, P. Elliott, J. Stamler, J. K. Nicholson, J. Draper, J. C. Mathers, E. Holmes, and G. Frost, “Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial,” *The Lancet Diabetes & Endocrinology*, vol. 5, no. 3, pp. 184–195, 2017.
- [156] E. Björnson, C. J. Packard, M. Adiels, L. Andersson, N. Matikainen, S. Söderlund, J. Kahri, C. Sihlbom, A. Thorsell, H. Zhou, M.-R. Taskinen, and J. Borén, “Investigation of human apob48 metabolism using a new, integrated non-steady-state model of apob48 and apob100 kinetics,” *Journal of Internal Medicine*, vol. 285,

- no. 5, pp. 562–577, 2019.
- [157] V. Sachdeva, A. Roy, and N. Bharadvaja, “Current prospects of nutraceuticals: A review,” *Current pharmaceutical biotechnology*, vol. 21, no. 10, pp. 884–896, 2020.
- [158] A. M. Valdes, J. Walter, E. Segal, and T. D. Spector, “Role of the gut microbiota in nutrition and health,” *Bmj*, vol. 361, 2018.
- [159] E. Biagi, C. Mengucci, M. Barone, G. Picone, A. Lucchi, P. Celi, G. Litta, M. Candela, G. Manfreda, P. Brigidi, *et al.*, “Effects of vitamin b2 supplementation in broilers microbiota and metabolome,” *Microorganisms*, vol. 8, no. 8, p. 1134, 2020.
- [160] D. Borda-Molina, J. Seifert, and A. Camarinha-Silva, “Current perspectives of the chicken gastrointestinal tract and its microbiome,” *Computational and structural biotechnology journal*, vol. 16, pp. 131–139, 2018.
- [161] S. Yadav and R. Jha, “Strategies to modulate the intestinal microbiota and their effects on nutrient utilization, performance, and health of poultry,” *Journal of animal science and biotechnology*, vol. 10, no. 1, pp. 1–11, 2019.
- [162] J. M. Diaz Carrasco, N. A. Casanova, and M. E. Fernández Miyakawa, “Microbiota, gut health and chicken productivity: what is the connection?,” *Microorganisms*, vol. 7, no. 10, p. 374, 2019.
- [163] P. Celi, A. Cowieson, F. Fru-Nji, R. Steinert, A.-M. Klüenter, and V. Verlhac, “Gastrointestinal functionality in animal nutrition and health: new opportunities for sustainable animal production,” *Animal Feed Science and Technology*, vol. 234, pp. 88–100, 2017.
- [164] J. G. Kers, F. C. Velkers, E. A. Fischer, G. D. Hermes, J. A. Stegeman, and H. Smidt, “Host and environmental factors affecting the intestinal microbiota in chickens,” *Frontiers in Microbiology*, vol. 9, p. 235, 2018.
- [165] D. W. Waite and M. Taylor, “Exploring the avian gut microbiota: current trends and future directions,” *Frontiers in microbiology*, vol. 6, p. 673, 2015.
- [166] A. De Cesare, I. F. do Valle, C. Sala, F. Sirri, A. Astolfi, G. Castellani, and G. Manfreda, “Effect of a low protein diet on chicken ceca microbiome and productive performances,” *Poultry science*, vol. 98, no. 9, pp. 3963–3976, 2019.
- [167] B. B. Oakley, H. S. Lillehoj, M. H. Kogut, W. K. Kim, J. J. Maurer, A. Pedroso, M. D. Lee, S. R. Collett, T. J. Johnson, and N. A. Cox, “The chicken gastrointestinal microbiome,” *FEMS microbiology letters*, vol. 360, no. 2, pp. 100–112, 2014.
- [168] V. Eeckhaut, F. Van Immerseel, S. Croubels, S. De Baere, F. Haesebrouck, R. Ducatelle, P. Louis, and P. Vandamme, “Butyrate production in phylogenetically diverse firmicutes isolated from the chicken caecum,” *Microbial biotechnology*, vol. 4, no. 4, pp. 503–512, 2011.

