**Teesside University**

# Exploration of machine learning approaches with genome-scale metabolic model-generated fluxes

Giuseppe Magazzù

A thesis submitted in partial fulfilment of the requirements of Teesside University for the degree of Doctor of Philosophy in Computer Science

Supervisor: Dr Claudio Angione
School of Computing, Engineering and Digital Technologies
February 2023

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where reference is made in the dissertation to any such works. This dissertation contains fewer than 60000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures. I certify that the statements above are correct.

Giuseppe Magazzù
February 2023

# Acknowledgements

As I approached the end of my Bachelor's Degree, I started to look around for opportunities at different universities. I had always wanted to do research, and a quick chat with Dr Claudio Angione, who had studied at the same university I was studying, convinced me that I could apply for a PhD at Teesside. I thank him for this opportunity, all the support (often, behind the curtain!) and the advice I have received. Under his guidance, I have learned how to be a good researcher, assess papers correctly and write and present ideas in a proper style.

I would also like to thank Dr Guido Zampieri, who welcomed me on my very first day at Teesside and helped me make the initial steps in this research field. Without him, I could never have become independent in my research.

Dr Yingke Chen and Dr Greg Atkinson have offered insightful comments during the annual reviews that have definitively improved my work. I am sincerely grateful to them for the time taken to review my work at different stages and for discussing with me the most unclear points.

I have been very lucky to have the opportunity to meet, discuss and interact with the bright minds that compose our research group: I thank them all for the insightful comments on my work and for the dinners out. To this group also belong all those staff members who have, in a way or another, given me suggestions on my work or career path, and the co-authors of my published papers.

A final thank goes to my family, my girlfriend and my friends, who have always been very supportive and have believed in my journey since the beginning. I would also like to take this opportunity to dedicate this thesis to my grandpa, who left us before he could see me a Doctor. *Ciao.*

# Abstract

Biology, from the ancient Greek, "study of life", is the most fascinating yet complex of all sciences. Its understanding is paramount in solving many current problems we face in our society, from curing seriously debilitating diseases, to safely devising new drugs, to determining guidelines for disease prevention. However, its complexity has so far hindered medical progress, and the objective impossibility of perfectly replicating experiments has contributed to less robust results and many non-reproducible claims. Lately, the advancement of computing technologies has led to the development of mathematical and computational methods which could shed light on the mechanisms of life which are still obscure to us to date. In particular, progress in Systems biology has opened the way to new techniques to be employed to solve long-standing problems in more efficient and robust ways. One of the newly designed methods, genome-scale metabolic models, can help simulate metabolic conditions and draw links between the molecular transformations happening at a small-scale level and the systemic modifications that organisms experience. In this work, we have explored the usefulness of such models and investigated possible approaches in the form of case studies, with the adoption of machine learning techniques. These techniques, which aim at discovering invisible patterns in the data the significance of which could lead to fundamental discoveries or directions for future medical research, currently represent the state-of-the-art approaches in countless of modern applications, and serve as the first choice when trying to advance in cutting-edge research scenarios. Our work demonstrates that some scope for these models exists, in particular in the field of precision medicine.

# Contents

# List of Figures

# List of Tables

# Data and code availability

## Using GSMM-generated fluxes with transfer learning for gene regulatory network reconstruction

The methods, data and results for this chapter are available at https://figshare.com/collections/Integrating_genome-scale_metabolic_modelling_and_transfer_learning_for_human_gene_regulatory_network_reconstruction/5237687.

## Concatenating transcriptomic and fluxomic data for yeast growth rate prediction

The novel methods developed for this chapter are available at https://github.com/Angione-Lab/HybridGroupIPFLasso_pc2Lasso.

## Integrating transcriptomic and fluxomic data with a late integration strategy for liver cancer diagnosis

The methods, data and results for this chapter are available at https://github.com/Angione-Lab/Hepatoblastoma-Children-Classification.

# List of publications

## Manuscripts in preparation

Occhipinti A., Vijayakumar S., Verma S., Doan L. M. T., Magazzù G., Moon P., Efthekhari N., Tarzi C., Angione C., 2023. Towards a unified machine learning framework for imaging, omics, and metabolic modelling. 2023

Verma S., Magazzù G., Eftekhari N., Occhipinti A., Angione C., 2023. Biologically interpretable and robust deep learning models for small imaging-omics-clinical datasets.

## Refereed journal articles

Magazzù, G., Zampieri, G. and Angione, C., 2022. Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods. Computers in Biology and Medicine, 151, p.106244.

Pio, G., Mignone, P., Magazzù, G., Zampieri, G., Ceci, M. and Angione, C., 2022. Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction. Bioinformatics, 38(2), pp.487-493.

Magazzù, G., Zampieri, G. and Angione, C., 2021. Multimodal regularized linear models with flux balance analysis for mechanistic integration of omics data. Bioinformatics, 37(20), pp.3546-3552.

## Book chapters

Vijayakumar, S., Magazzù, G., Moon, P., Occhipinti, A. and Angione, C., 2022. A Practical Guide to Integrating Multimodal Machine Learning and Metabolic Modeling. In Computational Systems Biology in Medicine and Biotechnology (pp. 87-122). Humana, New York, NY.

# Chapter 1

# The quest for better health

In the new century, the healthcare system has been facing serious challenges. Higher standards of quality of life, coupled with new discoveries in medicine, have increased the life expectancy in developed countries and thus the ageing population, which has already put a burden on healthcare systems in the entire world [3]. This, together with a chronic lack of staff [4, 5] and an increase in non-age-related disease rates [6, 7] means that in the near future access to quality healthcare will become more and more difficult. A possible answer to this has already been provided by the digitalisation of most services and the widespread adoption of IT-based solutions, which aims to organise and streamline the management of patients and staff's data. However, an alternative which is slowly making its way, is the adoption of precision medicine approaches.

Indeed, as a result of advancements in analytics and the increased availability and heterogeneity of data, data-driven methodologies have been applied to improve disease management, cohort discovery for clinical trials and early disease diagnosis, thus contributing to the development of personalised healthcare [8]. Precision medicine is a branch of the medical studies which aims at developing therapies and protocols that are tailored to the individual patients and may not be suitable for others. Precision medicine is an attempt at a more effective and efficient approach and an answer to rare diseases, and it is one of the most anticipated applications of big data in healthcare. It has recently received enormous attention for its potential in expanding the scope of disease management and prevention, by carefully monitoring the transition from non-diseased to diseased states, precisely identifying individuals at risk for disease, and individualising patient care by considering genetic, environmental, and lifestyle variability [9]. Precision medicine has been successfully applied for the selection of patient-specific treatments in complex diseases, including cancer [10, 11]. Specifically, when developing personalised therapies for cancer treatment, the mapping of molecular features for individual tumours and mutations can contribute to improving the selection of companion diagnostics and drug therapies. Consequently,

this can also provide a renewed understanding of how specific mutations arise, which in turn helps tailor more effective strategies for halting disease progression [12, 13]. Here, we use the term Phenotypic personalised medicine (PPM) to refer to an even more ambitious application that harnesses artificial intelligence to improve patient-specific medicine for existing monotherapies (therapies focusing on the use of one single drug or type of treatment) and combination therapies. PPM has aided in the design of novel drug combinations and new dosing strategies, as well as the identification of prospective markers and drug targets [14, 15]. Particularly, PPM has demonstrated potential in tailoring treatments for liver disease and acute lymphoblastic leukaemia [16, 17]. This work builds around the concept of phenotypic personalised medicine (later simply referred to as personalised medicine or precision medicine) and quests after computational approaches that allow a deeper understanding of metabolism biology at the individual level, with the aid of mathematical modelling and artificial intelligence. In particular, we will focus on Systems Biology as a viable approach to precision medicine. We will also combine this with data integration through machine learning models, in order to resolve many of the limitations of systems biology-based approaches. The rest of the chapter contains sections describing each of these aspects in detail, in the hope to provide the reader with the necessary background to understand the work later presented.

## 1.1 Biomedical data

The objective of precision medicine could not be accomplished without biomedical data, particularly patient-specific information. The high volume of this type of data is only destined to increase and the major problem already consists in developing better analytical approaches to take advantage of this.

Biomedical data are usually classified as structured or unstructured data. Structured data are those that have labels or database-like schemas. They are easily accessible via data management systems that facilitate computational interpretation and analysis. Any information recorded as measurements or signals is an example of structured clinical data. On the other hand, unstructured clinical data contain information that is not easily accessible by computational data management systems in their current form, meaning that they require ad-hoc computational methods to be processed and analysed.

Electronic health records (EHRs) usually include both structured data, such as patient demographics and clinical measurements of blood pressure, pulse and respiratory rate, and unstructured data, such as diagnoses and medications [18, 19]. EHRs would also include biomedical imaging data, e.g. computed tomography (CT), single-photon emission computed tomography (SPECT), or magnetic resonance imaging (MRI), some of which also provide high-resolution images [20].

Even if images, audio, and video streams constitute a major source of biomedical information, the largest source of unstructured information in the medical field is natural language text. This includes

published biomedical literature in the form of journal articles, EHRs, social media and other web-based sources, which are types of unstructured clinical textual data that can act as primary resources for natural language processing (NLP) [21]. In the following paragraphs we will present examples of structured and unstructured data that we used in our work, and whose presence can be found in the other chapters of the thesis.

### 1.1.1 Structured data

The structured data that the reader will find in this work can be divided into two categories: clinical data (demographics) and omics data. The following paragraph describes this second type of structured data, fundamental in all of our work.

**Omics data** The first challenge researchers in bioinformatics had to face was the production of DNA or RNA sequence reads (the so called primary genomic analysis). A second challenge consisted in the targeting of raw sequenced data, i.e. outputs from next-generation sequencing, by examining the alignment of these reads to a reference genome in order to locate gene mutations (secondary analyses). Lately, however, tertiary analyses have emerged as the most unsolved problem in bioinformatics. They regard pre-processed data (after the secondary analyses) and how heterogeneous, genomic regions interact with each other [22, 23]. Nonetheless, all three of these stages in genomic data analysis were necessary for scientists to develop the current approach that gives broader insights into the development of diseases such as cancer.

With the rapid development of next-generation sequencing technologies (also known as high-throughput sequencing techniques, a novel and faster DNA sequencing technology [24]), we are currently observing an unprecedented growth in data from various sources (consortium-based and large-scale projects), accelerated by the decreased cost of sequencing [25, 26, 22]. This exponential growth regards in particular what in biology is called omics, i.e. experimental profiles with large coverage over multiple biological domains. The principal '-omic' technologies focus on the detection of the following data in biological samples: the total gene content (for the field of genomics, i.e. the study of the genome), the total mRNA transcribed from the genome (field of transcriptomics), all the express proteins (i.e. the proteome), the complete lipid profile (for the field of lipidomics), all the compounds participating in metabolic reactions (i.e. the metabolome), the complete set of heritable phenotypic changes that do not modify DNA (field of epigenomics), and many more [27]. If the genome is all the genetic information of an organism, the transcriptome is instead the reflection of the expression level of the genes, and directly influences the frequency and quantity at which proteins are produced. The transcriptome can be measured fundamentally in two ways: with microarrays, and with next-generation sequencing technologies. While the former can be used to detect known sequences (and

their absence) by using multiple genetic probes targeting a single gene sequence (analysis which is carried out in parallel for multiple genes), the latter computationally reconstruct the original RNA by sampling a large number of nucleotide sequences (fragments of the RNA to sequence), in order to improve the sequencing accuracy. Study of the metabolome requires instead a combination of chromatographic techniques (to first separate the compounds) and mass spectrometry (for the detection of the individual metabolites by measuring the mass-to-charge ratio of their ions, after ionisation). Mass spectrometry is also used for protein detection in chemical samples.

Omics are becoming more and more important for the study of biological mechanisms and diseases, and we have made extensive use of this type of data in this work (and in particular of transcriptomic data). In each chapter, even though we did not analyse the original biospecimens directly, the collection and analysis of the data are briefly reported. Our analyses, which start from these raw measures, are explained in detail. In this work, we have focused on what we refer to as multi-omic data, i.e. collections of omic datasets, linked with each other by the fact that they all describe the same samples/patients/individuals. This approach is part of what goes under the name of multi-modal approaches, in which for every sample in our data we have several modalities (or views; omics, in our work) available. However, this strategy presents several challenges that will be described at the end of this chapter and tackled in the rest of the thesis, such as the problem of data incompleteness, i.e. some of the modalities are not present for some samples (and different modalities are missing from different samples), for which many approaches have been proposed [28].

### 1.1.2 Unstructured data

Biomedical unstructured data can be of three main types: images, videos and text. Its collection is expected to grow with the expansion of digitalised repositories compiling large volumes of data [29, 22]. These complex, distributed and often dynamic sets of biomedical data, which are increasing in availability, provide new opportunities for the development of personalised medicine and improvements of patient care through the implementation of appropriate computational techniques. The following paragraphs introduce the reader to these types of data.

**Images** In several biomedical fields, including radiology and oncology, the most commonly available data type is images. During medical image analysis, experts examine a range of bioimaging modalities, e.g. computed tomography scans, magnetic resonance imaging, and ultrasound imaging for diagnosis and treatment. These data, due to practitioners' subjective impressions and human error, are not as reliable as other patient-specific information (e.g. clinical or omic data). Nevertheless, to date, they remain the only window that clinicians have to examine patients' health status in several biomedical fields.

Positron emission tomography (PET), single-photon emission computed tomography, and functional magnetic resonance imaging are examples of functional therapeutic images that have been used to develop deep learning architectures (which will be introduced in Section 1.4) that supplement low-spatial resolution images with useful data-driven information [30].

Radiomics is the field of bioimaging which focuses on the extraction of quantitative features from images, in order to then use these features for diagnostic and prognostic purposes: cancer detection, prediction of response to treatment, monitoring disease status, personalised medicine [31, 32, 33, 34]. The extraction is performed by means of computer algorithms, and aims at uncovering information that the naked eye of the clinician cannot discover. Radiogenomics builds on top of radiomics in that it is an emerging field in personalised medicine that aims to stratify patients, evaluate clinical outcomes, and guide therapeutic strategies by combining medical images with genomic data [35].

**Videos** Biomedical videos are a type of unstructured data that share with images the characteristic of being necessary for certain diagnoses to be made. Many diseases, involving paralysis [36], ambulatory problems and in general affecting the dynamics and mobility of the human body can be ascertain only through careful and thorough observation of the patient during short time intervals.

Modern development has recently brought to light powerful approaches with applications spanning from the detection of abnormalities in the heart's dynamics [37] to the analysis of cell morphology [38], with further advancements likely to pave the way for more precise and faster diagnoses in the near future.

**Text** NLP is a growing field in biomedical research used to automatically find and interpret meaningful information in text. It is commonly applied in speech recognition and to extract information from narratives, but it has also shown potential for applications in imaging datasets [39]. Specifically, NLP can be used to automatically extract text-based information from imaging reports and convert it to a structured format that is easier to process [40, 41].

## 1.2 Systems Biology

One of the most promising fields to advance the knowledge of biology and precision medicine methods is systems biology, which takes an interdisciplinary approach to jointly analyse complex biological systems by computational and mathematical means, and can establish mechanistic connections between their components. Systems biology has its focus on analysing and understanding different factors and reactions in living organisms by examining biological processes from a global perspective, in

order to observe connections at the level of the individual cell, organism, or community [42, 43]. Systems biology can thus be viewed as a collaborative, interdisciplinary venture to examine changes in multiple systems simultaneously and under different conditions [44]. Ultimately, the aim is to improve understanding of how an organism's genetically inherited characteristics (genotype) affect its externally observable characteristics (phenotype) under a given set of conditions. The need to develop new mathematical approaches to model relations between the components of a system was recognised due to the complexity of metabolism, where metabolites are often involved in numerous chemical reactions [45, 46].

Among the tools available in systems biology, genome-scale metabolic models comprise mathematical reconstructions of biochemical reactions involving the exchange of metabolites within a cell, with the aim of conveying the changing functional states of living cells. This is achieved through predicting flux rates, i.e. the rates at which reactants are converted into products during metabolic reactions.

### 1.2.1 Genome-scale Metabolic Models

Although omics such as genomics and transcriptomics provide insights into the presence and expression of genes, pattern discovery and comparison only between genes are insufficient to gather a comprehensive understanding of disease development, therefore there has been a recent focus on the study of metabolism and metabolic networks. Studying alterations in metabolic pathways could explain the dysfunctional growth of malignant cells in addition to understanding the molecular mechanisms underpinning diseases [47, 43]. Within each cell, metabolism is the network of biochemical reactions that determines function [48, 49], and one of the main biological components affecting cellular phenotype.

With recent advances in mathematical and computational methods, we are now able to reconstruct all known pathways of these reactions as complete networks spanning the entirety of metabolic functions in living organisms [48, 49, 50]. The computational reconstruction of entire metabolic reaction networks has allowed to gain more knowledge about the interactions between genotype, environment, and phenotype for several species [51, 52]. It is a hierarchical process, composed of four steps, with the ultimate goal of elucidating the mechanistic relationship between the genotype and the phenotype [49]. The first step of the reconstruction process consists in creating an automated draft of the model starting from the annotated genome sequence of the organism of interest. The second step consists in the manual curation of the draft, and involves a time-consuming systematic acquisition of information (coupled with what already was included in the automated draft) about small molecules, proteins and in general all the available (and relevant) biochemical and genetic information for the process of interest. At the end of this step the model is chemically and thermodynamically self-consistent. The third step is the conversion of the manually curated model into a mathematical format compatible with the modern computational approaches. This entails obtaining a model which is a knowledge

base that can be queried for integrated cellular functions. Finally, the so-obtained model is tested for metabolic tasks. In particular, production of common/fundamental metabolites is simulated and it is checked that the execution of important metabolic processes works as expected. The model's predictions are also tested by comparing them with experimentally measured data. The forth step is the most laborious and it is a long-term improvement process for the model, whose aim is to align as well as possible its predictions with the real data. For this exact reason, the last process appears to be a cycle, as further modifications are due to the model every time the simulations results differ significantly from the experimental measurements.

These reconstructions are known as genome-scale metabolic models (GSMMs), in which all the known biochemical and genetic processes found in the genome are accounted for. Scope and content of these models have evolved throughout the years and expanded to include more and more details regarding the biochemical processes happening in the cell, to the point that now it is possible to trace the effects of the most various genetic parameters on computed phenotypic states [49]. This means that these models can be used to contextualise high-throughput data, assist biological system discovery, and simulate the effects of genetic and environmental perturbations. Specifically, GSMMs have been implemented to mathematically represent metabolic reaction networks and their relationships with associated enzymes and genes. The stoichiometric relationship between metabolites and accompanying chemical transformations is clearly captured in a GSMM, which can simulate metabolic processes and the corresponding gene activities [53]. The modelling of genome-scale metabolic networks usually relies on two assumptions based on the mass and charge conservation laws and the steady-state system condition. These assumptions guarantee, respectively, that the total mass of produced substrates is equal to the total mass of those consumed, and that the internal metabolite concentration is invariant over time. Using these assumptions, the flow of metabolites in each reaction of the metabolic networks (i.e. the reaction flux) can be estimated to gain a better understanding of both metabolic activity and wider biological phenomena. The omic concerning the metabolic reaction fluxes (whether they are simulated by metabolic modelling or experimentally measured) is named fluxomics, and aims to capture the in-vivo activity of the compounds involved in the metabolic pathways [54].

GSMMs are sufficiently flexible to accommodate data corresponding to specific cell types, processes or environments within the human body. In this case, we talk about constraint-based methods, which are currently the most commonly used method to estimate the flow of metabolites through a metabolic network and of which we have made extensive use in this work. Constraint-based modelling approaches typically consists in applying a series of linear constraints to calculate the range of feasible metabolic flux rates, depending upon the optimisation of one or many cellular objectives. These constraints, deriving from additional information, e.g. context-, condition- or patient-specific (hence personalised) experimental data, as well as various levels of omic and splice isoform information, help the model fully describe the metabolic functionality of a cell and aim at garnering more accurate predictions of phenotypes [55, 56] and identifying unknown phenomena in metabolism [57]. The more data types

included in the models, the more information is available to trace molecular components across multiple functional states and identify non-intuitive phenomena related to metabolism [58].

Transcriptional profiles are the most popular omic to build context-specific GSMM through two main classes of methods: (i) switch-based methods, which use a gene expression threshold to turn off reactions associated with lowly expressed genes [51]; and (ii) valve-based methods, which reduce the activity of lowly (or highly) expressed genes by adjusting the upper and lower bounds for their corresponding reactions [59]. For (ii), enzymatic activities can be used as constraints [60]. For example, an approach could consists in integrating kinetic and proteomic data using the GECKO tool [61]. In fact, since not all enzymes are active in each cell type or culture condition, context-specific or tissue-specific metabolic models can be developed from cell-specific RNA-Seq data using model extraction methods such as GIMME-like methods (e.g., GIMME, GIMMEp, and GIM3E), iMAT-like methods (e.g., iMAT, INIT, and tINIT), and MBA-like methods (e.g., MBA, mCADRE, FASTCORE, and rFASTCORMICS) [62], all of which belonging to (i). Throughout this work, we have used the second approach exclusively to generate context-specific models.