- [169] B. B. Oakley and M. H. Kogut, "Spatial and temporal changes in the broiler chicken cecal and fecal microbiomes and correlations of bacterial taxa with cytokine gene expression," *Frontiers in veterinary science*, vol. 3, p. 11, 2016.
- [170] S. Ranjitkar, B. Lawley, G. Tannock, and R. M. Engberg, "Bacterial succession in the broiler gastrointestinal tract," *Applied and environmental microbiology*, vol. 82, no. 8, pp. 2399–2410, 2016.
- [171] T. J. Johnson, B. P. Youmans, S. Noll, C. Cardona, N. P. Evans, T. P. Karnezos, J. M. Ngunjiri, M. C. Abundo, and C.-W. Lee, "A consistent and predictable commercial broiler chicken bacterial microbiota in antibiotic-free production displays strong correlations with performance," *Applied and environmental microbiology*, vol. 84, no. 12, pp. e00362–18, 2018.
- [172] V. Clavijo and M. J. V. Flórez, "The gastrointestinal microbiome and its association with the control of pathogens in broiler chicken production: a review," *Poultry science*, vol. 97, no. 3, pp. 1006–1021, 2018.
- [173] J. Z. von Martels, A. R. Bourgonje, M. A. Klaassen, H. A. Alkhalifah, M. Sadaghian Sadabad, A. Vich Vila, R. Gacesa, R. Y. Gabriëls, R. E. Steinert, B. H. Jansen, *et al.*, "Riboflavin supplementation in patients with crohn's disease [the rise-up study]," *Journal of Crohn's and Colitis*, vol. 14, no. 5, pp. 595–607, 2020.
- [174] S. Shahzad, M. A. Ashraf, M. Sajid, A. Shahzad, A. Rafique, and M. S. Mahmood, "Evaluation of synergistic antimicrobial effect of vitamins (a, b1, b2, b6, b12, c, d, e and k) with antibiotics against resistant bacterial strains," *Journal of global antimicrobial resistance*, vol. 13, pp. 231–236, 2018.
- [175] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics," *Analytical chemistry*, vol. 78, no. 13, pp. 4281–4290, 2006.
- [176] H. Wold, "Path models with latent variables: The nipals approach," in *Quantitative sociology*, pp. 307–357, Elsevier, 1975.
- [177] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [178] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, and J. Tan, "Asymmetric least squares for multiple spectra baseline correction," *Analytica chimica acta*, vol. 683, no. 1, pp. 63–68, 2010.
- [179] U. Z. Ijaz, L. Sivaloganathan, A. McKenna, A. Richmond, C. Kelly, M. Linton, A. C. Stratakos, U. Lavery, A. Elmi, B. W. Wren, *et al.*, "Comprehensive longitudinal microbiome analysis of the chicken cecum reveals a shift from competitive to environmental drivers and a window of opportunity for campylobacter," *Frontiers*

- in microbiology*, vol. 9, p. 2452, 2018.
- [180] S. Ahn, T.-E. Jin, D.-H. Chang, M.-S. Rhee, H. J. Kim, S. J. Lee, D.-S. Park, and B.-C. Kim, “*Agathobaculum butyriciproducens* gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of *eubacterium desmolans* as *agathobaculum desmolans* comb. nov.,” *International journal of systematic and evolutionary microbiology*, vol. 66, no. 9, pp. 3656–3661, 2016.
- [181] K. N. Haas and J. L. Blanchard, “*Kineothrix alysoides*, gen. nov., sp. nov., a saccharolytic butyrate-producer within the family *lachnospiraceae*,” *International journal of systematic and evolutionary microbiology*, vol. 67, no. 2, pp. 402–410, 2017.
- [182] M. Sakamoto, N. Ikeyama, M. Yuki, and M. Ohkuma, “Draft genome sequence of *faecalimonas umbilica* ta jcm 30896t, an acetate-producing bacterium isolated from human feces,” *Microbiology resource announcements*, vol. 7, no. 9, pp. e01091–18, 2018.
- [183] M. G. Dominguez-Bello and M. J. Blaser, “Do you have a probiotic in your future?,” *Microbes and infection*, vol. 10, no. 9, pp. 1072–1076, 2008.
- [184] F. Van Immerseel, V. Eeckhaut, R. J. Moore, M. Choct, and R. Ducatelle, “Beneficial microbial signals from alternative feed ingredients: a way to improve sustainability of broiler production?,” *Microbial biotechnology*, vol. 10, no. 5, pp. 1008–1011, 2017.
- [185] L. A. Rubio, “Possibilities of early life programming in broiler chickens via intestinal microbiota modulation,” *Poultry science*, vol. 98, no. 2, pp. 695–706, 2019.
- [186] R. E. Steinert, Y.-K. Lee, and W. Sybesma, “Vitamins for the gut microbiome,” *Trends in molecular medicine*, vol. 26, no. 2, pp. 137–140, 2020.
- [187] Y.-h. Luo, H.-w. Peng, A.-D. G. Wright, S.-p. Bai, X.-m. Ding, Q.-f. Zeng, H. Li, P. Zheng, Z.-w. Su, R.-y. Cui, *et al.*, “Broilers fed dietary vitamins harbor higher diversity of cecal bacteria and higher ratio of *clostridium*, *faecalibacterium*, and *lactobacillus* than broilers with no dietary vitamins revealed by 16s rna gene clone libraries,” *Poultry science*, vol. 92, no. 9, pp. 2358–2366, 2013.
- [188] V. A. Torok, R. J. Hughes, L. L. Mikkelsen, R. Perez-Maldonado, K. Balding, R. MacAlpine, N. J. Percy, and K. Ophel-Keller, “Identification and characterization of potential performance-related gut microbiotas in broiler chickens across various feeding trials,” *Applied and environmental microbiology*, vol. 77, no. 17, pp. 5868–5878, 2011.
- [189] D. Stanley, R. J. Hughes, M. S. Geier, and R. J. Moore, “Bacteria within the gastrointestinal tract microbiota correlated with improved growth and feed conversion: challenges presented for the identification of performance enhancing probiotic bacteria,” *Frontiers in microbiology*, vol. 7, p. 187, 2016.
- [190] M. Ventura, C. Canchaya, G. F. Fitzgerald, R. S. Gupta, and D. van Sinderen, “Ge-

- nomics as a means to understand bacterial phylogeny and ecological adaptation: the case of bifidobacteria,” *Antonie Van Leeuwenhoek*, vol. 91, no. 4, pp. 351–372, 2007.
- [191] J. M. Diaz Carrasco, E. A. Redondo, N. D. Pin Viso, L. M. Redondo, M. D. Farber, and M. E. Fernandez Miyakawa, “Tannins and bacitracin differentially modulate gut microbiota of broiler chickens,” *BioMed research international*, vol. 2018, 2018.
- [192] I. Biasato, I. Ferrocino, E. Biasibetti, E. Grego, S. Dabbou, A. Sereno, F. Gai, L. Gasco, A. Schiavone, L. Cocolin, *et al.*, “Modulation of intestinal microbiota, morphology and mucin composition by dietary insect meal inclusion in free-range chickens,” *BMC veterinary research*, vol. 14, no. 1, pp. 1–15, 2018.
- [193] H. Liu, R. Hou, G. Yang, F. Zhao, and W. Dong, “In vitro effects of inulin and soya bean oligosaccharide on skatole production and the intestinal microbiota in broilers,” *Journal of animal physiology and animal nutrition*, vol. 102, no. 3, pp. 706–716, 2018.
- [194] Y. Chen, J. Ni, and H. Li, “Effect of green tea and mulberry leaf powders on the gut microbiota of chicken,” *BMC veterinary research*, vol. 15, no. 1, pp. 1–6, 2019.
- [195] M. R. Islam, D. Lepp, D. V. Godfrey, S. Orban, K. Ross, P. Delaquis, and M. S. Dirarra, “Effects of wild blueberry (*vaccinium angustifolium*) pomace feeding on gut microbiota and blood metabolites in free-range pastured broiler chickens,” *Poultry science*, vol. 98, no. 9, pp. 3739–3755, 2019.
- [196] M. Zheng, P. Mao, X. Tian, Q. Guo, and L. Meng, “Effects of dietary supplementation of alfalfa meal on growth performance, carcass characteristics, meat and egg quality, and intestinal microbiota in beijing-you chicken,” *Poultry science*, vol. 98, no. 5, pp. 2250–2259, 2019.