The integration of transcriptional profiles into GSMM has allowed the development of novel methods for multi-tissue modelling to study intercellular interactions. For example, Bordbar *et al.* designed an advanced model to incorporate three human cell types: myocytes, adipocytes, and hepatocytes [63]. The established multi-tissue model was then used to investigate diabetes by integrating sequencing data to define discrepancies in metabolic functions between obese and type II obese bypass patients. Similarly, the MADRID pipeline was implemented to allow the construction of tissue-specific GSMMs using both transcriptomic and proteomic data, as well as existing knowledge from drug databases to identify potential therapeutic targets for various diseases [64]. Proteomic and metabolomic data have also been integrated with GSMMs, but there are fewer approaches due to the lower number of large-scale repositories required for model validation [65, 66].

Another important aspect and main reason why GSMMs were chosen for this work is that these models are mechanistically and biologically interpretable, which has proved them to be suitable for the interpretation of omics data and the generation of hypotheses that can be experimentally validated to support and drive further research [67, 68, 69].

Examples of GSMMs that we used in this thesis are Recon2.2 [70] and iMM1415 [71]. Recon2.2 is a genome-scale metabolic model for the human organism and is an extension of model Recon2.1 [72]. This model was an attempt at reducing the shortcomings of Recon1, the first GSMM for the human species [73]. Recon2.2 follows the legacy of Recon2.1 of improving genome-scale metabolic modelling by providing a more comprehensive annotation of genes and metabolites and by correcting the chemical reactions that, in the previous models, had been left unbalanced (in terms of mass as well as in terms of charge). In doing so, the model was expanded to contain 7785 reactions (twice as

many as in Recon1), 1675 genes and more than 5000 metabolites. The model was built by consensus, i.e. by integrating the available human models into a single GSMM [70]. Recon2.2, however, is not the last genome-scale metabolic model to have been developed for the human organism. Recon3D [74] and Human1 [75] have been since introduced to further advance metabolic modelling for the human species. However, even though these models do contain more information and detail than Recon2.2 (Recon3D almost doubles the number of reactions of Recon2.2 and contains functional annotations of more than 3000 proteins, while Human1 adds improvements in reaction reversibility and stoichiometric consistency to Recon3D), we have found it difficult, in practice, to use machine learning models and certain computational techniques (such as flux variability analysis) with them, due to their extremely high number of reactions. This would make most of the analyses presented in this thesis impractical in terms of computational time required. For this reason, given that Recon2.2 presented a sufficient level of biological detail for our work, we chose to use this model rather than the most updated ones. iMM1415 is, instead, a GSMM for the mouse (*Mus Musculus*) organism, built starting from Recon1 [71]. The structure of the human GSMM was filtered so as to retain only the reactions associated to homologous genes in the human and mouse organism, and the metabolic "skeleton" so obtained was then integrated with the addition of reactions known to occur in mouse metabolism and necessary to have a model capable of performing the most important metabolic tasks (production of biomass and ATP, for instance). This process was iterated several times and the added reactions cross-checked with the KEGG and Entrex Gene databases [76, 77] to make sure that the differences in metabolic functions between the two organisms were properly represented, thus obtaining a final model comprising 3724 reactions and 1415 genes. Just like for the human GSMM, we chose this model because a more up-to-date version (based on Recon3D, with a 7-fold increase in the number of reactions) would have been computationally intractable for our analyses [78]. The next two paragraphs explain how we used GSMMs (and these two models in particular) in the studies described in this work.

**Flux Balance Analysis** Flux balance analysis (FBA) is currently the most popular tool used to estimate the flow of metabolites in metabolic networks and identify the range of feasible flux values [79, 80]. FBA calculates the rate at which compounds are consumed or produced during metabolic reactions, and is an approach extensively used in the study of GSMMs [81, 82], in particular in the presence of constraints. The underlying idea is that constraints are imposed on the flow of metabolites through the network by assigning a lower and upper bound to each reaction in the network, which regulates the minimum and maximum amount of flux that would be allowed (valve-based approach). This feature also allows for flux predictions based on various experimental conditions [83].

The way FBA works is the following: in general terms, FBA solves a linear optimisation problem to find the value of the flux rates. Normally, however, due to the usually greater number of reactions than metabolites the problem is underdetermined and infinite solutions can satisfy it, which is why its resolution usually requires that one targets a subset of the reaction fluxes, whose rate has to be

optimised or minimised. In particular, since identifying the true objective/target reaction(s) for a cell remains a challenge, the maximisation of biomass is often considered a reasonable target for both bacteria and cancer cell models, whereas it may not be the best choice in the presence of mammalian cells [84, 85]. In fact, the true cellular objective might vary across cells in the same tissue, between tissues, and even within the same cell throughout time.

The FBA framework adopts two principles. The first, being this method based on GSMMs, is that the total sum of fluxes in the model has to be equal to zero (from which "Balance" in the name). The second principle, stemming from the first, is that the entire system is considered at steady-state [79]. This means that the concentration of metabolites in the model is considered constant, which in turn assumes that the metabolic fluxes are temporally invariant (do not change over time) and spatially homogeneous (if the model is composed of different compartments, each metabolite is accounted for independently in each of them, as if there were multiple different metabolites) [49]. Even though this hypothesis seems not applicable when trying to model the real world, the experience has demonstrated that the simplification of the models in such terms is still capable of offering meaningful insight in the functioning of biological systems. Moreover, one could say that this approach, which considerably streamlines the usage of these models, finds validation in the common approximation technique used in Physics which consists in analysing a system in such a small (*infinitesimal*) interval that all the temporally dependent variables can be considered constant. In our case, the steady-state principle means that we are observing the metabolic network in an extremely small time interval, in which indeed the concentration of the metabolites can be considered constant. Finally, an even more compelling reason why the steady-state assumption is important to make is that for most applications of FBA the accurate measurement of metabolic fluxes is not possible, especially for complex organisms such as the human one [86].

The general mathematical formulation of FBA is the following:

$$\begin{aligned} &\max\ \mathbf{c}^\top \mathbf{v} \\ &\text{subject to}\quad \mathbf{S}\,\mathbf{v} = 0, \\ &\qquad\qquad \mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub} \end{aligned} \tag{1.1}$$

where $\mathbf{v}$ represents the vector of metabolic fluxes, $\mathbf{c}$ is a constant vector of coefficients (usually boolean, used to select which reactions to optimise), $\mathbf{S}$ is the stoichiometric matrix, i.e. a matrix whose entries are the stoichiometric coefficients of the compounds produced and consumed by the metabolic reactions (associated to the columns of the matrix), and $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ are, respectively, the lower and upper bounds of the metabolic reactions.

The first condition, $\mathbf{S}\,\mathbf{v} = 0$, is the condition that imposes the steady-state, since the product of the stoichiometric matrix with the fluxes vector results in the vector of the variations of the metabolite

concentrations. By imposing that these variations are 0, we are binding the system to the solutions for which the metabolites concentration does not change, i.e. only the steady-state ones. Clearly this makes our problem easier, as the system becomes a linear one with constant coefficients. The second condition, instead, follows the constraint-based modelling approach. By applying experimentally-derived bounds to the reactions, we hope that the solutions we obtain from the model will be tailored towards specific individuals (patient-specific), environmental conditions or tissues, in the case of cells (tissue-specific). Moreover, this is often necessary to have biologically meaningful results.

In order to do so, we have to define a way to translate the activity at the gene level into activity at the reaction level. In particular, each metabolic reaction is usually controlled by more than one gene, in a specific combination named gene set. In a GSMM, the relation between gene sets and reactions is expressed in the form of gene-reaction rules, i.e. formulae combining the genes in the gene sets with the logical operators `AND` and `OR`. For example, if a reaction can be equally catalysed by two enzymes (namely, the two enzymes are *isozymes*), this relation will be encoded through an `OR` operator between the two corresponding genes. Conversely, an `AND` relation identifies enzymatic complexes where both genes are necessary for the reaction to occur. Following METRADE [87], we change the reaction bounds of the genome-scale metabolic model by assigning a gene expression value to each gene set, which then affects the lower and upper bound of the corresponding reactions. Such expression value is obtained by converting the logical operations into maximum/minimum rules, according to the following map:

$$\begin{aligned} \Theta(g_1 \wedge g_2 \wedge \cdots \wedge g_n) &= \min\{\theta(g_1), \theta(g_2), \ldots, \theta(g_n)\} \\ \Theta(g_1 \vee g_2 \vee \cdots \vee g_n) &= \max\{\theta(g_1), \theta(g_2), \ldots, \theta(g_n)\}, \end{aligned} \tag{1.2}$$

where $\theta(g)$ represents the expression level of gene $g$ and $\Theta$ represents the effective expression level of the gene set $\{g_1, g_2, \ldots, g_n\}$. This final value is then used to define the reaction bounds of the model according to some formulae, which are dependent on the available data and task at issue. This approach has been used extensively in this thesis.

As an example of the application of the above map, let us consider the following gene-reaction rule for the hypothetical reaction $r_1$:

$$r_1 : ((g_1 \wedge g_2) \vee g_3) \vee g_4$$

By iteratively applying Equations 1.2 to the above formula, starting from the innermost brackets to the outmost ones, we therefore obtain

$$r_1 : \max\{\, \max\{\, \min\{\, \theta(g_1), \theta(g_2)\}, \, \theta(g_3)\}, \, \theta(g_4)\}$$

An apparently more complicated example is the following:

$$r_2 : g_1 \wedge (g_2 \vee g_3 \vee (g_4 \wedge g_5 \wedge g_6)).$$

Here we have `AND` and `OR` relations between more than two genes, but the conversion of the formula

follows the same rules, and we obtain

$$r_2 : \min\{\, \theta(g_1), \max\{\, \theta(g_2), \theta(g_3), \min\{\, \theta(g_4), \theta(g_5), \theta(g_6)\}\,\}\,\},$$

The formulation in Equations 1.1, as we wrote, is in fact that of an optimisation problem, since we are maximising (or, equivalently, minimising) the flux rates of one or more reactions. This is necessary otherwise, as stated above, the solution of the linear system would be underdetermined, meaning infinite solutions would be available, from which the need for the new "constraint".

However, as previously explained, a clear choice of what reactions to optimise is not obvious in all cases. Since a correct decision is usually essential for the functioning and efficacy of the model, alternatives have been developed to remedy this.

To record changes in metabolism over time, several dynamic flux balance analysis (dFBA) techniques have been designed to increase the accuracy of predictions [88]. Within their computational frameworks, these approaches incorporate kinetic parameters and changes in the concentrations of specific metabolites over time. Other methods modelling unique solutions that are more consistent with observed metabolic states may consider conditional dependencies or thermodynamic uncertainties within the metabolic network [89, 90, 91, 92].

Data mining and constraint-based modelling have also been combined in unsteady-state FBA (uFBA) to estimate metabolic fluxes in dynamic conditions [93]. Principal Component Analysis (PCA) and linear regression were integrated to define an FBA model starting from metabolomics data. Since whole-metabolome measurements are generally difficult to obtain, uFBA includes an algorithm to estimate unmeasured metabolite concentration differences on the basis of measured ones. The obtained constraint-based model can be used for traditional FBA, variations of it, or related analysis in dynamic conditions.

Flux Variability Analysis (FVA) is an alternative FBA-derived approach which aims at investigating the metabolic limits of a model, by maximising and minimising alternately all the fluxes, subject to the optimisation of a group of reactions [94]. We used this approach in the case study described in Chapter 4, as in that case it helped differentiate patient-specific models better than how FBA could.

In general, however, when the assumption of maximum biomass is not appropriate, variations of FBA can be applied to characterise the solution space without choosing any objective. FVA is a borderline case, given that an objective function must be chosen anyway as further constraint to the multiple optimisation problems. A powerful alternative would be random flux sampling, which determines the feasible solution space for fluxes by approximating their distributions, while enabling a comprehensive understanding of the interplay of pathways [95]. The main problem of this solution is that, for big models, its computation of the flux distributions becomes prohibitive.

**Data and Models integration** Integration of FBA, FVA or sampling techniques with images, multi-omic and radiogenomic data, can potentially provide more biological knowledge for disease diagnosis and prognosis. Indeed, using GSMM-derived metabolic data in a multi-modal setting could be leveraged to investigate the correlation between these different biomedical data types. This correlation could also improve the predictive accuracy of high-risk patients and allow the implementation of more accurate approaches in disease detection or survival analysis, for instance.

The integration of histopathological/radiological images, multi-omic data, and constraint-based GSMM-derived data could potentially provide more comprehensive tissue-specific features and biomarkers for precise diagnosis and prognosis. So far, it has been shown that medical imaging, gene expression profiling, GSMM, and computer-aided diagnosis can play a significant role in early diagnosis and prognosis of several diseases and conditions, such as tumours or lesion detections [96, 97, 98, 99, 100], but no study has been carried out on multi-omic and imaging data integration in a GSMM framework. These data modalities could be investigated independently using advanced statistical models or they could be integrated through multi-modal approaches in order to develop more accurate models and potentially lead to interesting discoveries. However, despite the advancements in medical technologies, the integration of these data types remains a challenging task, due especially to data heterogeneity and high dimensionality.

When integrating multimodal data with GSMMs, the explainability and interpretability of the model used will also need to be taken into consideration. Most highly advanced artificial intelligence approaches are characterised by the lack of understanding of how they work, because of their complex interconnected processes. Model interpretability, therefore, represents a considerable challenge, especially when working with biomedical data, and becomes even more difficult with the integration of heterogeneous data sources. Even though, as we will see in the rest of the thesis, one of the reasons why the use of GSMMs has increased is that they can provide mechanistic information about the metabolism of the patients, in order to make the entire multimodal framework transparent and identify significant biological insights, robust and model-agnostic interpretation techniques may need to be applied (especially with the most complex models), such as SHAP, LIME, or permutation-based feature importance [101]. Considerable effort has been dedicated to tackling this issue, including the Explainable AI (XAI) program launched by DARPA [102]. There are several ongoing projects that aim to enhance the explainability of machine learning and deep learning models by proposing new statistical approaches and visualising features in order to gain the trust of the public and physicians. However, reliable multimodal approaches for interpretability and explainability in a general setting are currently lacking. For this reason, when integrating GSMMs with multimodal machine learning approaches, the models should be adapted to facilitate the identification of new biomarkers and the biological interpretation of important features that can improve model reliability[103]. Multimodal approaches will be presented more thoroughly in Section 1.5, and two case studies will be discussed in Chapters 3-4.

**Growth prediction** Choosing an objective function to optimise is difficult even when the task consists in predicting growth. Microbes, for instance, are characterised by high variability in their growth curves, which complicates the adoption of a GSMM-based approach. Growth variability in microbes can be due to genetic factors, environmental conditions and external stress factors. For example, bacterial populations have been found to replicate at different speed in different body sites [104] or in presence of inflammation [105]. Other "hidden" factors, such as the presence of rotating magnetic fields (and their frequency), were shown to influence growth dynamics and metabolic activity too [106]. To take into account this natural variability when predicting growth, [107] used a gaussian process. They controlled for all the possible "confounding factors" that affect growth such as genetic mutations, stress conditions and environmental perturbations explicitly by increasing the dimensionality of the kernel of the process, and were able to test for differential growth rigorously thanks to the probabilistic nature of the model. When using a GSMM, however, this approach would translate into integrating these additional "constraints" into the metabolic network. For example, environmental changes (as in the medium) are the simplest to take into account, since it is possible to force a minimal intake of a compound by modifying the bounds of the consuming reactions. However, more difficult is to adopt a proper definition of biomass, especially when it can change over time [108]. [109] have proposed to model this problem by considering a linear combination of all possible compounds of interest as biomass objective, with the coefficients being determined computationally. Finally, additional stress conditions can be integrated into a GSMM by considering the active/inactive metabolic pathways that such conditions would establish in the organism [110, 111]. However, even though some solutions have been proposed, more powerful techniques will still need to be developed, since integration of variability factors into predictive models can definitively make them more realistic but, on average, not necessarily more accurate [112].

### 1.2.2 Kinetic modelling

Genome-scale metabolic models are not the only possible modelling approach for biological systems. There is a much older alternative, called kinetic modelling, which instead focuses on the dynamics of such systems. When using flux balance analysis with a GSMM, the fundamental assumption is that the system is being studied at the steady state, which makes it very difficult to investigate time-variant biological mechanisms such as genetic/metabolic regulations and accumulation of metabolic intermediates [113]. Even variants of FBA such as dFBA cannot mitigate this problem, because the concentration of metabolites (albeit only the internal ones) is still considered constant [88]. The approach followed by kinetic modelling, instead, consists in using ordinary differential equations to investigate the dynamics of the system, and capture all of its time-dependent and ever-changing features (namely, the reaction kinetics). Unlike genome-scale metabolic models, this type of mathematical modelling heavily relies on precise knowledge of mechanistic relations between the various biological

entities in the system, and can describe instant changes at a very low-level. Kinetic modelling has successfully been used for the design of yeast mutants for fatty alcohol overproduction [114], for the analysis of dark fermentative hydrogen production [115], and to investigate molecular energy converters [116], among the many, and can either be structured (taking into account metabolic pathways) or unstructured (only used for microbial systems) [117]. Even though kinetic modelling can describe how metabolic networks change and evolve with time, its usage for many modelling tasks has been quite complicated. The use of ordinary differential equations instead of a linear system with constant coefficients (the concentrations of the metabolites), in conjunction with the necessity of very deep and precise knowledge of the mechanisms that govern biological systems has historically hindered the adoption of these models for large-scale analyses. Indeed, in order to formulate dynamic mass balance equations, a huge number of kinetic parameters are required, a number which grows exponentially at the increase in size of the system. Given that the values of many of these parameters cannot be measured directly, and that they can change with time, optimisation algorithms need to be used to estimate them, but the search space may easily become too large to solve the mathematical problem associated [49, 113]. For instance, a system with $>$ 1000 variables and $>$ 4000 parameters (size not uncommon among genome-scale metabolic models) would take thousands of hours of CPU to be solved, thus making kinetic modelling computationally intractable for moderate to large systems [118]. Finally, another complication comes from the fact that the values of kinetic constants can change over time because of evolution, and different individuals of the same species could have distinct values simply because of (epi)genetic differences [49]. Several solutions have been proposed recently: [119] have devised a framework to develop a kinetic model of the metabolic network of an organism starting from a GSMM, by simplifying the original model and using linlog kinetics to reduce the computational burden, while [120] have instead discussed reduction techniques to obtain smaller, yet functional kinetic models starting from bigger ones. It has also been suggested to use kinetic modelling in conjunction with genome-scale metabolic models to investigate in more quantitative detail the results obtained from the latter [113]. However, for all the above reasons, the adoption of kinetic modelling in the field of system biology is not as widespread as the usage of GSMMs, on which instead the work in this thesis is entirely based.

## 1.3 Machine Learning

Machine learning can be described as a subset of Artificial Intelligence comprising algorithms that can improve their performance on a task through experience, given a certain processable input from which they are able to learn and generalise. Beyond their potential, their widespread usage in bioinformatics and computational biology is also due to the limited assumptions they require compared to other statistical or computational approaches. This makes them essential in a number of tasks, ranging from the understanding of RNA folding to estimating the impact of mutations on splicing, and from the

exploration of gene expression profiles to reconstructing phylogenetic trees, and indeed we have used them extensively throughout this work [121, 122, 123, 124]. In this section, we will investigate the characteristics and issues of machine learning and describe some applications.

In recent years, machine learning has emerged as a key research tool for personalised medicine and the inspection, interpretation, and exploitation of multi-omic data. Both personalised medicine and the analysis of multi-omic data have benefited from it and are expected to further develop in the near future, helping overcome longstanding issues in the field of medicine [125, 126]. In particular, there has been an increase in the application of machine learning to extract more information from biological systems. The reason is that unlike statistical models, which are designed to mathematically formalise relations between variables for hypothesis testing and uncertainty estimation, machine learning techniques can be used to make predictions and also transform the data into shapes that can highlight some of their hidden characteristics, with the potential of learning not only from the input but also from the output of their analysis [127, 128]. Relying on modern powerful computing architectures, machine learning algorithms have become the go-to suite of methods in almost every field of data analysis. A number of recent developments in the application of machine learning to biological problems can be found here [129]. In the rest of this section we will describe some of the latest advances and introduce concepts that have been used in the case studies which compose this work.

### 1.3.1 Data preprocessing

The first aspect to take into account when working with machine learning models, is the available data. This is true for every experimental analysis: if the quality of data is not good enough, neither will be the quality of the results or, as usually stated by machine learning practitioner, "garbage in, garbage out".