- [197] C. Pineda-Quiroga, A. Camarinha-Silva, D. Borda-Molina, R. Atxaerandio, R. Ruiz, and A. García-Rodríguez, “Feeding broilers with dry whey powder and whey protein concentrate affected productive performance, ileal digestibility of nutrients and cecal microbiota community,” *Animal*, vol. 12, no. 4, pp. 692–700, 2018.
- [198] D. Stanley, S. E. Denman, R. J. Hughes, M. S. Geier, T. M. Crowley, H. Chen, V. R. Haring, and R. J. Moore, “Intestinal microbiota associated with differential feed conversion efficiency in chickens,” *Applied microbiology and biotechnology*, vol. 96, no. 5, pp. 1361–1369, 2012.
- [199] Z. Feng, R. W. Hanson, N. A. Berger, and A. Trubitsyn, “Reprogramming of energy metabolism as a driver of aging,” *Oncotarget*, vol. 7, no. 13, p. 15410, 2016.
- [200] P. Louis, G. L. Hold, and H. J. Flint, “The gut microbiota, bacterial metabolites and colorectal cancer,” *Nature reviews microbiology*, vol. 12, no. 10, pp. 661–672, 2014.



- [201] S. D. Jurburg, M. S. Brouwer, D. Ceccarelli, J. van der Goot, A. J. Jansman, and A. Bossers, “Patterns of community assembly in the developing chicken microbiome reveal rapid primary succession,” *MicrobiologyOpen*, vol. 8, no. 9, p. e00821, 2019.
- [202] P. Richards-Rios, J. Fothergill, M. Bernardeau, and P. Wigley, “Development of the ileal microbiota in three broiler breeds,” *Frontiers in veterinary science*, vol. 7, p. 17, 2020.
- [203] C. Fernández-Rubio, C. Ordonez, J. Abad-González, A. Garcia-Gallego, M. P. Honrubia, J. J. Mallo, and R. Balana-Fouce, “Butyric acid-based feed additives help protect broiler chickens from salmonella enteritidis infection,” *Poultry science*, vol. 88, no. 5, pp. 943–948, 2009.
- [204] G. Dalmaso, H. T. T. Nguyen, Y. Yan, L. Charrier-Hisamuddin, S. V. Sitaraman, and D. Merlin, “Butyrate transcriptionally enhances peptide transporter *pept1* expression and activity,” *PLoS one*, vol. 3, no. 6, p. e2476, 2008.
- [205] R. P. Heaney, “Factors influencing the measurement of bioavailability, taking calcium as a model,” *The Journal of nutrition*, vol. 131, no. 4, pp. 1344S–1348S, 2001.
- [206] B. Walther, A. M. Lett, A. Bordoni, L. Tomás-Cobos, J. A. Nieto, D. Dupont, F. Danesi, D. R. Shahar, A. Echaniz, R. Re, *et al.*, “Gutself: Interindividual variability in the processing of dietary compounds by the human gastrointestinal tract,” *Molecular nutrition & food research*, vol. 63, no. 21, p. 1900677, 2019.
- [207] E. Roanes-Lozano, A. González-Bermejo, E. Roanes-Macías, and J. Cabezas, “An application of computer algebra to pharmacokinetics: the bateman equation,” *SIAM review*, vol. 48, no. 1, pp. 133–146, 2006.
- [208] M. Assfalg, I. Bertini, D. Colangiuli, C. Luchinat, H. Schäfer, B. Schütz, and M. Spraul, “Evidence of different metabolic phenotypes in humans,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 5, pp. 1420–1424, 2008.
- [209] Q. Guo, A. Ye, N. Bellissimo, H. Singh, and D. Rousseau, “Modulating fat digestion through food structure design,” *Progress in Lipid Research*, vol. 68, pp. 109–118, 2017.
- [210] J. Ubbink, A. Burbidge, and R. Mezzenga, “Food structure and functionality: a soft matter perspective,” *Soft matter*, vol. 4, no. 8, pp. 1569–1581, 2008.
- [211] J. M. Aguilera, “The food matrix: implications in processing, nutrition and health,” *Critical Reviews in Food Science and Nutrition*, vol. 59, no. 22, pp. 3612–3629, 2019.
- [212] Z. Liu and M. Scanlon, “Predicting mechanical properties of bread crumb,” *Food and Bioproducts processing*, vol. 81, no. 3, pp. 224–238, 2003.
- [213] T. Maeda, M. Kokawa, M. Miura, T. Araki, M. Yamada, K. Takeya, and Y. Sagara,

- “Development of a novel staining procedure for visualizing the gluten–starch matrix in bread dough and cereal products,” *Cereal Chemistry*, vol. 90, no. 3, pp. 175–180, 2013.
- [214] H. Li, M. J. Gidley, and S. Dhital, “High-amylose starches to bridge the “fiber gap”: development, structure, and nutritional functionality,” *Comprehensive reviews in food science and food safety*, vol. 18, no. 2, pp. 362–379, 2019.
- [215] R. Van der Sman and A. Van der Goot, “The science of food structuring,” *Soft Matter*, vol. 5, no. 3, pp. 501–510, 2009.
- [216] C. De Kruif and T. Huppertz, “Casein micelles: size distribution in milks from individual cows,” *Journal of agricultural and food chemistry*, vol. 60, no. 18, pp. 4649–4655, 2012.
- [217] H. Singh, A. Ye, and D. Horne, “Structuring food emulsions in the gastrointestinal tract to modify lipid digestion,” *Progress in lipid research*, vol. 48, no. 2, pp. 92–100, 2009.
- [218] S. L. Turgeon and L.-E. Rioux, “Food matrix impact on macronutrients nutritional properties,” *Food Hydrocolloids*, vol. 25, no. 8, pp. 1915–1924, 2011.
- [219] H. Singh, A. Ye, and M. J. Ferrua, “Aspects of food structures in the digestive tract,” *Current Opinion in Food Science*, vol. 3, pp. 85–93, 2015.
- [220] R. F. Tester and S. J. Debon, “Annealing of starch—a review,” *International journal of biological macromolecules*, vol. 27, no. 1, pp. 1–12, 2000.
- [221] L. Tavares, L. Santos, and C. P. Z. Noreña, “Bioactive compounds of garlic: A comprehensive review of encapsulation technologies, characterization of the encapsulated garlic compounds and their industrial applicability,” *Trends in Food Science & Technology*, 2021.
- [222] E. Kirtil and M. H. Oztop, “1 h nuclear magnetic resonance relaxometry and magnetic resonance imaging and applications in food science and processing,” *Food Engineering Reviews*, vol. 8, no. 1, pp. 1–22, 2016.
- [223] H. Shi, Y. Lei, L. L. Prates, and P. Yu, “Evaluation of near-infrared (nir) and fourier transform mid-infrared (atr-ft/mir) spectroscopy techniques combined with chemometrics for the determination of crude protein and intestinal protein digestibility of wheat,” *Food chemistry*, vol. 272, pp. 507–513, 2019.
- [224] N. Cebi, M. Z. Durak, O. S. Toker, O. Sagdic, and M. Arici, “An evaluation of fourier transforms infrared spectroscopy method for the classification and discrimination of bovine, porcine and fish gelatins,” *Food Chemistry*, vol. 190, pp. 1109–1115, 2016.
- [225] P. M. Falcone, A. Baiano, A. Conte, L. Mancini, G. Tromba, F. Zanini, and M. A. Del Nobile, “Imaging techniques for the study of food microstructure: a review,” *Advances in food and nutrition research*, vol. 51, pp. 205–263, 2006.

- [226] A. García-García, M. Cambero, D. Castejón, R. Escudero, and M. Fernández-Valle, “Dry cured-ham microstructure: A t2 nmr relaxometry, sem and uniaxial tensile test combined study,” *Food Structure*, vol. 19, p. 100104, 2019.
- [227] M. Langton and A.-M. Hermansson, “Image analysis of particulate whey protein gels,” *Food Hydrocolloids*, vol. 10, no. 2, pp. 179–191, 1996.