In order to prepare raw data for machine learning, there is a variable level of pre-processing steps that each dataset is required to undergo. This includes data normalisation procedures, reduction of noise or removal of biases, feature extraction (i.e. the generation of new, possibly more relevant features, from the original feature set), and data labelling [130]. Among these, a more advanced technique is the matrix factorisation (or matrix decomposition) of the data, which is a common measure for reducing noise in datasets by breaking down/simplifying matrices into simpler constituents. A subclass of this is non-negative matrix factorisation, that has been used successfully to identify microbial guilds in metagenomic data for a microbial community [131].

When dealing with particularly high-dimensional datasets, it can also be helpful to perform feature selection to reduce their complexity and potentially obtain better predictive performance. The rationale behind this is that, in the presence of many variables in the dataset, relevant information for the

task is more likely to be contained in more than one variable (which leads to redundancy), with the vast majority of them containing little to no information. This is particularly true in biomedical applications dealing with omic data, which are usually highly correlated and noisy. Variables (also known as features) are defined as measurable properties of the observed process and are referred to as the input of modern machine learning algorithms. The main focus of feature selection is to select a subset of features that can effectively describe the data while reducing the effects of noise and removing irrelevant variables, e.g. correlated variables that do not include extra information and simply result in noise for the model. For instance, when two features are perfectly correlated, only one is necessary to describe the data. Hence, by removing any dependent variables the dimensionality of the data can be reduced significantly, which can improve model performance (at least in terms of computation time). To identify and remove redundant features, a suitable feature selection method that measures the relevance of each feature must be selected. This will depend on the problem under investigation and the type of machine learning algorithm to be applied [132]. In particular, feature selection methods can be divided into two groups: filter and wrapper methods.

Filter methods are feature selection techniques based on feature-ranking strategies. Specifically, a relevant ranking criterion is used to rank the features and then a threshold is used to select a subset of them. An example could be the ranking of the features based on their marginal association or correlation with the outcome variable [133]. Filter methods have been successfully applied to ordinary linear models with normal errors and generalised linear models [134, 135, 136]. Specifically, rank-based feature selection methods have been widely used in systems biology for metabolic data and genome-wide association studies [137]. Grissa *et al.* used a combination of rank-based methods to select the best $k-$features and discover predictive biomarkers in metabolic data [138]. The presented approach used feature selection methods from metabolomic data analysis based on the Pearson correlation coefficient and mutual information to discard highly correlated variables. In Chapter 4 we have used filter methods to remove redundant features from our data.

In wrapper methods, features are selected based on the performance of a predictive model [139]. Specifically, the selection criterion uses a search algorithm to find the set of features giving the highest predictive performance. Since evaluating all the $2^N$ subsets that can be selected from $N$ features is usually computationally very expensive, sequential search or evolutionary search algorithms such as genetic algorithms (GAs) are usually applied [140]. While sequential search algorithms start with an empty set and add (or remove) features until the defined criterion is satisfied, heuristic search algorithms move between different feature subsets based on a predefined heuristic.

Finally, some models (such as Lasso [141]) conduct a feature selection step automatically, as part of their learning process. We have focussed on some of them in Chapter 3.

Data preprocessing is fundamental in machine learning and has been applied in all the case studies

presented in this work, albeit in different instances, and fashions.

### 1.3.2 Types of learning

Depending on the task and the available data, machine learning models may necessitate different types of learning paradigms. This usually means that for different tasks and different datasets, certain models are better than others, or cannot even be used. The following paragraphs describe what the typical learning approaches are and how the training is structured.

**Supervised learning** Supervised learning is probably the most common learning paradigm for machine learning models. It is called supervised because the available data to train and test the model contain the variable that we want to predict (whether it be categorical or continuous). This means that the model can "see", during the training, what the correct prediction would be for each of the samples it is training on, and that it is possible to define a performance metric that estimates how accurate the model is in its predictions. Examples of models that can be trained in a supervised setting are support vector machines, random forests and neural networks.

The training procedure follows this structure: the available data are split into two sets, named training and test set. The former is used to train the model (and, depending on the model itself, to choose its hyperparameters), which will learn the patterns present in the data, while the latter will be used to obtain an unbiased estimate of the model's performance on unseen data. We used the supervised learning paradigm in Chapters 3-4.

**Unsupervised learning** Unsupervised learning is a learning paradigm in which the data, unlike in the supervised setting, do not present a "ground truth" value, meaning that there is no exact, unique way to estimate the performance of the model. In this case the machine learning practitioner has to choose one or more heuristics to determine if the model is performing well or not. Usually this type of learning is used when the purpose is not to make a specific prediction for an individual sample but instead when we want to obtain insights from the data, generally in the form of patterns. An example of an unsupervised learning task is clustering, and an example of unsupervised learning algorithm is Principal Component Analysis.

In this case all the available data are used, since there is no way to determine a priori whether the correct patterns have been detected (meaning that leaving some data out to check the model's conclusions would not be useful). For instance, in clustering tasks, it is not always possible to determine the meaning of the newly found clusters, and more than one clustering solution might be possible, meaning that, as stated above, heuristics should be adopted to determine if the results are satisfactory

or not.

In the case studies presented, PCA was adopted in Chapter 2 to reduce the dimensionality of the data, and in Chapter 4 for data exploration.

**Semi-supervised learning** Semi-supervised (or weak) learning is a learning paradigm in which only a small portion of the available data is labelled. This means that semi-supervised learning can be considered to be halfway between supervised and unsupervised learning. The models that can be used in this settings can have very different architectures, we presented an approach in Chapter 2 for dealing with known and unknown gene regulations.

**Reinforcement learning** Reinforcement learning is a learning paradigm with apparently limited scope but very powerful. It is an optimisation technique based on the concept of feedback and agent-environment interaction, with the task to solve being framed as follows: an agent (or more) interacts with the environment in which it exists (which can be totally virtual or, in case of robots, even physical) and is rewarded for its actions according to a policy function that aims at solving, in a finite sequence of steps (or states) the original problem. Every time the agent commits to an action, the interaction with the environment determines the reward which, in turn, influences the future actions of the agent. This type of learning is used in many practical applications such as robotics, self-driving cars and video games. The formulation of the reinforcement learning paradigm is so general that it is an approach usually suitable for the most disparate tasks, and it is indeed studied in many other disciplines.

In this case, unlike the previous two methods, the use of data may not be even necessary, as all the required information is already encoded in the agent-environment interaction mechanism.

### 1.3.3 Types of prediction tasks

Once the data have been preprocessed, the machine learning algorithm can be applied to solve the task at issue. In particular, there are three general types of problems (excluding the more general reinforcement learning paradigm) that the machine learning practitioner can attempt to solve: classification, regression and clustering. The following paragraphs aim at providing a brief introduction to each of them.

**Classification** A classification problem is a task in which we are trying to assign a category (or class) to an individual entity (sample). Common classification tasks in biomedicine can consist, for instance, in determining if a patient is ill or not. Classification algorithms include support vector machines

(SVMs), k-Nearest Neighbours (kNN), self-organizing maps, random forests, and locally weighted learning. For example, an SVM model was proposed in [142] to classify genes as essential/non-essential, given a training set of flux-coupled features. Together with kNN, it is an instance-based classification algorithm, since it bases its prediction on instances seen during the training phase and stored in memory. The case studies described in Chapters 2-4 are classification problems, with the latter using an SVM as main prediction model.

**Regression** Regression is a learning technique that can be used to model and predict continuous variables, as opposed to classification which instead is used to predict categorical ones. Indeed in this type of problem, the input features are used to determine the numerical value which is the objective of the prediction task. The most common regression model is linear regression, which assumes a linear relationship between the input features and the value to predict. Similarly, multivariate linear regression can be applied when the several dependent variables depend on multiple explanatory variables [143]. In general, however, when dealing with multiple independent variables the model becomes more prone to overfitting (i.e. to find spurious correlations/patterns in the data) and ad-hoc approaches must be applied to deal with cases of data multi-collinearity (i.e. correlation among the input variables) and data noise [144], such as feature selection. Regression techniques usually can be easily interpreted, and they can provide insights into the relationships between the predicted value and the input features. As a consequence, regression models have been used in the clinical setting for medical diagnosis, process control, quality assurance, process optimisation, and quality control for years [145]. The case study described in Chapter 3 is a regression problem.

**Clustering** The term clustering refers to a particular type of machine learning problem which consists in grouping (or clustering, hence the name) the data in a way that samples in the same cluster are more similar to each other than to samples in other clusters. It can be considered a type of classification when no information regarding the classes (not even how many they are) is available. However, unlike classification, this approach does not require labelled data, i.e. there is no need to have information about which samples constitute which groups (in the presence of labelled data, this would indeed be a classification task). Therefore, this type of problem is an unsupervised task in machine learning, as opposed to supervised tasks such as classification and regression. The general approach to clustering consists in using a mathematical function to identify a degree of similarity (or dissimilarity) between the samples being clustered. Several algorithms can be applied to identify the clusters, such as k-means, affinity propagation, and hierarchical clustering [146]. PCA is also an unsupervised technique which has been widely used as a clustering technique within the context of metabolic modelling [147]. It is usually employed for dimensionality reduction or feature extraction, but was used to identify data similarities from multidimensional biological datasets in [148] as well. In this case, one can simply use PCA to compute a set of principal components representing the patterns encoding the highest

variance in the dataset, and by analysing the variables correlating with the principal components, they can try to identify hidden patterns and potentially cluster the samples. Hierarchical clustering, a type of clustering which assumes the existence of a hierarchy among the clusters in the data, was applied on metabolomics data to separate samples from different origins (such as wild-type or knockout mutant samples) and to identify further relations within the data in [149]. As already mentioned, PCA was used in Chapters 2-4, while hierarchical clustering was used in Chapter 3. Finally, clustering can also be used for anomaly detection as in [150], in which the authors used it to discover hidden patterns and outliers in the metabolic reaction fluxes.

### 1.3.4 Applications with biomedical data

Machine learning has been increasingly applied in bioinformatics problems and with biomedical data. For example, classical machine learning techniques such as random forests and support vector machines have been adopted to predict cardiovascular events and diagnose acute coronary syndrome [151, 152]. Multivariate logistic regression has been applied to measure the association between imaging and genomic features from 48 patients with breast cancer [153]. Then, the subtypes that showed a positional association with the image features identified in the first step were analysed, thus finding that luminal B cancer (a subtype of breast cancer) is associated with a subset of MRI features. Supervised and unsupervised NMF models have also been used to identify the features of microbial communities and infer the ecological interaction networks of different gut communities, starting from high-dimensional metagenomic samples [131]. NMF allowed to transform the complex microbial data into a low-dimensional representation, thus simplifying the analysis by searching for temporal patterns in the microbiome. Furthermore, it allowed to capture the differences between the microbial compositions of two groups of communities, while retaining biological interpretability.

Still, when using gene expression data, some extra adaptations need to be made. Gene expression is a stochastic process, meaning that in a collection of cells (a tissue, for example) the level of gene expression varies in a probabilistic way, with its variability being influenced by regulatory factors, gene state variables and by the state of the overall biological system [154, 155]. One possible explanation for such variability is that it provides higher adaptability/plasticity and therefore can increase the fitness of the cells, especially when exposed to fluctuating conditions [156, 157]. This characteristic can prove very useful, for example, for cells of the immune system [154, 158] and for cancer cells, as it gives them higher chances to survive and therefore replicate in a hostile environment [157, 158, 159, 160, 161, 162]. Genes with high expression variability have also been shown to be linked to various diseases, thus drawing a link between this "flexibility" and the susceptibility to diseases and treatments among different individuals [156, 163]. Due to these reasons, gene expression variability has started to be studied in the hope that this will lead to a better understanding of gene regulation mechanisms [154, 162, 164, 165, 166, 167].

However, the source of variability in gene expression can also be technical (when it is measured), and not only biological, in which case it should be estimated and corrected for [154, 155]. In such a case, we usually talk about batch effect, which can be caused by the different times of the day the same experiment is conducted at, by the choice of reagent lot (or, indeed, batch), or by other non-biological factors. When using a machine learning model, however, this may be problematic, since it could affect the model's generalisation ability. For this reason, several approaches have been adopted to deal with gene expression variability of technical origin, each affecting differently distinct groups of genes [168]. In [169], for example, it was noted that the use of gene modules (obtained from a gene co-expression network analysis) can be more robust than raw gene expression when the machine learning model is trained on a dataset and tested on another. However, when joining more datasets together, the performance of raw gene expression can be superior. In [170], instead, a hierarchical Bayesian mixture model was successfully used to simultaneously correct for the cell-type dependent technical variation and the biological variation of single-cell RNA-seq data. In Chapter 4, when we used gene expression samples from cancer tissues from three different datasets (which can only exacerbate the variability of technical origin in the measured data), we instead used ComBat [171] and cross-validation to properly "amalgamate" samples from different datasets in order to maximise the generalisation ability of our machine learning models (we could not use gene modules in our GSMM and did not have single-cell data).

## 1.4 Deep Learning

Deep learning is a subfield of machine learning which investigates the capabilities and usage of Artificial Neural Networks (ANNs). The history of the field is long, with two periods of dormancy. The first use of ANNs dates back to 1943, with the threshold logic unit (TLU) [172]. In 1957, the first perceptron was created, while Widrow invented the Adaptive Linear Neuron (ADALINE) [173]. The perceptron was an algorithm devised to classify an object into one of two classes (i.e. it was a binary classifier). It calculated the class of the object by using a linear threshold function, and therefore had the limit of being capable of learning only linearly separable patterns [174]. This entailed that such models had very basic limitations in terms of what they could learn: for instance, the XOR issue (networks back then could not approximate the simple XOR function) almost sanctioned the end of the entire field, and considerably slowed down the research, even though other solutions had already been devised in the same period, such as the MultiLayer Perceptron (MLP) [175]. Only in the last decade, with the advancement of hardware technology and new ideas, the development and usage of increasingly sophisticated neural networks has started to accelerate at unimaginable speed. This second generation of neural networks, which includes MLPs, convolutional neural networks (CNNs) and long short-term memory units for recurrent neural networks, learns mostly through the back-propagation mechanism. Lately, many novel advanced types of deep neural network architectures have been defined: restricted

Boltzmann machines, deep belief networks, autoencoders, and deep convolutional neural networks [176, 177, 178].

Inspired by the human brain, neural networks are weighted graphs in which each input node is connected to one or more output nodes via a hidden layer that allows the detection of non-linear relationships between the input and output variable(s) [179]. Deep neural networks contain multiple hidden layers comprising of many nodes, also called neurons, which enable them to represent functions of increasing complexity and can generate extremely accurate predictions when correctly optimised [180, 181]. The neurons/nodes are combined to produce an output (such as a neuron firing) at a later layer, based on an activation function, which introduces a nonlinearity in the network (and biologically simulates the rate of the neuron's action potential). From a theoretical point of view, the flexibility that these models are endowed with is such that they can approximate any arbitrary function - provided that they contain a sufficient number of computational units (nodes/neurons) and network layers (this remark results from the so-called universal approximation theorems) [182]. The layered structure of ANNs allows these computational models to learn and represent data at different levels of abstraction, aiming at reproducing the brain's functionalities when processing complex information, such as images, text and speech. One of the main advantages of these models is that there is virtually no need for expert-curated input features, as they automatically learn appropriate representations of the data. The MLP is a particular case of ANN, in which the layers are composed by perceptrons (as nodes). However, in some if not all of these, unlike in the original perceptron formulation, the threshold function is nonlinear. This small expedient, together with the use of more than one layer in the network, is already sufficient to make the MLP a universal approximator, meaning that it can potentially approximate any function (even non-linear ones).

Deep learning is a broad term that includes not only neural networks, but also hierarchical probabilistic models and other different unsupervised and supervised feature learning algorithms. The recent increase in interest in deep learning architectures is mostly due to their proven capacity to outperform prior state-of-the-art techniques on several tasks, facilitated by technological advances in computing power which have made deep neural networks faster and more accessible than ever.

### 1.4.1 Applications with biomedical data

Applications of deep learning approaches with biomedical data have been numerous and various: for instance, tasks such as integrative clustering [183], drug response prediction [184], cancer survival prediction [185], cancer subtyping [186] and psychiatric disease risk prediction [187] have been successfully tackled with deep neural networks.

One of the most common and successful deep learning architectures is autoencoders: they are artifi-

cial neural networks composed of two parts - an encoder and a decoder. The encoder tries to compress the information onto a lower-dimensional space in order to retain only the relevant features of the data, while the decoder controls this process indirectly. This is because the decoder has to reconstruct the input data starting from their compressed representation, which is only possible when their important information has been retained. The primary benefit of this architecture is that it can extract meaningful characteristics and filter out irrelevant data throughout transmission, therefore it can be used as a preprocessing technique [180]. Additionally, certain types of autoencoders are generative models, i.e. they can be used to effectively reproduce the data distribution [188]. Autoencoders have been used successfully mainly in classification settings [189, 190], whereas some studies have started to use them for omics data integration [191]).

The most widely applied model for extracting information about targets (tumours, organs, or tissues) from medical images is instead convolutional neural networks [192]. The human visual brain has inspired the design of CNNs, with filters and layers that mimic the geometric properties that human image recognition system possesses. Recently, CNNs have been fused with comprehensive attention approaches to produce an attention-based CNN that improves performance and explainability during biomedical image segmentation tasks [193, 194, 195]. CNN with transfer learning approaches (e.g. RestNet-101, VGG16, and InceptionV3), have been also used to detect tumours from radiological and histopathological images [196, 197, 198]. In these cases, identifying the size and location of the tumour is crucial for developing any treatment plan, and the U-Net architecture has been effectively used to segment tumours and extract the region of interest from biomedical images [199, 200, 201].

Several deep learning models have also been recently implemented in order to simulate biological processes. DCell is a visible neural network (VNN) capable of simulating cellular growth by using prior gene ontology knowledge to investigate genotype-phenotype associations [202]. During the training phase, genes' perturbations propagate through the network, giving rise to functional changes that inform the phenotypic response predicted by the model. A similar phenotype prediction model has been proposed by Guo *et al.* (DeepMetabolism), where an autoencoder-based neural network method integrates unsupervised pre-training with supervised training to predict phenotypic outcomes [203]. The connections between the layers of the network were used to model the relation between gene expression and phenotype, and were regulated using biological prior knowledge to reduce the risk of overfitting and increase training speed. Finally, deep neural networks have also been merged with differential search algorithms for gene deletion interventions in *E.coli* for the production of xylitol [204]. Similarly, convolutional and recursive neural networks have been merged to train prokaryotic models on ribosomal profiling data and binding site patterns for more precise annotations of open reading frames and translation initiation sites [205]. Neural networks have been adopted in Chapters 3-4 of this thesis.

## 1.5 Multimodal Machine Learning

Multimodal (or multi-view) machine learning is a branch of machine learning that combines multiple facets (modalities/views) of the same entity in a single setting, in an attempt to offset their limitations when used in isolation [206, 51]. In many fields, especially in biomedicine, the vast abundance of available data could not be exploited previously due to computational limits, with the key problem being how to extract desirable knowledge from large, complex and heterogeneous datasets. Manual analysis of data is considered to be difficult, largely ineffective and inefficient even with the support of statistical methods, therefore the challenge of managing and integrating large, multi-dimensional datasets is still an open problem, and new analytical tools are required to utilise these data to their full potential [207, 208]. A possible answer to this is the development of multimodal machine learning approaches, which focus on trying to link several datasets describing the same samples, i.e. different aspects of the same entity, for predictive purposes, with the expectation that these can provide more complementary information and thus improve the performance on the task.

In general all integration strategies can be traced back to three types of approaches: early integration, intermediate integration, and late integration, and can be adopted to merge heterogeneous data and develop predictive models. In early integration, the modalities are merged as a single data vector before being fed to the model, whereas in intermediate integration they are all processed simultaneously, and the new features are then merged and fed to the subsequent predictive model. Finally, in late integration, each modality is analysed independently first, and the results are then combined to get the final consensus results [209]. Multimodal approaches could prove to be an effective strategy when dealing with multi-omic datasets, as all types of omic data are interconnected. Data values may be directly concatenated as single sample matrices into one large matrix, transformed into a common intermediate format, or analysed separately with multiple models with different training sets for each data type. These integrative approaches have been investigated in the case studies presented in this work: in particular, Chapter 3 presents a case of early integration, while Chapter 4 investigates a late integration approach.

The stage at which data integration is carried out must be carefully considered, as this may have an impact on the transformation of data: for instance, the initial concatenation of all samples (early integration) usually results in increased noise unless regularisation is performed (which we have extensively explored in Chapter 3). Therefore, it is often preferable to build a similarity matrix between data types (intermediate integration) or analyse each data type separately (late integration) prior to the application of machine learning techniques [210, 60].

Table 1.1 contains the most recent multimodal machine learning and deep learning approaches with biomedical applications. For each study, several details are reported, including the purpose of the investigation, the type of data used in the analysis, the disease being scrutinised, and the link

to the source code (where available). In the following we will describe the application of multimodal approaches to omics data and to fluxomics in particular. Multi-omics models and the use of GSMM-generated data in machine learning models have been extensively investigated in this work, and various case studies are reported as chapters in the rest of this thesis.

Table 1.1: **Multimodal machine learning and deep learning approaches for biomedical applications**

| Ref | Year | Prediction task | Omic | Disease | Purpose of study | Deep-Learning | Source Code |
|---|---|---|---|---|---|---|---|
| [211] | 2022 | Regression | T, G, C | UCC | Drug response | No | |
| [212] | 2022 | Classification | T, C | AD, PC | Diagnostic classification | No | github.com/dmcb-gist/MOMA |
| [213] | 2022 | Clustering | G, T, E | PC | Cancer samples cluster | No | |
| [214] | 2022 | Classification | G, T, C | BC, KC | Subgroup identification | Yes | github.com/Lifoof/MoGCN |
| [32] | 2022 | Classification | I, C | BC | Cancer detection | Yes | github.com/bensteven2/HE_breast_recurrence |
| [215] | 2022 | Regression | T, G | GC | Survival prediction | Yes | github.com/huyy96/ RDFS |
| [216] | 2022 | Classification | T, P, M | COV19 | Patient outcome prediction | No | |
| [217] | 2022 | Regression | T, M | BC | Subgroup identification | No | |
| [218] | 2021 | Classification | T, M, C | PC | Biomarker discovery | No | github.com/kemplab/ML-radiation |
| [219] | 2021 | Classification | G, T | PC | Subgroup identification | Yes | github.com/NabaviLab/ GCN-on-Molecular-Subtype |
| [220] | 2021 | Classification | T, E, C | PC | Subgroup identification | Yes | github.com/SomayahAlbaradei/ MetaCancer |
| [99] | 2021 | Classification | G, R, C | AD | Subgroup identification | Yes | |
| [28] | 2021 | Classification | T, R | OPSCC | Survival prediction | Yes | |
| [221] | 2021 | Classification | G, R | AD | Subgroup identification | Yes | |
| [222] | 2021 | Classification | T, G, E | BC | Drug repositioning, target gene prediction, cancer subtypes prediction | Yes | autogenome.com.cn/ AutoOmics/AutoOmics.html |
| [33] | 2021 | Regression | T, G, E, C, I | PC | Survival prediction | Yes | github.com/luisvalesilva/ multisurv |
| [223] | 2021 | Regression | T, G, E | BLD | Survival prediction | Yes | |
| [224] | 2021 | Classification | G, T | CLL, ILD, PC | Cancer detection | Yes | github.com/duttaprat/ DeePROG |
| [225] | 2021 | Regression | T, E, G, P | PC | Survival prediction | Yes | github.com/wangyuanhao/ DeFusion |

*Continued on next page*

Table 1.1 – *Continued from previous page*

| Ref | Year | Prediction task | Omic | Disease | Purpose of study | Deep-Learning | Source Code |
|---|---|---|---|---|---|---|---|
| [226] | 2021 | Classification, regression | T, E, C | PC | DR, tumour classification, survival prediction | Yes | github.com/zhangxiaoyu11/OmiEmbed/ |
| [227] | 2021 | Classification, regression | G, T, E, C | BC | Subgroup identification, Survival prediction | No | |
| [34] | 2021 | Classification | I, E, C, T | SC | Subgroup identification | Yes | sys-med.de/en/ |
| [228] | 2020 | Classification | R, I | BT | Brain metastasis detection | Yes | |
| [229] | 2020 | Classification | T, E | AD, BT, KC, BIC | Biomaker discovery, Subgroup identification | Yes | github.com/txWang/MOGONET |
| [230] | 2019 | Classification | G, I | AD | Subgroup identification | Yes | |
| [231] | 2019 | Classification | T, P, E, C | BUC, BLGG | Treatment outcome prediction | Yes | github.com/BeautyOfWeb/Multiview-AutoEncoder |
| [232] | 2019 | Regression | G, C | GBM | Survival prediction | Yes | github.com/DataX-JieHao/PAGE-Net |
| [233] | 2019 | Regression | T, R, C, I | CC | Survival prediction | No | |
| [185] | 2019 | Regression | T, C, I | PC | Survival prediction | Yes | github.com/gevaertlab/ MultimodalPrognosis |
| [184] | 2019 | Classification | G, T | PC | Drug response | Yes | github.com/hosseinshn/ MOLI |
| [234] | 2019 | Regression | T, G, C | BC | Survival prediction | Yes | github.com/huangzhii/SALMON/ |
| [201] | 2018 | Segmentation | T, R, I | LC | Tumour Detection, Segmentation | Yes | |
| [100] | 2018 | Regression | T, E | LC | Subgroup identification | Yes | |
| [235] | 2017 | Regression | T, P, C, I | LC | Pathways analysis, Survival prediction | No | |
| [236] | 2016 | Regression | G, I | LC | Survival prediction | No | |

The table contains the most recent studies that present multimodal machine learning applications on biomedical data. For each study, the following details are reported: reference; year of publication; type of prediction task; data modalities used in the study (C: Clinical; E: Epigenomics; G: Genomics; I: Imaging; M: Metabolomics; P: Proteomics; R: Radiomics; T: Transcriptomics); disease investigated; the purpose/focus of the study; whether the study uses a deep learning approach; link to the source code (where available). The abbreviations used to identify the disease types are listed below. AD: Alzheimer's disease, BC: breast cancer, BIC: breast invasive carcinoma, BLC: bladder cancer, BLGG: brain lower grade glioma, BUG: bladder urothelial carcinoma, BT: brain tumour, CC: colorectal cancer, CITE-Seq: cellular indexing of transcriptomes and epitopes by sequencing, CLL: chronic lymphocytic leukaemia, COV19: Covid-19, GC: gastric cancer, ILD: interstitial lung disease, KC: kidney cancer, LC: Liver Cancer, OPSCC: oropharyngeal squamous cell carcinoma, OSC: ovarian serous cystadenocarcinoma, PC: prostate cancer, SC: somatic comorbidity, UCC: Ulcerative colitis and Crohn.

### 1.5.1 Multi-omic machine learning

With the technological advancements that have characterised omics data collection in recent years, the availability of high-throughput biological data has rapidly increased. Specifically, multi-omic data, including transcriptomics, proteomics, and metabolomics, are now easily accessible and of potential use for diagnostic and prognostic prediction. Since each of these data types is inextricably linked with the others in a coherent, mechanical way, it is of paramount importance that they can be combined in order to achieve a better understanding of the entire biological system. The more data types included in models, the more information is available to trace molecular components across multiple functional states [58]. However, integration of biomedical data is challenging, which is why specific machine learning models have been designed to achieve accurate predictions by integrating data obtained from multiple modalities [237]. Each of the modalities in a multimodal system is expected to add value by contributing information that is not present in any of the other modalities, and by highlighting possible interplays between them, consequently improving the model's accuracy. For instance, multimodal deep learning has been used in structural and functional neuroimaging and with metabolic modelling, in the attempt to provide a more comprehensive mechanistic understanding of the brain and its disorders [238, 239].

Figure 1.1 provides an outline of how multimodal raw data can be used as input for prediction purposes. First, multimodal raw data is collected (panel A); this can include clinical data, multi-omic data, radiological, and histopathological images. Once the data have been collected, preprocessing steps are applied to structure the data in a format suitable for machine learning applications (panel B); this process could include dealing with missing values (either by removing them or using imputation techniques), encoding text-based variables into numeric values, or normalising the data distributions. Part of the preprocessing phase consists in selecting a subset of suitable features to use as final input for the machine learning procedure, by retaining the meaningful ones and discarding the ones that might add noise to the data, as already explained in Subsection 1.3.1. This process is called "feature selection and engineering", and is reported in the figure as a separate step for clarity (panel C). The main goal of this phase is to select, manipulate, and transform the data into features that can be used in the next stages. The last step of the pipeline includes the model training and evaluation phases (panel D). Once one or more suitable machine learning models have been selected, a training-testing procedure that estimates, given the input features, the performance of the models on a test dataset can be used. It is fundamental that the procedure be robust and not include any statistical bias, whether be it given by the datasets or by the procedure itself. In some cases, some additional parameters (hyper-parameters) need to be estimated, which further complicates the training. A possible solution to this is the use of $k$-fold nested cross-validation, which can be employed to identify the optimal hyperparameters (inner loop) and evaluate the model performance (outer loop). During this process, the dataset is randomly split into $k$ folds, where $k-1$ folds are used to train the model and the left-out

fold is instead used to assess the performance of the model on unseen data (outer loop). This process is repeated $k$ times allowing each fold to be used as test set. The average performance over the $k$ test folds is then returned, together with other sample statistics. The optimal hyperparameters used to train the model in the outer loop are identified in the inner loop, where the $k-1$ training folds are randomly split into subfolds (training and validation folds). Cross-validation is then applied to train the model on the training folds and evaluate it on the validation fold using different combinations of hyperparameters. The optimal hyperparameters are then used in the outer loop for training and deploying the final predictive model. Values of $k$ like 5, 8 or 10 are commonly used to partition the dataset. At the end of this pipeline, the final best model (with optimal hyperparameters, assuming the stability and consistency of the training process) can be used to perform the predictive task on any new data sample. Cross-validation was used in all the case studies presented in this thesis (together with the rest of the described pipeline), and nested cross-validation was in particular adopted in Chapter 4. It is important to remark that the structure of this pipeline is the same for uni-modal machine learning tasks as well (the only differences would be the use of only one data source and the absence of any integration step, not explicitly reported above as it is dependent on the model adopted and the task).

**Examples of machine learning applications with multi-omics data** Machine learning-based integration techniques for multi-omic data have been implemented to identify and investigate disease states by condensing complementary information provided by the different omics, thus increasing the effectiveness of disease analysis and diagnosis [240, 241, 242, 243, 211, 212, 213, 214, 244, 245, 215, 246, 217, 222, 223, 224, 225, 226, 227, 234]. Wang *et al.* proposed a deep learning pipeline (MOGONET) that uses graph convolutional networks (GCN) to process similarity networks (DNA methylation, miRNA, and mRNA) whose outcomes are then combined in a cross-omics discovery tensor and fed to a view correlation discovery network [247]. This model can exploit the correlation among samples and their multi-omic data, while allowing biological interpretation of the results. GCNs have also been applied to extract local features from gene-gene and protein-protein interaction networks and gene-coexpression networks, where node features were RNA-seq expression and copy number variation data [219]. Global features from these networks were obtained by using a fully-connected network, and then concatenated into a single layer with softmax activation function for classification purposes. Convolutional variational autoencoders (CVAEs) have also been adopted to integrate miRNA, mRNA, DNA methylation, and clinical pan-cancer data to extract features (later inputted into a deep neural network) for predicting whether a tumour was primary or had metastasised [220]. The omics were concatenated before being fed to the CVAE. A similar multi-view factorisation autoencoder was implemented for feature learning and generation of patient representations, which allowed the construction of patient similarity networks using miRNA, gene and protein expressions, and DNA methylation data [231]. Domain knowledge was also injected into the model via feature interaction matrices.

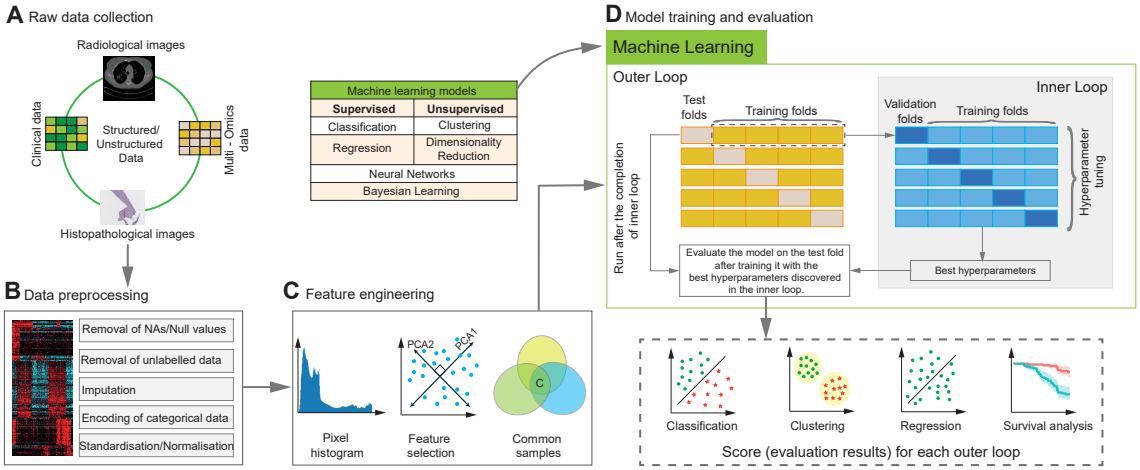



Figure 1.1: **General machine learning pipeline for multimodal biomedical data applications.** (A) Structured and unstructured raw data are collected and (B) preprocessed by removing NAs/null values and applying normalisation techniques. (C) Feature selection is then performed to extract the region of interest from images or to reduce the high dimensionality of multi-omic data. (D) A multimodal machine learning architecture can be trained and validated using a nested cross-validation approach. The inner loop is used for tuning the hyperparameters, while the outer loop is used for model evaluation. The final model can be used for regression or classification purposes (e.g. predicting patient-specific survival probability or classifying patients into risk groups). The predictions made by the machine learning algorithms can be further analysed by clinicians and researchers with computational techniques (e.g. using interpretability approaches) to inform therapeutic interventions and contribute to the development of personalised medicine [1].

**Examples of machine learning applications with multi-omics and imaging data** Following recent advances in computer-aided diagnosis [248], much progress has been made towards developing more efficient forms of integration for multi-omic and imaging data. Specifically, the most recently developed integration methods are based on the intuition that information about other omics can be inferred from biomedical images [249]. In order to improve the predictive ability of image-based deep learning models, new deep learning architectures for the integration of multi-omics, clinical, and image data have been proposed [250]. For example, bioimaging data have been integrated with clinical data and patient history of treatments to generate a comprehensive view of patient health in [251]. Complex tasks such as survival prediction have also been successfully addressed by integrating images and omic data [236]. In [233], the relationship between radiomic features and gene expression data in colorectal cancer was investigated by using CT images to extract radiomic features from the tumour area (i.e. intensity, shape, and texture). The extracted radiomic features were then integrated with clinical, histopathological, and genomic data for survival prediction. Instead, the model

PAGE-Net was proposed in [232] to integrate histopathological images, gene expression data, and age of glioblastoma multiforme patients. Several multimodal neural network-based models have also been implemented to predict patient survival using a combination of clinical data, mRNA expression data, microRNA expression data, and histopathology whole slide images (WSIs), as in [185]. Classical machine learning approaches have too been used in a multimodal setting with multi-omics and imaging data. For example, genomic, transcriptomic, proteomic, and histopathological features from tumour samples have been integrated through a random forest model to identify molecular pathways associated with histopathological patterns [235]. In [252], instead, a sparse linear regression approach (Lasso) was applied to identify the highly correlated image features and metagenes (i.e. aggregated patterns of co-expressed genes) for survival prediction, based on the assumption that the biological processes associated with morphological changes can provide insights on the molecular mechanisms of many cancers. Finally, in another study, an ensemble learning strategy based on Gaussian kernels was proposed to classify healthy controls/Alzheimer's patients by integrating MRI and PET images, biomarkers in cerebrospinal fluid, and age information [253]. Linear discriminant analysis was used separately on all omics for feature extraction, whereas ensemble models were used to discriminate among the classes in a decision tree-like approach.

### 1.5.2 GSMMs and machine learning

As we have seen so far, machine learning is a useful tool that can be used to deconstruct biological complexity and extract relevant outputs for clinical biomedicine when dealing with the high volume and heterogeneity typical of modern multi-omic data. In particular, machine learning algorithms can be leveraged to reduce the dimensionality of the data and elucidate cross-omics relationships. However, this very often translates to models enjoying high computational predictive power at the expense of interpretability. Although some steps have been taken towards the development of an interpretative framework [254, 255], a definitive approach has not yet taken hold. The complementary characteristics of GSMMs and machine learning and their common mathematical bases make them suitable to be used in combination to solve this issue. As the features introduced by GSMMs are fully informative, given that they provide biological information in terms of stoichiometry and genetic control of the biochemical reactions, a combined approach can help to address the lack of interpretability associated with machine learning models in biology.

In the context of this work, constraint-based modelling is ideally poised to bridge the gap between biological datasets and biologically-agnostic machine learning models, thus consolidating these two computational frameworks in order to reveal novel insights relating to the biology of metabolism. Together, machine learning algorithms and constraint-based models can improve omics-fuelled statistical and machine learning analyses by supplementing the learning process with biological knowledge and refining phenotypic predictions. Moreover, machine learning has proved to be able to improve

metabolic models reconstructions [256].

Several studies have shown that supervised classification models can benefit from the integration with information generated by metabolic models [257, 216, 258]. For example, Chien *et al.* reconstructed the GSMMs of 21 *Pseudomonas* species living in the endosphere and in the rhizosphere, and 12 media formulations were simulated to predict and classify bacterial ecological niches [259]. SVM, artificial neural networks, and non-negative matrix factorisation were cross-compared, and it was found that SVM was the most effective model in capturing the ecological niche of these bacteria. Importantly, metabolic features were more predictive than purely genomic features. The integration of metabolic models and machine learning within a classification framework has also been shown to correctly identify side-effects of inhibitory drugs [260]. Drug-specific actions were simulated by *in-silico* gene deletions, which were propagated to metabolic reaction bounds via logical rules. Flux Variability Analysis was then used to estimate the metabolic fluxes that were then fed to an SVM model showing improved results when compared with traditional predictive models. Taking these results as a lesson, in Chapter 4, we have used SVM with FVA-generated metabolic fluxes in a classification setting for the prediction of liver cancer in children.

Elastic Net regression was applied to optimise a metabolic model of *Pseudomonas putida* with the aim of maximising rhamnolipid production while Ajjolli *et al.*, instead, built an artificial neural network to estimate fluxes using enzyme concentrations for the upper part of the glycolytic pathway as input [147, 261].

The relation between bacterial central metabolism and metabolism-affecting factors such as carbon sources, oxygen condition, and genetic background, was studied in [262] with the integration of metabolic fluxes facilitated by three machine learning models (i.e. SVM, kNN, and decision tree). Roy *et al.*, instead, combined existing tools including Omics Mock Generator (OMG), Inventory of Composable Elements, Experiment Data Depot, and the Automated Recommendation Tool (ART) to store, visualise, and leverage multi-omic data to predict bioengineering outcomes [263]. The fluxes in the OMG model were computed by using FBA with growth rate maximisation, and were then integrated into the machine learning model to predict isoprenol production. With the integration of ART, the proposed approach reported an increase in biofuel production of 23%.

Genetic and population-based algorithms have been successfully merged with metabolic modelling to identify optimal strategies for metabolic engineering when taking into account multiple cellular objectives simultaneously [264, 265], and non-trivial multiple gene knockouts affecting cell growth [266, 267, 268]. Cavill *et al.*, instead, showed how using GAs with metabolic data could improve classification performance by over 9% whilst also halving computation time [269].

Agglomerative hierarchical clustering (AHC) and k-means clustering have been used on transcriptomic data and fluxomic profiles in order to characterise the ageing process in CD4 T-cells [270].

Principal Component Analysis, which is commonly used to reduce dimensionality in large datasets by identifying a small number of dimensions (i.e. principal components) that can be used to perform a change of basis on the data, has also been used numerous times for clustering. For example, Jalili *et al.* used PCA and random forest to cluster the flux states characterising different cancer types and reveal reactions containing the most relevant information [271]. In [272, 273], PCA was instead integrated with k-means to combine GSMM-derived data and transcriptomics and elucidate key mechanisms used by *Cyanobacteria* that could have not been detected using transcriptomics alone.

### 1.5.3 GSMMs and multimodal approaches

Original attempts at integrating genome-scale metabolic models and machine learning algorithms have recently taken the lead especially in the stream of multimodal approaches that are now being used with omic data. In the field of biomedicine, this has translated so far in the use of different omics, fluxomics included, as different modalities (or views) to potentially highlight distinct biological mechanisms that might go unnoticed when not using the metabolic information provided by GSMMs. Indeed, integrating mathematical models or biological networks with additional data has been recognised as an important tool for gaining novel insights from large amounts of biological data [274, 275], and has shown performance improvements compared to omics-only settings [255, 261, 147, 270, 276, 272, 257, 268, 277]. We explored this approach in the thesis, and in particular in Chapters 3-4.

Figure 1.2 presents the general pipeline for the integration of multimodal imaging, omics and GSMMs-generated data in a machine learning framework. Patient-specific multi-modal data are collected and pre-processed to be represented in a suitable format for machine learning applications. Some of the data modalities (e.g. transcriptomics and proteomics) can be also integrated into a GSMM to generate patient-specific fluxomic data. Once the data have been preprocessed, all the three different approaches (early, intermediate and late) can be applied for a multi-modal data integration through machine learning models [240, 250]. Finally, the machine learning results can be further analysed for therapy optimisation and to generate biological insights. Survival analysis, pathway analysis, or explainability techniques are usually applied to identify relevant features highly affecting the target variable [218, 278].