- [228] M. Langton, A. Åström, and A.-M. Hermansson, “Influence of the microstructure on the sensory quality of whey protein gels,” *Food Hydrocolloids*, vol. 11, no. 2, pp. 217–230, 1997.
- [229] A. S. Szczesniak, “Texture is a sensory property,” *Food quality and preference*, vol. 13, no. 4, pp. 215–225, 2002.
- [230] J. E. Clark, “Taste and flavour: their importance in food choice and acceptance,” *Proceedings of the nutrition society*, vol. 57, no. 4, pp. 639–643, 1998.
- [231] C. Tournier, C. Sulmont-Rossé, and E. Guichard, “Flavour perception: aroma, taste and texture interactions,” 2007.
- [232] R. Pereira, H. Singh, P. Munro, and M. Luckman, “Sensory and instrumental textural characteristics of acid milk gels,” *International Dairy Journal*, vol. 13, no. 8, pp. 655–667, 2003.
- [233] M. N. Corstens, C. C. Berton-Carabin, A. Kester, R. Fokkink, J. M. van den Broek, R. de Vries, F. J. Troost, A. A. Masclee, and K. Schroën, “Destabilization of multi-layered interfaces in digestive conditions limits their ability to prevent lipolysis in emulsions,” *Food structure*, vol. 12, pp. 54–63, 2017.
- [234] Y. Tan and D. J. McClements, “Improving the bioavailability of oil-soluble vitamins by optimizing food matrix effects: A review,” *Food Chemistry*, p. 129148, 2021.
- [235] Y. Tan, Z. Zhang, J. M. Mundo, and D. J. McClements, “Factors impacting lipid digestion and nutraceutical bioaccessibility assessed by standardized gastrointestinal model (infogest): Emulsifier type,” *Food Research International*, vol. 137, p. 109739, 2020.
- [236] S. Salentinig, “Supramolecular structures in lipid digestion and implications for functional food delivery,” *Current Opinion in Colloid & Interface Science*, vol. 39, pp. 190–201, 2019.
- [237] J. Calvo-Lerma, V. Fornés-Ferrer, A. Heredia, and A. Andrés, “In vitro digestion of lipids in real foods: influence of lipid organization within the food matrix and interactions with nonlipid components,” *Journal of food science*, vol. 83, no. 10, pp. 2629–2637, 2018.
- [238] K. Žolnere, M. Arnold, B. Hull, and D. W. Everett, “Cheese proteolysis and matrix disintegration during in vitro digestion,” *Food Structure*, vol. 21, p. 100114, 2019.
- [239] S. Dhital, R. R. Bhattarai, J. Gorham, and M. J. Gidley, “Intactness of cell wall

- structure controls the in vitro digestion of starch in legumes,” *Food & function*, vol. 7, no. 3, pp. 1367–1379, 2016.
- [240] H. Li, M. J. Gidley, and S. Dhital, “High-amylose starches to bridge the “fiber gap”: development, structure, and nutritional functionality,” *Comprehensive reviews in food science and food safety*, vol. 18, no. 2, pp. 362–379, 2019.
- [241] Y. Ogawa, N. Donlao, S. Thuengtung, J. Tian, Y. Cai, F. C. Reginio Jr, S. Ketnawa, N. Yamamoto, and M. Tamura, “Impact of food structure and cell matrix on digestibility of plant-based food,” *Current opinion in food science*, vol. 19, pp. 36–41, 2018.
- [242] A. P. Pallares, B. A. Miranda, N. Q. A. Truong, C. Kyomugasho, C. M. Chigwedere, M. Hendrickx, and T. Grauwet, “Process-induced cell wall permeability modulates the in vitro starch digestion kinetics of common bean cotyledon cells,” *Food & function*, vol. 9, no. 12, pp. 6544–6554, 2018.
- [243] A. Romano, V. D’Amelia, V. Gallo, S. Palomba, D. Carputo, and P. Masi, “Relationships between composition, microstructure and cooking performances of six potato varieties,” *Food Research International*, vol. 114, pp. 10–19, 2018.
- [244] J. E. Fannon, R. J. Hauber, and J. N. BeMILLER, “Surface pores of starch granules,” *Cereal Chem*, vol. 69, no. 3, pp. 284–288, 1992.
- [245] J. Tian, Y. Ogawa, J. Shi, S. Chen, H. Zhang, D. Liu, and X. Ye, “The microstructure of starchy food modulates its digestibility,” *Critical reviews in food science and nutrition*, vol. 59, no. 19, pp. 3117–3128, 2019.
- [246] E. De La Hera, C. M. Rosell, and M. Gomez, “Effect of water content and flour particle size on gluten-free bread quality and digestibility,” *Food chemistry*, vol. 151, pp. 526–531, 2014.
- [247] E. Capuano and A. E. Janssen, “Food matrix and macronutrient digestion,” *Annual Review of Food Science and Technology*, vol. 12, pp. 193–212, 2021.
- [248] K. Nyemb, C. Guérin-Dubiard, S. Pézenec, J. Jardin, V. Briard-Bion, C. Cauty, S. M. Rutherford, D. Dupont, and F. Nau, “The structural properties of egg white gels impact the extent of in vitro protein digestion and the nature of peptides generated,” *Food Hydrocolloids*, vol. 54, pp. 315–327, 2016.
- [249] G. A. Mueller, S. J. Maleki, K. Johnson, B. K. Hurlburt, H. Cheng, S. Ruan, J. B. Nesbit, A. Pomés, L. L. Edwards, A. Schorzman, *et al.*, “Identification of maillard reaction products on peanut allergens that influence binding to the receptor for advanced glycation end products,” *Allergy*, vol. 68, no. 12, pp. 1546–1554, 2013.
- [250] M. Teodorowicz, J. Van Neerven, and H. Savelkoul, “Food processing: The influence of the maillard reaction on immunogenicity and allergenicity of food proteins,” *Nutrients*, vol. 9, no. 8, p. 835, 2017.
- [251] M. Toda, M. Hellwig, T. Henle, and S. Vieths, “Influence of the maillard reaction

- on the allergenicity of food proteins and the development of allergic inflammation,” *Current allergy and asthma reports*, vol. 19, no. 1, p. 4, 2019.
- [252] M. L. A. Lustria, S. M. Noar, J. Cortese, S. K. Van Stee, R. L. Glueckauf, and J. Lee, “A meta-analysis of web-delivered tailored health behavior change interventions,” *Journal of health communication*, vol. 18, no. 9, pp. 1039–1069, 2013.
- [253] J. Norton, Y. G. Espinosa, R. Watson, F. Spyropoulos, and I. Norton, “Functional food microstructures for macronutrient release and delivery,” *Food & function*, vol. 6, no. 3, pp. 663–678, 2015.
- [254] R. Sharma, B. Dar, S. Sharma, and B. Singh, “In vitro digestibility, cooking quality, bio-functional composition, and sensory properties of pasta incorporated with potato and pigeonpea flour,” *International Journal of Gastronomy and Food Science*, vol. 23, p. 100300, 2021.
- [255] D. E. Garcia-Valle, L. A. Bello-Pérez, E. Agama-Acevedo, and J. Alvarez-Ramirez, “Structural characteristics and in vitro starch digestibility of pasta made with durum wheat semolina and chickpea flour,” *LWT*, vol. 145, p. 111347, 2021.
- [256] H. M. Romero and Y. Zhang, “Physicochemical properties and rheological behavior of flours and starches from four bean varieties for gluten-free pasta formulation,” *Journal of Agriculture and Food Research*, vol. 1, p. 100001, 2019.
- [257] G. Somaratne, M. J. Ferrua, A. Ye, F. Nau, J. Floury, D. Dupont, and J. Singh, “Food material properties as determining factors in nutrient release during human gastric digestion: A review,” *Critical reviews in food science and nutrition*, vol. 60, no. 22, pp. 3753–3769, 2020.
- [258] R. van der Sman, S. Houlder, S. Cornet, and A. Janssen, “Physical chemistry of gastric digestion of proteins gels,” *Current Research in Food Science*, vol. 2, pp. 45–60, 2020.
- [259] M. Langton and A.-M. Hermansson, “Image analysis of particulate whey protein gels,” *Food Hydrocolloids*, vol. 10, no. 2, pp. 179–191, 1996.