In [218], the metabolic fluxes of 915 TCGA cancer patients were generated using FBA and integrated with clinical, genomics, and transcriptomics data. The final dataset was fed to a gradient boosting machine classifier to predict the radiation response in individual patient tumours. The integration of genome-scale metabolic modelling and machine learning approaches showed an improvement in identifying new biomarkers and predicting tumour response to radiation. Lee *et al.* integrated immune cell transcriptomes with the Recon3D human metabolic model using iMAT [216]. FVA was performed to generate the metabolic fluxes, which were then integrated with plasma metabolites to

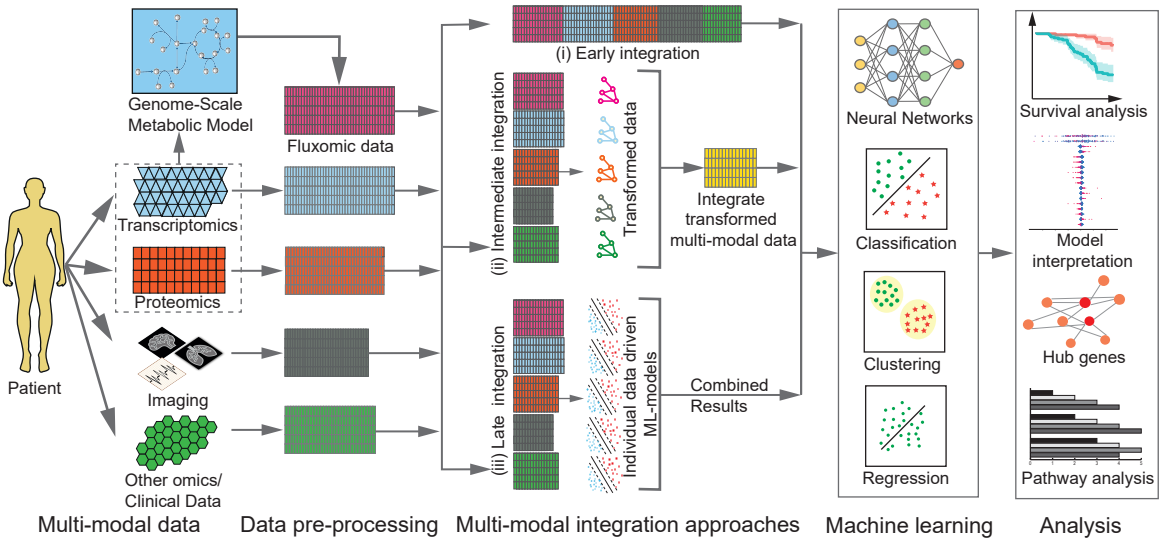



Figure 1.2: **Integrating multi-modal machine learning and GSMM.** Patient-specific multi-modal data (e.g. transcriptomics, proteomics, imaging, genomics, and clinical data) are collected and pre-processed. Simultaneously, a patient-specific GSMM can be employed to generate fluxomic data. The multi-modal pre-processed data and the fluxomic data can be used as input of a machine learning model for patient-specific diagnosis and prognosis, using the three different integration approaches that we have introduced at the beginning of this section: (i) early integration, where the preprocessed data modalities are concatenated before being used as input of the machine learning model; (ii) intermediate integration, where the individual modalities are first jointly transformed to reduce data dimensionality or extract meaningful features (e.g. using cross-modality approaches), and then integrated to be used as input of the machine learning model; (iii) late integration, where each data modality is used as input of an individual machine learning model, and the results from each model are then combined and further analysed (e.g. using ensemble approaches). The final model's results can be further analysed to generate biological insights and identify disease-related biomarkers through survival, pathway and enrichment analysis, and explainability approaches.

predict Covid-19 patient survival outcomes using random forest. The integration provided more accurate results than using metabolic fluxes alone for classifying disease severity and predicting clinical outcomes.

A multimodal deep learning approach that integrates gene expression profiles and metabolic models to predict the cell growth rate in *Saccharomyces cerevisiae* was proposed by Culley *et al.* [255]. The results indicated a noticeable improvement in performance after integrating fluxomic data from FBA, compared to the exclusive use of gene expression. We have built on this and conducted additional work described in Chapter 3, in which the generated flux rates were concatenated with the gene expression profiles and two neural network models were compared with a range of regularised linear models.

Another application of machine learning for genome-scale metabolic models was the generation of the context-specific metabolic profiles through the integration of gene expression measurements with drug-specific small intestine epithelial cell metabolic models [279]. In this instance, gene expression and flux measurements were used as features for a multi-label support vector machine to predict the occurrence of gastrointestinal side effects and to cluster drugs in order to reveal similarities that go beyond a chemical or pharmacological classification. Kim *et al.* [280] used instead recurrent neural networks (for the transcriptomic data), Lasso regression (for the proteomic data) and FBA to generate growth rate predictions for *Escherichia coli*, and obtain a weighted consensus result.

Sparse Lasso regularisation approaches have also been integrated with regression models to identify growth-boosting and limiting characteristics for *Synechococcus* [272], to train multimodal artificial neural networks using the metabolic modelling of *Saccharomyces cerevisiae* [281, 255, 282], and to investigate phenotypic extreme currents (ECs) based on a combination of metabolic network features and gene expression data [283].

Finally, GSMMs have also been integrated with Cox regression models to investigate cancer metabolism and provide personalised survival predictions and cancer development outcomes [278], while ensemble learning has been successfully applied to integrate clinical, gene expression, metabolic (from GSMMs), and mutation data, as in [218]. In particular, each omic type was used to train a different decision tree-based XGBoost model (base learner), whose predictions were then used by another decision-tree based XGBoost model to identify the optimal base learner for each sample. The final prediction was given by a linear combination of the predictions of the initial models.

Although results from integrating machine learning into biological models seem promising, there is still much leeway for improvements in terms of refining phenotypic predictions, and while several models have been proposed, integration remains challenging [60].With this in mind, we argue that GSMMs can be used both as a foundation for the integration of multi-omic data originating from different domains and as a source of interpretable features for machine learning algorithms. Cuperlovic [284] summarised the main prerequisites essential for the successful implementation of machine learning as

follows: the proper selection of learning attributes, construction of training and test sets, selection of the appropriate learning algorithm(s), careful design of the learning approach and an accurate evaluation of predictive performance. Consequently, it is important to consider how these key attributes can be adhered to when considering the application of machine learning to GSMMs. In the biomedical setting, understanding the biological phenomenon is almost always necessary. However, machine learning algorithms suffer from a lack of biological interpretability in spite of their optimal predictive power. It is important to note that this does not only depend on the data-driven models but extends to the preprocessing of their inputs as well. This is critical to keep in mind, since Wolpert has demonstrated that the performance of machine learning algorithms on a specific task depends heavily on the available data (and therefore on the preprocessing as well) associated [285], which means that when they are used with biomedical data, the preprocessing choices need to be made not only considering the final predictive performance, but also potential interpretability issues. For this reason, we envisage the ideal framework as one that can leverage the quantitative power of machine learning with the biological interpretability provided by the mechanistic GSMMs. Overall, the success of biomedical research is dependent on its capacity to deal with the sheer volume of multi-omic data currently generated. This framework could aid in the elusive goal of creating personalised medicine and more resilient healthcare systems, and it is also the focus of the work of this thesis.

## 1.6 Current issues

In the previous sections, we have introduced and described biomedical data, genome-scale metabolic models, machine learning and multimodal integrative approaches. However, each of these aspects presents significant challenges that need to be tackled in order to achieve success in the quest for better healthcare systems. In the following paragraphs, we will present these challenges and possible solutions, referring when possible to how we solved them in this work.

**Biomedical data management** The increasing numbers in structured and unstructured biomedical data collected are paving the way to a future where medical treatments can be tailored to the individual patient, thus streamlining current national healthcare systems and leading to a future of personalised healthcare. The huge volume of data is contributing to increasing the understanding required to extract knowledge from different sources and create the 'golden triangle' of treatment (i.e. the right target, the right chemistry, and the right patient), with the overarching goal of advancing research and continually improving the quality of patient care whilst reducing costs. However, in order to exploit these high volumes of data in an AI-assisted healthcare, better computational infrastructures, new data interpretation methods, and unique collaborative approaches are required, meaning that several challenges need to be overcome [286]. For example, clinical data are available in heterogeneous

formats but the wider computational infrastructure to generate, maintain, transfer, and analyse these data securely (and correspondingly, to integrate clinical data with omics data) is lacking, also because the cost of data generation is usually much lower than that of data analysis and storage. Another challenge concerns the data transfer between different locations. Emailing data related to individual patients or sharing information on a single cloud environment undoubtedly raises concerns about the security and privacy of individuals both before and after the data transfer [287]. Several solutions have been proposed to create better security systems that include encryption methods and de-identification algorithms. For example, distributed processing can be implemented where multiple computers that undertake different parts of the same task in different locations are connected via a communication network under the control of a central server [288]. Another major development is the emergence of federated learning, a type of decentralised modelling format where a shared global model is trained within a central server while training and combining individual local models and keeping all the sensitive data in local institutions close to where they originated [289]. Such a format holds great promise for the healthcare sector where much of the data are highly fragmented and sensitive. However, model aggregation methods within federated learning pipelines must still be carefully evaluated to preserve robustness and privacy that may be compromised by biased local datasets, faulty clients, and cyber attacks following local updates [290].

**Generation and processing of fluxomic data** Genome-scale metabolic models are one of the main frameworks striving to bridge the gap between genotype and phenotype by incorporating prior biological knowledge into mechanistic models. Nevertheless, they require additional experimental measurements to refine parameters within feasible limits and increase predictive performance [264]. When developing computational techniques for biomedical purposes, achieving an accurate and robust prediction is essential. However, to date, GSMMs and multi-omic integration studies have relied on the generation of personalised constraint-based fluxomic data through FBA, which not always provides a unique solution. Indeed, different solvers can produce slightly different results when solving linear problems using FBA, which might lead to serious issues when using the full flux distribution for further analysis, or as a part of a prediction algorithm, when not taken into account. Moreover, in metabolic modelling, choosing a suitable objective function remains elusive. In systems biology, it is often assumed that cells optimise their metabolic networks to maximise their growth rate (biomass). However, as we wrote in Section 1.2, the cellular objective might vary depending on the type of tissue, the species, or even throughout time. As a result, optimisation techniques that consider multiple objectives should be preferred, when computationally feasible. Furthermore, FBA may contain thermodynamically infeasible loops in the network. Flux Variability Analysis can overcome these limitations by providing a flux range (i.e. a range of activity for all reactions) that could better represent the metabolic potential of the network [50, 53]. In particular, in the case studies presented in this thesis, we have always adopted a variant of FBA to deal with these issues, and have used specifically FVA

in the work presented in Chapter 4. Moreover, the large size and complexity of GSMMs makes it difficult to analyse their output without losing sight of the big metabolic picture. A solution to this could consists in the generation of a smaller metabolic network. For example, Erdrich *et al.* proposed an algorithm that applies a pruning step that removes iteratively the reactions with the smallest flux range until no further reactions can be deleted without eliminating "protected" parts of the network [291]. This algorithm is thereby able to capture fundamental characteristics of the central metabolism or other metabolic modules of interest and perform a fast, unbiased, and exact network reduction, which consequently reduces the number of metabolic features, making further analyses more feasible. In the case study described in Chapter 3 we used parsimonious FBA (pFBA) with a similar objective of "simplifying" the metabolic network, simulating at the same time the evolutionary tendency of optimising (i.e. reducing) enzyme usage with equal cell growth [292].

An accurate prediction is also the result of clean, experimentally valid data. In addition to considering the quality of data when generating replicates across different omics, a balance needs to be struck between achieving sufficient statistical power and coverage of biological variability whilst reducing the batch effect [293]. Moreover, combining omic datasets increases the risk of introducing redundancies across different types of data, which must be resolved in order to prevent generalisation issues on other data samples. This is particularly applicable in the case of fluxomic data, which are derived as a direct result of other omics incorporated within metabolic models. To tackle this issue, in each of the following chapters we have conducted different preprocessing steps depending on the data and the task in question.

As a consequence, we can easily state that the development of new generative techniques or regularisation approaches specifically for flux outputs is necessary. Although all omics are, in theory, directly interpretable from a biological perspective, their integration within machine learning tasks remains limited by the interpretability of the results generated from the most complex techniques, such as neural networks, which act as "interpretability bottleneck". This calls for the development of better, more interpretable models tailored to the analysis of multi-omic data and GSMM-generated fluxomic data. Regarding GSMMs, the prospective research directions that could be followed include the building and testing of new mathematical models that would, for instance, allow the prediction of the behaviour of different types of cancer under diverse environmental conditions in a parametric way. Further studies could also devise new ways to incorporate signalling and regulatory networks (something that we attempted to do in Chapter 2), thus introducing additional information in relation to factors that are not included in the human genome but still affect cellular functions.

**Multi-omic data challenges** The integration of different omic datasets reserves many challenges which have been solved only partially to date. The most frequent obstacle when working with multi-omic data is related to the cross-omics interactions and how these can be incorporated into a model.

Other challenges can instead derive from the sample size, which might not be large enough to provide statistical power [294, 295]. Moreover, to estimate and reduce the presence of potential confounders and biases, cross-validation-like approaches need to be adopted, which increases the computational power and time required in the model training phase [295]. Finally, different omics might have different sample distributions, meaning that different normalisation and preprocessing techniques have to be considered or even designed [296, 297, 298]. This has to be coupled with the high dimensionality of the data, their sparsity, and the missingness of specific omic types for some samples in the dataset [299, 300]. The problem of the high-dimensionality of the data (known among machine learning practitioners as the "curse of dimensionality"), when working with omics, is further exacerbated by the connections between different omics, which may prevent a naive use of dimensionality reduction techniques. Some approaches have been developed to overcome this issue and improve the models accuracy [301, 132], and have been used accordingly in the work presented in the rest of this thesis.

The unavailability of the same number of data modalities for all the samples is even more common in this field than in others because of production costs, privacy, ethics regulations and datasets sizes [302]. For example, due to stochastic gene expression and technical noise, the missing value rate in RNA-seq single-cell dataset can be usually observed at approximately 30% [303]. Proteomics datasets also show similar issues, since missing values account for a substantial amount of experimentally-acquired data. Statistical techniques such as k-nearest neighbour and mean value imputation are usually applied to impute missing data, however, these techniques generate data from the known statistical distribution of the provided samples and can only be utilised for the same data modality [28]. Recently, deep learning methods such as generative adversarial networks and autoencoders have been applied to handle missing data in imaging modalities, while transfer learning has been used to impute missing gene expression data from DNA methylation [304]. An attempt to solve this issue was made by Zhou *et al.* who adopted a three-stage deep neural network that integrated different combinations of multimodal data (i.e. MRI, PET images, and single nucleotide polymorphism data) while using the maximum number of training samples available for each omic combination [230]. A similar problem was addressed in [228], where a clustering-like approach was applied directly in the learned subspace to obtain compactness and separability of the different classes, for samples with MRI and PET images. Finally, in order to maximise the number of available instances in raw omic and neuroimaging data, an integrative method based on linear interpolation to fill the missing attributes for each incomplete instance was proposed in [221]. An analogous approach based on an incomplete multi-modality data fusion technique that utilises the consistency between modalities has also been applied for multimodal brain image fusion [305]. Although the reported approaches have been shown to handle incompleteness issues in multimodal machine learning, more research is required to identify an effective and general approach to deal with missing modalities when integrating imaging, multi-omics, and radiogenomics data. The problem of cross-omics interactions, instead, has been recently addressed by integrating feed-forward neural networks and tensor factorisation decoders [241]. In our

work, we have mainly used datasets with complete multimodal samples. When this was not possible, as in the case study presented in Chapter 4, we imputed the missing data or directly removed the missing modality depending on the extent of the problem.

**Reproducibility** When integrating multimodal data (including imaging data) and GSMMs, benchmarking and reproducibility are other significant challenges. Many studies reported difficulties in reproducing published AI biomedical results since these works failed to provide test data, sufficiently documented methods, or source code [306, 307, 308]. Moreover, published code can sometimes lack sufficient description or miss internal dependencies to reproduce the results [307], which might significantly reduce the ability to validate and improve the study. In order to address this problem, several guidelines and methodologies for scientific reproducibility and reporting have been proposed in the biomedical research field. Particularly, the FAIR (findability, accessibility, interoperability, and reusability) standard [309] is the principle endorsed by global organisations and could be leveraged to address the reproducibility issues. Data sharing, code sharing, workflow sharing, and environment sharing are approaches that could contribute to FAIR research. For example, researchers are encouraged to use public repositories such as GitHub, GitLab, and Bitbucket to enable FAIR sharing of code. Besides, computational reproducibility can also be handled by applying container technology tools such as Singularity or Docker (a container is a method to package an application and allow it to run together with its data and code dependencies) [310]. Although the adoption of such technology for GSMMs is not widespread yet, these approaches are promising for integrating more comprehensive bioinformatics pipelines and developing multimodal architectures for GSMMs. All the case studies presented in this thesis have indeed public repositories and code for full reproducibility and transparency, as reported in Data and code availability.

## 1.7 Related work and final remarks

Part of this chapter has been published as a tutorial in Computational Systems Biology in Medicine and Biotechnology, Springer [311]. I was personally responsible for the sections about multi-omic data integration and machine learning (text, code and figures). A good majority of the content of this chapter will be also part of a review published in the foreseeable future. As a joint first-author, I prepared the written manuscript in collaboration with the co-authors. In particular, I wrote the sections related to multi-omics integration, deep learning and multimodal machine learning. In the context of this work, the original content has been re-organised and complemented with additional sections and paragraphs to provide a more comprehensive introduction to the subject of the thesis.

## 1.8 Aims of the thesis

In this chapter of the dissertation, a comprehensive literature review was provided, which lists the main entities and actors in the field. The following chapters represent different case studies and will take the reader through a journey of investigation of machine learning-based integrative approaches for constraint-based metabolic modelling and omics data.

In particular, the main aim of this thesis is to explore how the integration of multi-omic data through the application of machine learning techniques can greatly enrich the scope of traditional constraint-based metabolic modelling.

We will show that:

- GSMM-generated metabolic fluxes contain information which is different than the information present in gene expression data (Chapters 2-3-4)
- it is possible to successfully extract such information from GSMMs with machine learning approaches (Chapters 2-3-4)
- GSMM-generated metabolic fluxes can be combined (integrated) with gene expression data to improve machine learning models' predictions and results' interpretability (Chapters 3-4)
- The quality of such integration depends heavily on the model used and the task to perform (Chapter 3)
- In this scenario, simpler machine learning models can perform as well as more complex ones, with the merit of being easier to interpret (Chapters 3-4)
- Different combinations of omics data influence models' performance to different extents, depending on the samples' clinical characteristics (Chapter 4)

All of these chapters have been published as separate papers and hope to provide the reader with a coherent exploration of the possibilities that GSMMs promise in such a rich and complex field of machine learning-led computational biology, with a particular focus on how these technologies can pave the way to more standard approaches in personalised medicine.

# Chapter 2

# Using GSMM-generated fluxes with transfer learning for gene regulatory network reconstruction

## 2.1 Introduction

The aim of this thesis is to investigate the use of constraint-based metabolic modelling in the context of precision medicine, focusing in particular on machine learning techniques for the integration of multi-omic data and their leverage. We should start this journey by asking:

- Do GSMM-generated metabolic fluxes contain different information than experimentally-measured gene expression data?
- Is it possible to extract such information? Under what conditions?
- How can machine learning be leveraged to achieve this?

These questions are very general but essential to answer, since the rest of this work will depend on them. In this chapter, we will try to answer them by looking at a case study, it being understood that the results, even if our experimental framework is quite specific, will be valid in other settings as well. This is true for the results of the other chapters too: we will demonstrate that, under certain conditions, integration of genome-scale metabolic models and other omic data can take forward precision medicine approaches and make their use more widespread.

However, in this chapter, we will not discuss a precision medicine application in the classical way that precision medicine is understood. Usually, when using the term "precision medicine", one alludes to approaches that provide tailored solutions to different patients. Due to the nature of the experiments we will, instead, present an approach in which the "patients" are distinct genetic mutations of the same individual. In the remaining chapters, we will show precision medicine examples in the commonly accepted meaning.

The contributions of this chapter are the following: we demonstrate that GSMM-generated metabolic fluxes contain different information than gene expression data, and introduce a machine learning framework within which this information can be extracted and exploited. In the next chapters we will therefore use this newly acquired knowledge to investigate whether the information held by metabolic fluxes can be employed in conjunction with the information present in gene expression data (as opposed to being used by itself, as in this chapter) to improve the performance of machine learning models in a precision medicine setting.

The case study presented is a gene regulatory network reconstruction task, in a transfer learning scenario. The experiments presented were performed in collaboration with multiple co-authors. In particular, I performed all the analyses relating to the generation of the metabolic features and the analysis and interpretation of the results, as well as contributing to the data preprocessing of the gene expression data and the metabolic fluxes. The co-authors downloaded the gene expression data, conducted an initial preprocessing, and trained and tested the machine learning models (each of these steps is described in detail in the chapter).

## 2.2 Background

Living organisms need, for their survival and replication, a gene regulatory system responsible for their maintenance, development and response to changing environmental conditions. Gene regulation is orchestrated by large sets of regulator molecules with specific targets, which collectively form a gene regulatory network (GRN). The mapping of GRNs was recently propelled by the surge of high-throughput data, that led to both the discovery of unknown biological interactions and a deeper understanding of known structures [312, 313, 314], thus benefiting basic biology but also related disciplines, such as biomedicine and biotechnology [315]. Given its importance, the reconstruction of the network of regulatory mechanisms existing in the human body is key to elucidating pathogenic processes and identifying molecular drug targets.

Several computational methods for GRN reconstruction have been proposed in the literature, including graphical Gaussian models [316], Bayesian networks [317], as well as approaches that consider and exploit causality phenomena [318, 319] or knowledge derived from related organisms [320]. To pre-

dict unknown relationships, GRNs have also been mathematically integrated with metabolic networks, which mediate interactions between gene regulation and environmental cues [321].

Unlike other approaches, such as genome-wide association studies (GWAS), systems biology techniques can provide mechanistic information to exploit in the reconstruction of GRNs [322]. In particular, genome-scale metabolic models (which we have introduced in Section 1.2) are well-suited for integration with GRN networks, as they complement them [323, 324]. Indeed, as we have shown in Chapter 1, GSMMs allow the capturing of long-range phenomena on the scale of cellular systems thanks to the functional information they contain, which is encoded in their metabolic pathway and reaction representations [325]. *In silico* metabolic information has been adopted in reconstructing GRNs [315, 326], and to infer gene relationships by analysing the metabolic effect of simultaneous gene KO [327, 268]. However, metabolic network modelling has not been used so far to inform GRN inference methods in combination with transfer learning.

Following this line of research, in this chapter we investigate the potential of exploiting metabolic information while reconstructing the human GRN, in an integrated transfer learning framework. In particular, we reconstruct the human GRN by leveraging the knowledge about an additional model organism [328], i.e., the mouse, and exploit both a set of known/verified regulations as well as a large set of still unstudied gene regulations. The two considered organisms are linked by considering their orthologous genes, i.e., genes inherited in both species from a common ancestor gene. Such genes are integrated within a constraint-based model (the GSMM) that simulates their artificial knockout and determines how this perturbation propagates over the corresponding metabolic network, thus creating a precision medicine-like scenario in which the metabolic fluxes do not correspond to the metabolism of different individuals, but to the metabolism of the same individual with different gene mutations. This approach allows us to catch possible analogies between the two organisms in terms of their fluxes, in both known and still unknown regulations.

Our experimental evaluation, described in detail in Section 2.4, empirically proves the effectiveness of the proposed integrated approach. This approach resulted in increased accuracy of the reconstruction and the generation of mechanistic insights coming from the analysis of the most important metabolic features contributing to the GRN reconstruction, related to either the human or to the mouse organism. With this chapter, we hope to show the reader the promising potential of GSMM-generated metabolic fluxes in machine learning-based precision medicine applications. In the following chapters, we will take a further step in complexity by investigating their use in conjunction with gene expression data as well.

## 2.3 Materials and Methods

In this section, we first describe how we built the dataset under analysis, from the collection of the gene expression levels for both human and mouse genes to the construction of metabolic features. Then we provide the methodological details of the proposed transfer learning approach. A graphical overview of the proposed approach is shown in Figure 2.1.

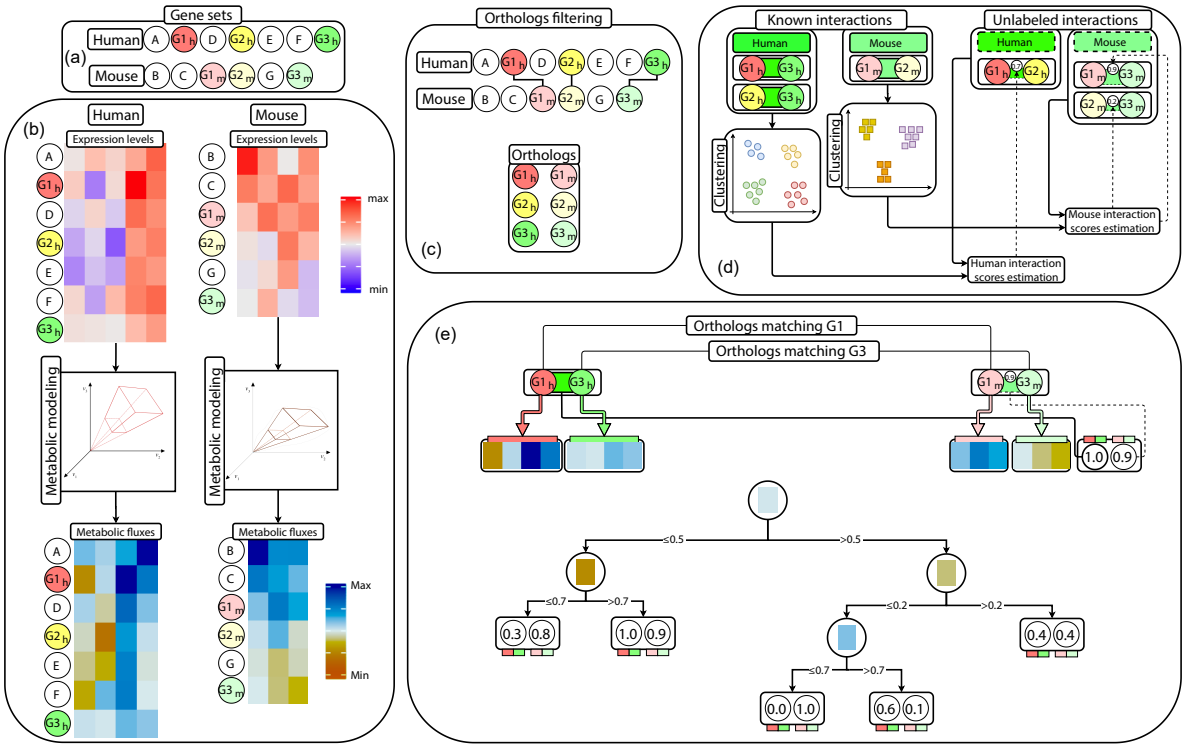



Figure 2.1: **Transfer learning pipeline.** Starting from the selected gene sets for the human and mouse organisms *(a)*, we compute metabolic fluxes from gene expression levels through genome-scale metabolic modelling of gene knockouts *(b)* using TRFBA. Genes are then filtered to consider only the subset of orthologous genes for human and mouse *(c)*. For both organisms, we estimate the confidence of existence on unlabelled (i.e. untested) interactions through a clustering-based procedure, and in this way obtain a set of interaction confidence scores *(d)*. Finally, we build multi-target training instances and train a multi-target regression tree *(e)* to maximise the homogeneity both in the input and in the output spaces, between gene regulations of both human and mouse. The values in the circles of the regression tree represent the prediction (for the human and for the mouse organisms) provided to a gene pair falling into a specific leaf of the regression tree.

### 2.3.1 Gene expression levels

We collected raw expression data from the Gene Expression Omnibus - GEO (https://www.ncbi.nlm.nih.gov/geo/). We considered the platform GPL570 for the human organism and the platform GPL1261 for the mouse organism. We took only control samples to reconstruct the gene regulatory networks, without the potential influence of disease conditions. The complete list of the considered GEO Accession Numbers can be found in an excel file at this link: Chapter 2 Supplementary Data. Quantitatively, for the human organism, we collected $54,675$ probesets, described by 180 samples (that correspond to features in our case): 17 for bone marrow, 37 for brain, 6 for breast, 4 for heart, 7 for liver, 45 for lung, 64 for skin. These samples were obtained by using Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays. As for the mouse organism, we collected $45,101$ probesets described by 171 features, distributed as follows according to the organs: 14 for bone marrow, 8 for brain, 10 for breast, 8 for heart, 124 for liver, 4 for lung, and 3 for skin. These samples were instead obtained by using the Affymetrix GeneChip Mouse Genome 430 2.0 Array technology. To correct for the batch effect, the raw samples were processed according to the workflow proposed in the DREAM5 (Dialogue on Reverse Engineering Assessment and Methods) challenge, which was the fifth annual set of DREAM challenges, consisting in inferring the transcriptional regulatory network of E. coli, S. Cerevisiae, S. aureus and of a simulated network by starting from anonymised gene expression transcripts [329]. The preprocessing protocol, used to reduce noise and ensure robustness of the data, consisted in microarray normalisation Robust Multichip Averaging (RMA) [330]. RMA is a three-step process involving background adjustment, quantile normalisation and median polish. In particular, background adjustment (or correction) is performed to remove local artifacts from the fluorescence intensities. Quantile normalisation is then used so that measurements from different samples become comparable, while the median polish (consisting in repeatedly subtracting an overall median from the rows and columns of the data, arranged in a tabular format) is used to combine the probe intensities across the samples. For each organism, considering one batch per organ, we performed RMA using the Affymetrix Expression Console Software. Finally, in order to more easily interpret the fold change, the data were $\log_2$-transformed. We then mapped the Affymetrix probeset IDs to gene symbols with the use of the Affymetrix libraries (when multiple probesets mapped to the same gene their expression values were aggregated through the arithmetic mean). The so-processed data consisted of 23035 genes per 180 samples (human organism) and of 21681 genes per 171 samples (mouse organism).

### 2.3.2 Metabolic features

To construct the metabolic features, we first filtered out genes with no corresponding HGNC ID (MGI ID for the mouse organism) [331]. We also removed all the genes for which we did not find any regulatory information according to the RegNetwork database [332], which resulted in a total of 16272

and 14067 genes for the human and mouse models, respectively. Finally, to obtain an expression fold change for constraining the metabolic model, each gene expression value was normalised against its median value across all the samples.

In order to include the regulatory information into the metabolic features explicitly, we used TRFBA [324], which integrates a transcriptional regulatory network and the related organism genome-scale metabolic model. We used Recon2.2 [70] and iMM1415 [71] as the human and mouse metabolic models respectively. The solution selected by TRFBA lies in the feasible solution space defined by the following constraints:

$$\begin{aligned}
&\mathbf{S}\,\mathbf{v} = 0 \\
&\mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub} \\
&\sum_{i \in R_j} v_i \leq E_j \times C \\
&s_I \times \sum_{r \in G_T} E_r - E_T + U \times w_{I,1} + U \times w_{I,2} \geq -IN_I \\
&\sum_{r \in G_T} E_r + U \times w_{I,1} - U \leq \lambda_I \\
&\sum_{r \in G_T} E_r - U \times w_{I,2} + U \geq \lambda_{I+1},
\end{aligned} \tag{2.1}$$

where $\mathbf{S}$ is the stoichiometric matrix associated with the species' organism's metabolism, $\mathbf{v}$ is the vector of metabolic flux rates, $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ are the lower and upper bounds of the metabolic fluxes respectively, $R_j$ is the set of indices corresponding to the reactions which are associated with metabolic gene $j$, $G_T$ is the set of indices of the regulatory genes of target gene $T$, $E_i$ indicates the gene expression of gene $i$, $U$ is a very large number (in our experiments, it was set to the maximum observed expression level multiplied by 5) and $s_I$, $IN_I$, $w_{I,1}$, $w_{I,2}$, $\lambda_I$ and $\lambda_{I+1}$ are parameters computed directly by the method from the gene expression levels [324]. The hyperparameter C, used to convert the expression levels of the genes to the upper bounds of the reactions, was set to 0.00014 as suggested by [333]. All the other parameters of TRFBA and the boundary constraints for the metabolic models were left to the default values. Therefore, TRFBA adds to the stoichiometric matrix of a GSMM further reactions representing the transcriptomic regulations among the genes, which we exploited by computing the single gene-knockouts and the resulting metabolic fluxes via FBA [49] for all the genes in the datasets. This was performed for both organisms, obtaining a flux distribution for each knockout.

To account for the tolerance of the solver Gurobi, we then eliminated all the obtained fluxes whose value was lower than $10^{-7}$ for all the samples, and applied PCA to reduce the dimensionality of the flux distributions. In both cases we retained an explained variance >99%, obtaining 250- and 150-dimensional features for the human and mouse samples, amounting to 1.7% and 0.92% of the original features respectively (14705 and 16383). These steps were conducted using the COBRA toolbox [334] in Matlab R2017b.

### 2.3.3 Transfer learning for the reconstruction of the human GRN from metabolic features

We here describe our transfer learning approach for the reconstruction of the human GRN, which also exploits the information conveyed by the mouse organism. Our approach learns a model that is able to predict a score in $[0, 1]$, representing the degree of certainty about the existence of a given regulation between two genes. The synergies among the two considered organisms are captured by resembling to a multi-target prediction model, which aims at predicting the existence of a given regulation between two genes in the two organisms simultaneously. Although predicting the existence of a given regulation for the mouse organism is not of specific interest in this study, this strategy allows us to exploit possible correlations between the organisms not only in the input space, but also in the output space [335].

Methodologically, we focused on *orthologous* genes, i.e. different genes of the human and mouse organisms that originated from a single common ancestor gene. Each possible pair of orthologous genes corresponds to a unit of analysis for the predictive task at hand, namely to a possible regulation activity between the two genes. The descriptive attributes of a gene pair correspond to the concatenation of principal component features calculated from flux rates obtained after the respective single-gene knockouts. On the other hand, the value of each target attribute (i.e., the degree of certainty of the existence of such regulation, in the human and in the mouse organisms, respectively) was set to 1.0 if the corresponding gene regulation was experimentally validated according to the BioGRID database [336], or estimated through a clustering-based solution [320] if such regulation has not yet been studied (i.e., it is an unlabelled example). This setting corresponds to the so-called Positive-Unlabelled setting, that is a subclass of the semi-supervised setting as well as a different way to model a one-class classification task [337]. We note that, for the descriptive attributes of each gene pair, one can in principle compute a flux distribution after a double gene-knockout for the pair, rather than concatenating single-gene knockout fluxes; however, this would require prohibitive computational resources for the dataset and metabolic model at hand (several years of computational time), but it could be a viable approach for smaller models.

Specifically, the known regulations were grouped into clusters, whose number was optimised via a silhouette analysis [338]. The value of the target attributes for unlabelled pairs of genes was then estimated according to the similarity with their closest cluster, computed on the descriptive attributes. Formally, given the descriptive feature vectors $x_h \in \mathbb{R}^p$ (for the human organism) and $x_m \in \mathbb{R}^q$ (for the mouse organism) for the same gene pair, we computed the value of the target variables $t_h$ (for the human organism) and $t_m$ (for the mouse organism) to use during the training of the predictive model, as follows:

$$\begin{aligned} t_h(x_h) &= \max_{c \in C_h} sim_p(x_h, cent(c)) \\ t_m(x_m) &= \max_{c \in C_m} sim_q(x_m, cent(c)), \end{aligned} \tag{2.2}$$

where $C_h$ and $C_m$ are the sets of clusters identified for the human and mouse organisms, respectively; $cent(c)$ is the feature vector of the centroid of the cluster $c$; $sim_k$: $\mathbb{R}^k \times \mathbb{R}^k \rightarrow [0,1]$ is a vector similarity function working on arbitrary $k$-dimensional vectors, based on the Euclidean distance after applying a min-max normalisation (in the range $[0,1]$) to all the descriptive features. Formally, $sim_k(a,b) = 1 - 1/k \cdot \sqrt{\sum_{i=1}^{k}(a_i - b_i)^2}$. In this way, we exploited both the information on verified regulations and the information conveyed by a large set of unlabelled examples, according to their similarity with respect to labelled examples.

Finally, we built a predictive model in the form of a multi-target regression tree, by exploiting the system CLUS [339], which is based on the predictive clustering framework. Predictive clustering approaches appear adequate to solve the task at hand, since they have proven to be generally effective in detecting different kinds of autocorrelation phenomena [340], including network autocorrelation phenomena usually exhibited by data organised in network structures [341, 342].

The multi-target regression tree was built via a standard procedure for the top-down induction of regression trees, where the tests of the internal nodes are greedily chosen by considering the reduction of variance achieved by partitioning the examples according to this test. In our case, the model aims to reduce the variance of both target attributes $t_h$ and $t_m$. More formally, for a given internal node of the tree under construction, it aims to maximise the reduction of the average variance over the target attributes due to the split, namely

$$Var_X(t_h, t_m) - (Var_{X'}(t_h, t_m) + Var_{X''}(t_h, t_m)), \tag{2.3}$$

where $X, X', X''$ are the sets of examples in the parent, left child and right child nodes, respectively, and $Var_Z(t_h, t_m) = \frac{Var_Z(t_h)+Var_Z(t_m)}{2}$ is the average variance on the target attributes $t_h$ and $t_m$, computed over the set of examples $Z$. As a result, we maximised the homogeneity of the defined subsets of examples, that also depends on the correlations, both in the input and in the output spaces, between gene regulations of both the human and the mouse organisms.

## 2.4 Results and Discussion

The final network under consideration consists of 512,576 possible interactions. Among them, 507,656 are unlabelled, while 4,920 are labelled/known interactions from BioGRID. Therefore, the proportion of labelled:unlabelled interactions is ∼1:100.

We compare the results obtained by our framework based on metabolic features, hereafter referred to as TRANSFER, with those achieved by two different settings:

- **Expression Levels**. We adopt the same workflow proposed in this paper, but directly using

the expression level features instead of metabolic features. This setting allows us to evaluate the actual contribution provided by the metabolic features, thus helping us understand whether genome-scale metabolic models effectively augment machine learning models with additional biological knowledge.

- **NOTRANSFER**. We only exploit features related to the genes of the human organism. This setting allows us to evaluate the contribution provided by information conveyed by the mouse organism as well as the effectiveness of the proposed transfer learning solution.

The experiments were performed through 10-fold cross-validation. In particular, each fold consisted of 9/10 positive examples for training and 1/10 positive examples for testing, while all the unlabelled examples were used for both training and testing purposes. Therefore, coherently with the *semi-supervised transductive* setting [343, 344], at training time the methods knew the examples for which they have to make a prediction, i.e., they may have already observed and exploited the value of descriptive attributes, but not the actual value of the target attributes. We note that the confidence scores estimated by our method are not adopted to define a ground truth for unlabelled examples, but only as an intermediate step for the construction of the multi-target regression tree.

The results were evaluated in terms of recall@$k$ (r@k), the area under the recall@$k$ curve (AUR@K), the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPR). We note that, while r@k and AUR@K do not introduce any bias on the existence of a regulation activity on unlabelled gene pairs, the computation of the AUROC and AUPR requires considering the unlabelled examples as negative examples.

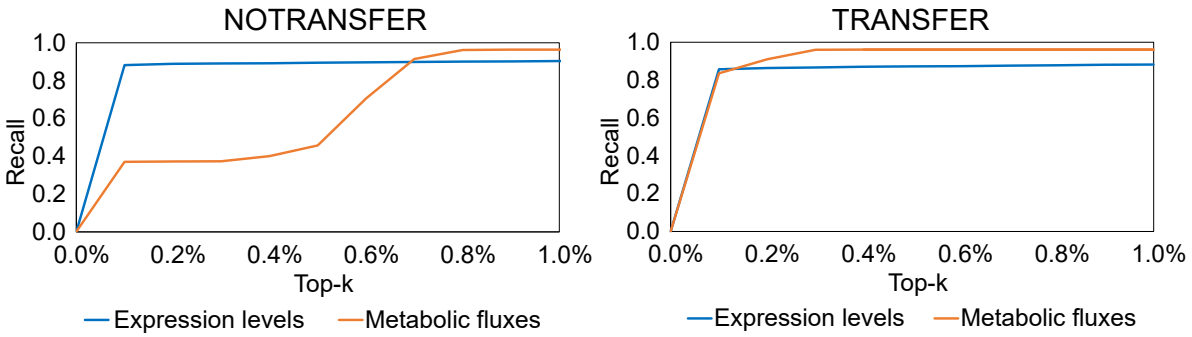



Figure 2.2: **Recall@$k$ comparison in the different experimental settings.** Recall@$k$ measured in the range $[0, 1\%]$ for the reconstruction of the human GRN, by considering different sets of features. The NOTRANSFER approach does not exploit data of the mouse organism, while the TRANSFER approach exploits also the mouse GRN knowledge.