- [260] J. F. Fundo, M. A. Quintas, and C. L. Silva, “Molecular dynamics and structure in physical properties and stability of food systems,” *Food Engineering Reviews*, vol. 7, no. 4, pp. 384–392, 2015.
- [261] V. Morris and K. Groves, *Food microstructures: Microscopy, measurement and modelling*. Elsevier, 2013.
- [262] V. Bolnykh, J. M. H. Olsen, S. Meloni, M. P. Bircher, E. Ippoliti, P. Carloni, and U. Rothlisberger, “Extreme scalability of dft-based qm/mm md simulations using mimic,” *Journal of chemical theory and computation*, vol. 15, no. 10, pp. 5601–5613, 2019.
- [263] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, “Qsar-based virtual screening: advances and applications in drug

- discovery,” *Frontiers in pharmacology*, vol. 9, p. 1275, 2018.
- [264] S. K. Chakravarti and S. R. M. Alla, “Descriptor free qsar modeling using deep learning with long short-term memory neural networks,” *Frontiers in artificial intelligence*, vol. 2, p. 17, 2019.
- [265] F. Bonomi, M. G. D’Egidio, S. Iametti, M. Marengo, A. Marti, M. A. Pagani, and E. M. Ragg, “Structure–quality relationship in commercial pasta: A molecular glimpse,” *Food Chemistry*, vol. 135, no. 2, pp. 348–355, 2012.
- [266] D. Bernin, T. Steglich, M. Röding, A. Moldin, D. Topgaard, and M. Langton, “Multi-scale characterization of pasta during cooking using microscopy and real-time magnetic resonance imaging,” *Food research international*, vol. 66, pp. 132–139, 2014.
- [267] Y. Tao, Y.-C. Lee, H. Liu, X. Zhang, J. Cui, C. Mondo, M. Babaei, J. Santillan, G. Wang, D. Luo, *et al.*, “Morphing pasta and beyond,” *Science Advances*, vol. 7, no. 19, p. eabf4098, 2021.
- [268] V. Gallo, A. Romano, and P. Masi, “Does the presence of fibres affect the microstructure and in vitro starch digestibility of commercial italian pasta?,” *Food Structure*, vol. 24, p. 100139, 2020.
- [269] M. C. Gonçalves and H. R. Cardarelli, “Changes in water mobility and protein stabilization of mozzarella cheese made under different stretching temperatures,” *LWT*, vol. 104, pp. 16–23, 2019.
- [270] G. M. Bosmans, B. Lagrain, N. Ooms, E. Fierens, and J. A. Delcour, “Biopolymer interactions, water dynamics, and bread crumb firming,” *Journal of Agricultural and Food Chemistry*, vol. 61, no. 19, pp. 4646–4654, 2013.
- [271] M. R. Serial, M. B. Canalis, M. Carpinella, M. C. Valentinuzzi, A. León, P. D. Ribotta, and R. H. Acosta, “Influence of the incorporation of fibers in biscuit dough on proton mobility characterized by time domain nmr,” *Food chemistry*, vol. 192, pp. 950–957, 2016.
- [272] F. A. Manthey and A. L. Schorno, “Physical and cooking quality of spaghetti made from whole wheat durum,” *Cereal Chemistry*, vol. 79, no. 4, pp. 504–510, 2002.
- [273] V. Bortolotti, R. Brown, P. Fantazzini, G. Landi, and F. Zama, “Uniform penalty inversion of two-dimensional nmr relaxation data,” *Inverse Problems*, vol. 33, no. 1, p. 015003, 2016.
- [274] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
- [275] G. Simonetti, A. Padella, I. F. do Valle, M. C. Fontana, E. Fonzi, S. Bruno, C. Baldazzi, V. Guadagnuolo, M. Manfrini, A. Ferrari, *et al.*, “Aneuploid acute myeloid leukemia exhibits a signature of genomic alterations in the cell cycle and protein

- degradation machinery,” *Cancer*, vol. 125, no. 5, pp. 712–725, 2019.
- [276] T. Shlomi, M. N. Cabili, and E. Ruppin, “Predicting metabolic biomarkers of human inborn errors of metabolism,” *Molecular systems biology*, vol. 5, no. 1, p. 263, 2009.