In Figure 2.2 we show the measured recall@$k$ in the range $[0, 1\%]$, that is the range of the top-1% most reliable interactions returned by all the approaches considered. Our results show that the adoption

of metabolic fluxes is beneficial, with respect to directly adopting the raw gene expression levels, both when exploiting the knowledge coming from the mouse organism (TRANSFER) and when ignoring such additional information (NOTRANSFER). Specifically, such an improvement amounts to 6.6% in the case of NOTRANSFER and to 8.73% in the case of TRANSFER, when observing the recall@1%. Moreover, it is noteworthy that, in the TRANSFER setting, we identify existing gene regulations much earlier in the returned ranked list of interactions. Specifically, we identify 96% of the known interactions of the testing set in the top-0.3% interactions returned in the case of the TRANSFER setting, whereas we need to consider 0.8% of the list of the returned interactions to identify the same amount of known interactions in the NOTRANSFER setting. This behaviour emphasises that the knowledge coming from the mouse organism can fruitfully be exploited to improve the accuracy of the reconstruction of the human GRN.

A more comprehensive overview is reported in Figure 2.3, where we show boxplots representing AUR@K, AUROC and AUPR measured over the 10 folds of the cross-validation. These show that the predictive models trained via metabolic fluxes can better exploit the mouse gene regulation knowledge leading to more stable predictive models (i.e., with a lower variance observed over different folds of the 10-fold CV). Furthermore, the area under all the considered curves is higher and more stable when adopting the metabolic fluxes in combination with the TRANSFER setting. Conversely, when adopting metabolic fluxes in the NOTRANSFER setting, we observe worse results with respect to directly using expression levels. This phenomenon indicates that the metabolic fluxes of the human organism alone are not able to describe the regulatory activities as well as expression levels, but the exploitation of mouse and human metabolic fluxes in combination provides our framework with a significant advantage, leading to the best overall results. This observation confirms that the proposed workflow, which synergically exploits metabolic fluxes and the knowledge of the mouse GRN, provides significant advantages in terms of the quality of the reconstruction of the human GRN, and demonstrates that GSMM-generated data contain different information than gene expression data, which highlights how promising their use is in the field.

To understand the contribution provided by human and mouse metabolic features in human GRN reconstruction, we performed additional experiments in the TRANSFER and NOTRANSFER settings. Specifically, we considered the approach proposed by [345], based on the evaluation of the (negative) effect of noise. We purposely introduced noise in a given feature by randomly permuting its values over the examples, evaluating the effect on the predictive performance of the tree: the greater the performance degradation, measured through the relative increase of the predictive error, the higher is the importance/contribution of the feature. We produced a descending ranking of the features for each fold of the 10-fold cross-validation, and analysed the average ranks (an excel file containing such computed ranks can be found at the link Chapter 2 Supplementary Data). In the rest of this thesis, instead, the alternative approach of looking at the model's weights was used to determine the feature contribution/importance.

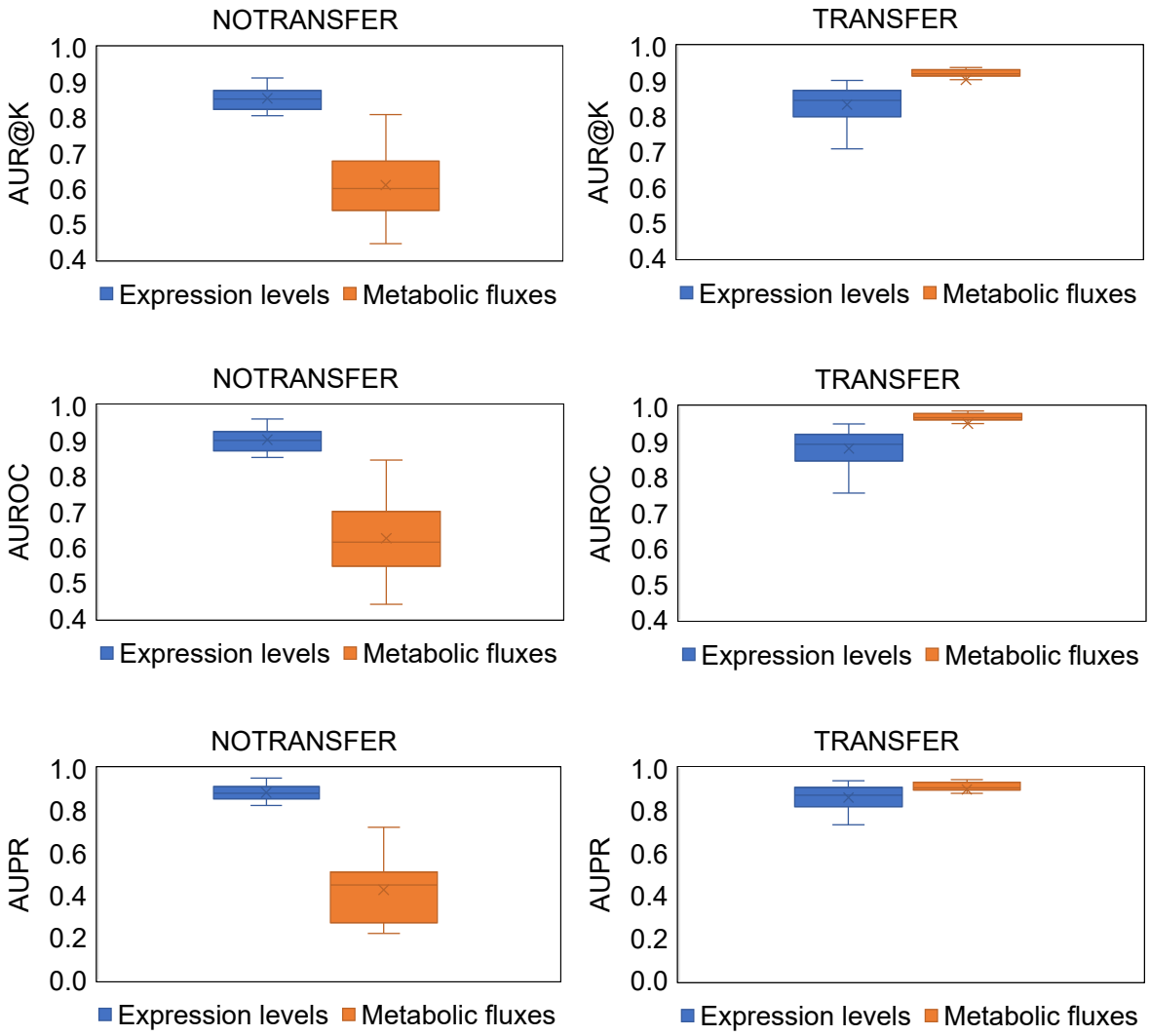



Figure 2.3: **Boxplots for the 10 folds of the human GRN reconstruction task.** Each row corresponds to a measure, i.e., AUR@K, AUROC, AUPR, respectively, measured in the range [0, 1%] of the top-k ranked interactions. Each column corresponds to a learning setting, i.e., without and with the exploitation of the mouse GRN knowledge, respectively.

As shown in Figure 2.4(c), the metabolic features related to the mouse organism dominate the upper half of the ranking and are therefore assigned a higher relevance than those related to the human, in the setting TRANSFER. Conversely, when directly using gene expression levels, many features from human retain a high relevance when combined with those from mouse. This finding further confirms the advantage provided by the adoption of our transfer learning technique on GSMM-derived information, and suggests that the regulatory mechanisms present in the mouse metabolism are relevant in the human metabolism as well.

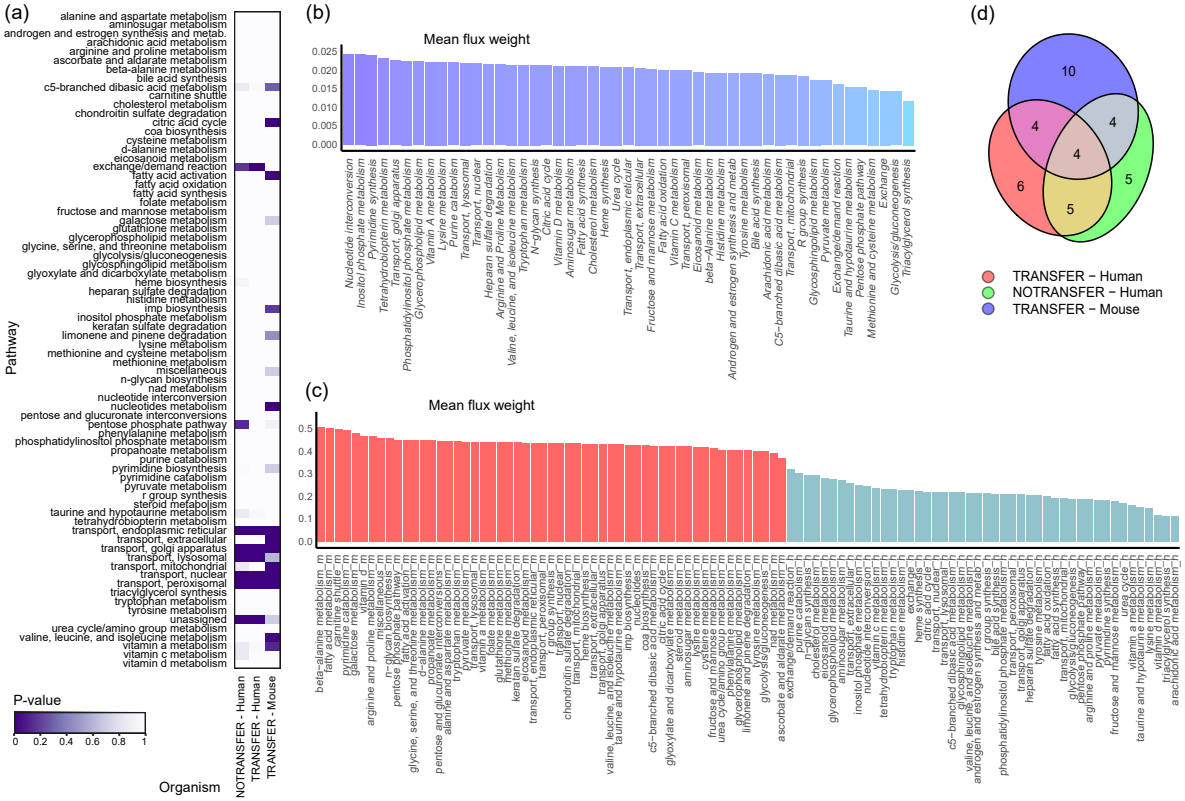



Figure 2.4: **Analysis of the metabolic features.** (a) Enrichment $p$-values (corrected through the Benjamini-Hochberg procedure for multiple hypothesis testing) for the pathways assigned to the 10% most relevant metabolic features in the three experimental settings considered. (b) Mean flux weight across pathways for the human metabolic features used in the setting NOTRANSFER. (c) Mean flux weight across pathways for the human (blue) and mouse (red) metabolic features used in the setting TRANSFER. (d) Euler-Venn diagram that summarises the overlap in terms of biological pathway enrichment (pathways with associated corrected $p$-value $\leq 0.05$) for the 10% most relevant metabolic features in the three considered settings.

Further, a consistent number of metabolic features (295/800) present a relevance score equal to zero, as opposed to the more gradual decline in gene expression feature relevance. This is in line with

previous experimental results from another data integration task, where metabolic features displayed a highly skewed relevance distribution compared to transcriptomic ones [255]. A possible explanation is given by the structure of metabolic networks and by the method used to estimate its activity, which is based on a linearly constrained MILP problem that generates collinearity and redundancy among the features.

Consistently, the addition of mouse-related features impacts the importance of human-related features to a varying degree depending on their type. When comparing the TRANSFER and NOTRANSFER settings, human metabolic features have indeed an average importance difference of $2.12 \pm 2.80$, whereas for human transcriptomic features such difference is $2.99 \pm 1.25$. In other words, human gene expression features that are considered poorly (or highly) relevant in the NOTRANSFER scenario have on average a higher chance to be considered more (or less) relevant in the TRANSFER setting – and by a larger extent – as compared to human metabolic features. However, the difference in importance for the latter is highly variable and reaches the highest values. The addition of mouse-related features therefore appears to drastically change the learned model when using metabolic features.

As a complementary approach to determine the importance of the input features, we inspected the metabolic pathways associated with the most relevant reactions adopted in the construction of the metabolic features with the means of a Flux Enrichment Analysis (FEA). Enrichment analysis is a statistical testing technique that computes the probability that a set of fluxes/genes belongs to a specific subsystem or pathway of the cell, i.e. that a class of fluxes/genes is over-represented in a large set. The purpose of this is to understand whether the class is overly present in the pool "by chance" or because of the existence of real biological mechanisms linked to it. In general, these tests make use of the hypergeometric cumulative distribution:

$$F(x|M,K,N) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}, \tag{2.4}$$

where $M$ is the total size of the population (genes or fluxes in our case), $K$ is the size of the population belonging to a certain subsystem/pathway/class, $N$ is the size of the set we are considering for the analysis, and $x$ is the number of elements, among those $N$, which belong to the group of size $K$ (i.e. the number of elements which share a specific characteristic and are the subject of the analysis). Since $F$ is a cumulative distribution, it indicates the probability of finding *up to* $x$ elements belonging to the same group in the considered set. The associated $p$-value therefore indicates the probability of obtaining such a result by chance alone. Here, we conducted the flux enrichment analysis using the MATLAB Bioinformatics Toolbox on the subset of reactions which, for each organism in the two experimental settings (TRANSFER and NOTRANSFER), had been given a weight above the 90th percentile. The weight for the $j$-th reaction was computed as

$$\theta_j = \sum_i |l_{ij} \times \sigma_i^2 \times (rank_{1j} + rank_{2j})|, \tag{2.5}$$

where $l_{ij}$ is the linear coefficient of the $j$-th feature/reaction with respect to the $i$-th principal component deriving from the PCA (adopted to generate the metabolic features), $\sigma_i^2$ is the variance explained by the $i$-th principal component, while $rank_{1j}$ and $rank_{2j}$ are the rankings of the $j$-th feature, computed using the approach proposed by [345], when considered in the first and second position, respectively, in the gene pair. From these values, we computed the average flux weight for each metabolic pathway as the average weight of its reactions. These weights can be considered a proxy of the relevance of the input features to the model.

As shown in Figure 2.4(a)(d), the number of enriched pathways (associated $p$-value $\leq 0.05$, corrected through the Benjamini-Hochberg procedure for multiple hypothesis testing) is higher for the metabolic features of the mouse, while it is almost equal for the human ones. Indeed, reactions enriched in the human features employed when building the model without the mouse features (NOTRANSFER-Human) were all included in the pool of enriched reactions from the human features used in the TRANSFER setting. In particular, in this setting, the enrichment also includes exchange/demand reactions ($p$-value $> 0.05$ for the NOTRANSFER-Human features), indicating that adding features from a different organism increased the importance of the features associated with internal production/-consumption reactions and extracellular/intracellular transport reactions. Conversely, mouse features share all the transport pathways of the human ones, except for that relating to lysosomal transport, and also encompass the pathways associated with the citric acid cycle, nucleotide metabolism, fatty acid activation and the metabolism of leucine, isoleucine and valine (see Figure 2.4 (a)).

Overall, these results demonstrated the effectiveness of the proposed approach, which exploits metabolic information coming from two organisms through our transfer learning method. Moreover, the analysis of the contribution of the metabolic features emphasised the new information introduced by the mouse features. We believe that our results pave the way towards the exploitation of knowledge of multiple model organisms - across several omic layers - while reconstructing the GRN of a target organism, but more importantly demonstrate that genome-scale metabolic models provide information which is not present in (or easily extractable from) gene expression data, and that such information can be leveraged by machine learning models to improve their performance. In particular, this may be due to the fact that metabolic fluxes approximate the production (and therefore abundance) of transcription factors better than gene expression directly, or to the fact that the metabolic network of a GSMM enforces regulatory mechanisms among genes by constraining the values of metabolic fluxes via the use of FBA (therefore reflecting regulatory activity exclusively on the reaction fluxes). In the following chapters we will continue our investigation by trying to understand whether the integration of GSMM-generated metabolic fluxes with gene expression data can be beneficial to machine learning models more than the independent use of these two omics.

## 2.5 Conclusion and future directions

In this chapter, we presented a novel method for the reconstruction of the human gene regulatory network that fruitfully exploits the information conveyed by *in silico*-generated metabolic fluxes of both mouse and human organisms. Specifically, we exploit a transfer learning method to capture analogies between the metabolic responses in mouse and human upon simulated deletion of their orthologous genes.Our aim was to demonstrate how GSMM-generated features could provide different information from the more commonly used gene expression data, and that this information is more easily interpretable than other types of structured or unstructured data.

Our results show that metabolic features, computed from gene expression levels and metabolic modelling, improve the performance and the stability of the trained predictive models when exploited in combination with our transfer learning approach. This emphasises that the underlying regulatory patterns are better captured when (both known and possible) gene regulations are described through metabolic features, computed through genome-scale metabolic model simulations, on both the human and the mouse organisms. To the best of our knowledge, this is the first attempt to exploit metabolic features and a transfer learning approach for the reconstruction of the human GRN, and our results support the adoption of the developed method as a state-of-the-art tool for solving this task.

However, this study is not devoid of limitations. For example, the use of TRFBA requires some regulatory knowledge for the computation of the metabolic fluxes. This means that, in the absence of this information, reconstruction of the regulatory network with only the metabolic features may be less effective. Moreover, this information should be complete enough: knowledge of around 50% (for instance) of the regulatory network would generate worse fluxes than by simply using standard FBA. Another limitation, linked again to the use of the genome-scale metabolic models, is the use of single gene-knockouts. In this case study, we have characterised gene pairs with the concatenation of their metabolic features obtained from single gene-knockouts. However, this has prevented us from exploiting the combined effect that multiple gene mutations could have on the metabolism of the organism, since we are considering these mutations separately (and then concatenating the resulting metabolic profiles). Unfortunately, in this instance this problem has no solution, since the size of the model and dataset could not allow for a higher computational workload. Finally, previous experiments have demonstrated how the percentage of labelled samples used in a semi-supervised experiment can influence, positively or negatively, the final performance of the machine learning model [346]. Since this effect strongly depends on the task and dataset being used, in the future this further parameter should be taken into account when possible. In our case, the ratio labelled/unlabelled examples proved successful, but a different outcome could result with different data.

As future work in the field of GRN reconstruction, it would be interesting to design a multi-source approach to capture possible dependencies among multiple organisms and to simultaneously recon-

struct their GRNs, even when the knowledge about their orthologous genes is limited. In conjunction with multi-omic integration strategies, this could lead to refined GRN reconstructions, thus expanding the current knowledge on the biological mechanisms of metabolic regulation.

## 2.6 Related work, funding and final remarks

The work presented in this chapter has been published in Bioinformatics [347], and the analyses were completed in collaboration with multiple co-authors, as described in the introduction to this chapter.

Part of this work was supported by the Ministry of Universities and Research through the project "Big Data Analytics", AIM 1852414-1(line 1), by the UKRI Research England's THYME project, by a Children's Liver Disease Foundation Research Grant and by the Apulia Region through the "Research for Innovation - REFIN" initiative (Grant n. 7EDD092A).

# Chapter 3

# Concatenating transcriptomic and fluxomic data for yeast growth rate prediction

## 3.1 Introduction

In the previous chapter we have shown that GSMM-generated metabolic fluxes can reveal, within the right framework, information which is different than the one present in gene expression data. The next step would consist of determining whether this additional, distinct information can be combined with the information held by gene expression data in order to improve machine learning models' performance (as opposed to using it by itself), with the pleasant "side effect" of having the opportunity to interpret the results more easily in a mechanistic way. For this reason, the questions that we are trying to answer in this chapter are the following:

- Can we combine GSMM-generated metabolic fluxes and gene expression data to exploit their different information in order to improve our machine learning models?
- Can this "integration" be as simple as a concatenation?
- Which types of machine learning models are better suited for following this approach, if any?

As in the previous chapter, we will use a case study to answer the above questions, but unlike the preceding pages, the case study we present here will be a precision medicine application in the classical

way the term is intended, except for the fact that we will not deal with patients but with yeast strains.

The contributions of this chapter are the following: we show that under certain conditions the integration of metabolic fluxes and gene expression data performs better than the use of these two omics by themselves, and demonstrate that simple models such as linear ones can be a better practical choice than more complex and flexible models in this setting. In this chapter we also show a case in which the integration of these two data types is a simple concatenation, while in the next chapter we will explore a more complex integrative approach.

The presented case study is a classic regression task, in which we are trying to predict yeast growth rate. The entirety of the study was conducted by me. This means that I performed all the analyses relating to the generation of the metabolic features, I devised, implemented, trained and tested the models, and I conducted the analysis and interpretation of the results as well.

## 3.2 Background

Understanding and controlling cellular growth is fundamental in biotechnology for the development of efficient cell factories [348, 349]. CRISPR/CAS-enabled genetic engineering gives the ability to modify DNA with single-nucleotide precision *in vivo* [350, 351], making the engineering of strains that maximise a desired output possible for industrial purposes. Yet, the identification of such strains is still a complex issue [352] which requires considerable amount of time and notable costs. In the past, the problem of cellular growth prediction has mainly focused on mechanistic representations of biomolecular processes. These, however, require detailed knowledge of uptake rates from the environment in order to achieve accurate estimates. On the other hand, it is also possible to find correlations between gene expression and cell growth using only data-driven machine learning methods. Previous research focused on building linear predictive models for yeast growth [353], and more recently, machine learning models both for *E. coli* and *S. cerevisiae* [354]. The metabolic activity of *S. cerevisiae* in combination with machine learning techniques was only evaluated in recent times [255].

With the technological advances of the past two decades, we now have access to enormous amounts of biological data, which in Section 1.1 we referred to as omics. Each of these omic types represents a different facet of an organism and its functioning, which suggests the presence of shared patterns and intertwined mechanisms among them. For this reason, the development of multi-modal learning methods in a biological setting has been recently promoted [355]. Thanks to the flexibility of machine learning approaches (especially deep learning), this subfield of research has been applied to several tasks: transfer learning (framework within which we have operated in the previous chapter) [356], integrative clustering [183] and drug response prediction (among the many) [184].

However, as already highlighted in Section 1.6, these technologies are mainly used as black boxes and, depending on their architecture, may not be able to produce new knowledge on the underlying biological mechanisms. In many situations, the use of appropriate linear models for high-dimensional data (and interpretability purposes) can hence be a preferable option. In this chapter, we continue the investigation started in Chapter 2 to devise and compare multimodal regression methods that utilise both transcriptomic data and strain-specific metabolic models to predict cellular growth of *Saccharomyces cerevisiae*, one of the main eukaryotic platforms for bio-industrial production [357]. We combine regularised statistical learning methods with flux balance analysis for omic data integration, in a regression setting designed to exploit the different information present in the two different views (the gene expression data and the metabolic fluxes). To this end, we use a compendium of 1,143 single gene knock-out yeast strain expression profiles to predict cell doubling rates. We leverage the GSMMs at a steady state (given that gene expression maintains a steady state during the exponential growth phase [358], predicting growth in such a simplified setting is reasonable) with a parsimonious implementation of flux balance analysis to generate strain-specific reaction flux rates, which are then added to the gene expression profiles as additional features (Figure 3.1).

We investigate a range of regularisation techniques, proposing expansions of previous frameworks and empirically evaluating them on a common benchmark and show that, in this setting, group and view-specific regularisations achieve higher performance than principal component regularisation, outperforming multimodal neural networks. On the other hand, the latter obtains a larger performance improvement when concatenating transcriptomic and fluxomic data. Overall, our results demonstrate the competitiveness of multimodal regularised linear models compared to data-hungry neural network-based methods in a multi-omic task using experimental and model-generated omic data. At the same time, we highlight the lack of a clearly superior method for effective and transparent omic data integration through concatenation, further underlying the importance of a bespoke selection of both features and machine learning models for each case study. In the next chapter we will take a final step towards a more complex integrative approach, and investigate how different omics combinations can be more appropriate for different individuals.

## 3.3 Materials and Methods

### 3.3.1 Dataset

We used a transcriptomic dataset generated in a previous study [2], which contains two-channel microarray profiles for 1,484 single-gene deletion strains of *S. cerevisiae* during early mid-log phase. In the original study, each deletion strain consisted of four replicates: two biological ones from two independent cultures, each profiled in technical replicates (each gene was represented twice in the

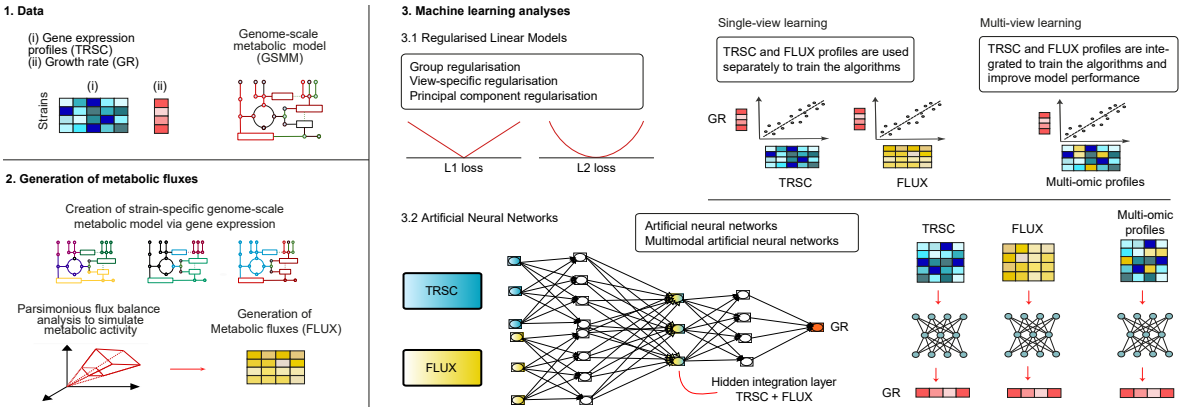



Figure 3.1: **Machine learning pipeline.** Pipeline adopted in this study. From 1,143 *S. cerevisiae* strains, the gene expression was used as a starting point [2]. A genome-scale metabolic model was then used (panel 1) to generate strain-specific GSMM models. From these GSMMs, metabolic fluxes were generated via parsimonious flux balance analysis (panel 2, see Subsection 3.3.2). The machine learning methods were applied in two different settings: single-view and multi-view regression. In the former case, transcriptomics and fluxomic data were used separately as input for regularised linear models and artificial neural networks, while in the latter the two omics were concatenated to let the two classes of methods leverage the different information of both sources (panel 3).

microarray, therefore $2 \cdot 2 = 4$ measurements for each mutant). The biological replicates were compared against each other with the help of the R package limma [359] for quality control purposes: in case a significant overlap in the expression was not found by a hypergeometric test when the genes changing significantly in the mutant strain were more than seven, the hybridization was repeated. In order to control for day-specific effects and monitor batch effects, over 400 wild-type cultures were grown in parallel alongside the mutant strains. These were used to determine whether the effects observed in the mutant strains were specific to the strains or not. Notably, hypergeometric tests were used to determine whether a relevant overlap in the expression of the significantly changed genes in the mutant strains was present between these and the wild-type ones grown on the same day. In this case, if the number of such genes was more than seven, the hybridization was repeated. A common reference design with wild-type reference RNA was applied in dye-swap to control for the dye bias as well. Finally, microarray data normalisation was performed using print-tip LOESS [360]. We downloaded the data from the supplementary materials of a second study providing relative growth rates compared to the wild type for 1,312 of the strains grown on the same days, expressed as $\log_2$ of the doubling times ratio (the doubling times of the two biological replicates were averaged together) [361]. The final (fold-change) gene expression dataset was composed of those strains for which the flux balance analysis formulation described below provided feasible solutions (1,143 samples), and is here denoted as TRSC. The distribution of the relative growth rates (in terms of $\log_2$ of the doubling

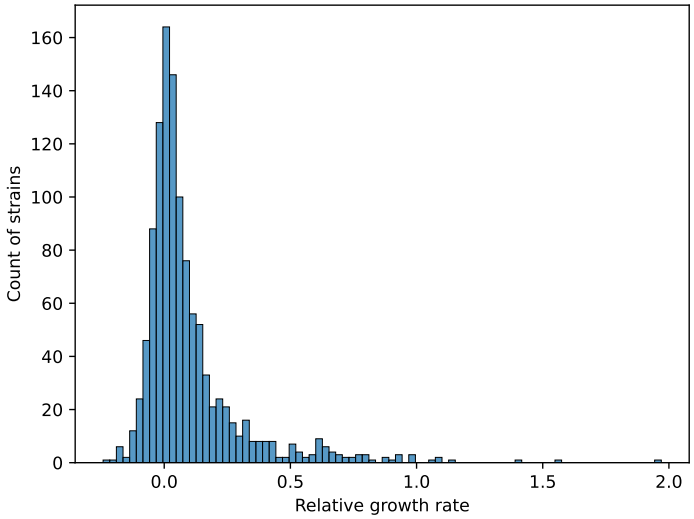



Figure 3.2: **Distribution of the relative growth rates for the strains considered.** The relative growth rate of each mutant strain was computed as the $\log_2$ of the ratio of the doubling times (mutant vs wild type). For the biological replicates, these values were averaged together.

times ratio) can be found in Figure 3.2. Pre-processing was applied separately on the fluxomic data (denoted as FLUX) and the gene expression profiles. For the fluxomic data, all the reaction fluxes for which the value was $< 10^{-7}$ for all the samples were discarded (negligible fluxes in all samples). All data were standardised by subtracting the mean from each feature and dividing by the standard deviation, which yielded better results following a preliminary exploration of normalisation techniques (namely min-max normalisation, log normalisation, and normalisation in [-1, 1]). Finally, in addition to these two datasets, a third one was built by joining (i.e. concatenating) the previous two (TRSC + FLUX). This was done so that the integration of transcriptomic data and fluxomic data could be compared with the use of one of the two omics alone, in the hope to demonstrate the usefulness of GSMM-generated data with machine learning models.

### 3.3.2 Genome-scale metabolic modelling

As genome-scale metabolic model for this investigation we utilised the *i*Sce926 yeast GSMM, which includes 926 genes, 3494 reactions and 2223 metabolites [362]. Among all the genes in the TRSC data, a total of 908 (98%) were present in our transcriptomic dataset. Information on the simulated medium is reported in Table 3.1.

**Parsimonious flux balance analysis (pFBA).**

Table 3.1: **Composition of the simulated medium**

| **Medium component** | **Exchange reaction name** | **Exchange reaction ID** |
|---|---|---|
| ammonium | ammonium exchange | r_1654 |
| sulphate | sulphate exchange | r_2060 |
| biotin | biotin exchange | r_1671 |
| (R)-pantothenate | (R)-pantothenate exchange | r_1548 |
| folic acid | folic acid exchange | r_1792 |
| myo-inositol | myo-inositol exchange | r_1947 |
| nicotinate | nicotinate exchange | r_1967 |
| 4-aminobenzoate | 4-aminobenzoate exchange | r_1604 |
| pyridoxine | pyridoxine exchange | r_2028 |
| H+ | H+ exchange | r_1832 |
| riboflavin | riboflavin exchange | r_2038 |
| thiamine(1+) | thiamine(1+) exchange | r_2067 |
| sulphate | sulphate exchange | r_2060 |
| potassium | potassium exchange | r_2020 |
| phosphate | phosphate exchange | r_2005 |
| sulphate | sulphate exchange | r_2060 |
| sodium | sodium exchange | r_2049 |
| L-alanine | L-alanine exchange | r_1873 |
| L-arginine | L-arginine exchange | r_1879 |
| L-asparagine | L-asparagine exchange | r_1880 |
| L-aspartate | L-aspartate exchange | r_1881 |
| L-cysteine | L-cysteine exchange | r_1883 |
| L-glutamate | L-glutamate exchange | r_1889 |
| L-glutamine | L-glutamine exchange | r_1891 |
| glycine | glycine exchange | r_1810 |
| L-histidine | L-histidine exchange | r_1893 |
| L-isoleucine | L-isoleucine exchange | r_1897 |
| L-leucine | L-leucine exchange | r_1899 |
| L-lysine | L-lysine exchange | r_1900 |
| L-methionine | L-methionine exchange | r_1902 |
| L-phenylalanine | L-phenylalanine exchange | r_1903 |
| L-proline | L-proline exchange | r_1904 |
| L-serine | L-serine exchange | r_1906 |
| L-threonine | L-threonine exchange | r_1911 |
| L-tryptophan | L-tryptophan exchange | r_1912 |
| L-tyrosine | L-tyrosine exchange | r_1913 |
| L-valine | L-valine exchange | r_1914 |
| oxygen | oxygen exchange | r_1992 |
| adenine | adenine exchange | r_1639 |
| uracil | uracil exchange | r_2090 |

List of nutrients allowed to be imported when performing flux balance analysis, together with their corresponding exchange reactions in the *i*Sce926 metabolic model [362]. These correspond to commonly used media [363, 364].

We used a variation of FBA, namely parsimonious FBA (pFBA) to control the global metabolic activity of the cell through an L1-regularisation for maximising our objective, at the same time making the solution as sparse as possible. The complete optimisation problem with constraints is

$$\begin{aligned} &\min_{\mathbf{v}} \|\mathbf{v}\|_1 \\ &\text{subject to } \mathbf{c}^\top \mathbf{v} = g_{max}\,, \\ &\mathbf{S}\,\mathbf{v} = 0\,, \qquad \mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub}\,. \end{aligned} \tag{3.1}$$

where $\|\mathbf{v}\|_1$ is the 1-norm (or L1-norm) of $\mathbf{v}$, $\mathbf{c}$ is a one-hot encoding vector identifying the biomass pseudo-reaction as the unique objective, and the reaction bounds are defined in the following way, as in [255]:

$$\begin{aligned} \mathbf{v}_{ub} &\leftarrow \mathbf{v}_{ub}\,\Theta^{\gamma} \\ \mathbf{v}_{lb} &\leftarrow \mathbf{v}_{lb}\,\Theta^{\gamma}\,, \end{aligned} \tag{3.2}$$

where $\gamma$ is a hyperparameter expressing the relevance of the gene expression in influencing the reaction bounds. We set $\gamma = 1$ according to [255], as this value minimises the linear correlation between predicted biomass accumulation rates and experimentally-available relative doubling times over all strains. As already explained in Section 1.2, these constraints are the mathematical representation of the several genetic or environmental factors under which the cell has to operate, and give a context-specific metabolic model that is consistent with experimental data, whereas $\Theta$ is a function representing the gene expression level of the gene sets associated to the reactions (see Equations 1.2). Finally, $g_{max}$ is the maximal growth rate achievable under these conditions. To perform the optimisation of Equation 3.1, the COBRA toolbox 3.0 [334] was used with the PDCO solver. The solutions provided steady-state flux levels $\mathbf{v}$ for each yeast strain and every reaction in the *i*Sce926 GSMM, which collectively constitute a fluxomic profile (whereas the full set of transcription levels of a sample represents the transcriptomic profile of that sample). The rationale behind this is that it should be possible to predict the growth of the strain by looking at its metabolism.

### 3.3.3 Regularised linear models for omic data

The models that we investigated belong to two different categories of machine learning techniques: statistical learning algorithms and neural networks. From the former group, we decided to consider only regularised linear models (RLMs) due to their inherent interpretability. The following multi-view approaches on the original omic profiles were employed:

**IPF-Lasso L1.** Integrative Lasso with Penalty Factors [365] is a variation of Lasso [141] that accounts for different modalities being used. Specifically, it uses penalty factors $\lambda_m$ to weigh the $L_1$

penalty applied to the $m$-th modality. The objective to minimise is thus

$$\sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M}\sum_{j=1}^{p_m} x_{ij}^{(m)}\beta_j^{(m)}\right)^2 + \sum_{m=1}^{M}\lambda_m\|\beta^{(m)}\|_1, \tag{3.3}$$

where $M$ is the number of modalities, $p_m$ the number of covariates of the $m$-th modality, $\beta$ the regression coefficients and $n$ the total number of samples. The rationale behind this approach is that each modality has, in general, a different proportion of relevant variables, hence each contribution is weighted differently.

**IPF-Lasso L2.** We extended the originally proposed IPF-Lasso algorithm, replacing the $L_1$ norm with an $L_2$ norm, which was not tested in the original paper.

**pcLasso.** Principal component Lasso is a variation of elastic net that biases the solution coefficient vector towards the leading singular vectors of the feature matrix (or, in case of grouped features, towards the leading singular vector of each matrix associated with a group) [366]. In other words, the solution is pushed towards the most important/identified pattern to improve prediction accuracy. The objective to minimise is

$$\frac{1}{2}\left\|Y - \sum_{p=1}^{P} X_p\beta_p\right\|^2 + \lambda\|\beta\|_1 + \frac{\theta}{2}\sum_k \beta_k^T(V_k D_{d_{k_1}^2 - d_{k_j}^2} V_k^T)\beta_k, \tag{3.4}$$

where $k$ is a non-overlapping group (fluxomic or transcriptomic data in this study) whose columns have rank $m_k$, $\beta_k$ is the subvector of $\beta$ corresponding to group $k$, $V_k$ are the right singular vectors of the columns of $X$ corresponding to group $k$, and $D$ is a diagonal matrix with entries $d_{k_1}^2 - d_{k_j}^2$, which are the singular values of the columns of $X$ related to group $k$ (the former associated with the leading singular vector, the latter with $j = 1, 2, \ldots, m_k$).

**pc2Lasso.** We also modified pcLasso and tested a new version, which shrinks the vector of coefficients towards the first and the second singular vectors associated with the two largest singular values. In our implementation, the entries $d_{k_1}^2 - d_{k_j}^2$ are substituted with $\alpha_1 d_{k_1}^2 + \alpha_2 d_{k_2}^2 - 2d_{k_j}^2$, where $d_{k_2}^2$ is the second-largest singular value, while $\alpha_1$ and $\alpha_2$ represent the quantity of variance explained by the first and the second largest singular values respectively.

**Group Lasso.** Group Lasso is a variation of Lasso regression in which the model is forced to include or disregard entire groups of variables defined by the user [367]. Notwithstanding the similarity with IPF-Lasso, there are two main differences: first, the groups are defined by the user without necessarily following a strict logic such as the one regarding the modalities; second, the algorithm makes a binary choice for each group, i.e. whether to include it or disregard it. In biological applications, this strategy can be justified based on the relationships among genes (e.g. whether they code the same protein, or regulate the same genes). In this investigation, the groups were defined looking at the correlation among the data in both views separately (TRSC and FLUX), while the number of groups was chosen

to encourage a larger granularity. This was set to 50 groups for the fluxes, already fairly correlated, and 500 for the transcriptomic data. We tested different values of these two parameters, but a greater number of groups would lead to non-significant clusters, while a smaller number would lose information about the potential aggregations. We conducted hierarchical clustering by using the R function *hclust* with default parameters. When using both data sources, we used the same groups we had already defined when using the sources separately. The minimisation problem we solved is

$$\frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_j \beta_j\right\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j}, \tag{3.5}$$

where $\|\beta\|_{K_j} = (\beta^{-1} K_j \beta)^{\frac{1}{2}}$, while $X_j$ and $K_j$ are respectively a design matrix identifying a group of covariates and an associated kernel. We chose $K_j$ as the identity matrix multiplied by the square root of the size of the group, as suggested in the original paper, thus obtaining an L2 penalty.

**Hybrid Group-IPF Lasso.** We developed a hybrid method to take into account both the two modalities and the possible relationships within each of them. To this end, we combined the L1 penalty of IPF-Lasso and the L2 penalty of Group Lasso on the two different omic levels. We chose the same groups chosen for the Group Lasso algorithm to make a fair comparison of the methods. The objective to minimise is therefore

$$\begin{aligned} &\sum_{i=1}^{n} \left( y_i - \sum_{m=1}^{M} \sum_{j=1}^{p_m} x_{ij}^{(m)} \beta_j^{(m)} \right)^2 + \\ &\sum_{m=1}^{M} \lambda_m \|\beta^{(m)}\|_1 + \sum_{j=1}^{J} \lambda_j \|\beta_j\|_{K_j}, \end{aligned} \tag{3.6}$$

where $\lambda_j = 1$ for $i = 1, 2, 3, ..., J$ to reduce the computational burden.

**Artificial Neural Networks.** As we already wrote in Section 1.4, ANNs are models capable of approximating any function, provided they are endowed with enough layers and/or neurons. An ANN is composed of an input layer, an output layer and one or more hidden layers in between. Each layer is made up of neurons, which are linked to the neurons assembling the other layers of the network. When the neurons are perceptrons (with or without a linear activation function), the ANN is called MLP (see Section 1.4). When a neural network has more than one hidden layer it is defined as a Deep Neural Network (DNN). For this study, numerous architectures were devised and studied, optimising several hyperparameters (e.g. number of layers, learning rate, optimisation strategy) to choose the best neural network architecture among the possible ones. We used a feed-forward neural network as a regressor, defining its architecture via hyperparameter optimisation (the exact procedure is described in Subsection 3.3.4). We also explored feature selection techniques prior to applying ANN models, but we did not proceed further as we obtained a performance decrease in all cases, as also observed before [255]. In the end, we used an ANN with two hidden layers (therefore with a total of four layers), whose hyperparameters are described in the next section.

Table 3.2: **Hyperparameter spaces for the ANN explored during Grid/Random Search**

| **Hyperparameter** | **Hyperparameter search space** |
|---|---|
| batch size | $\{32, 64, 128\}$ |
| epochs | $\{400, 800, 1200, 1600, 2000, 2400\}$ |
| learning rate | $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ |
| no. neurons of first hidden layer | range depending on the input data |
| no. neurons of second hidden layer | range depending on the input data |
| optimiser | $\{ADAM, SGD, RPROP, ADADELTA\}$ |
| dropout | $\{0, 0.3, 0.6\}$ |
| loss | $\{L1, MSE, Smooth\_L1\}$ |

For not mentioned parameters, default values were used.

**Multi-Modal Artificial Neural Networks.** Multi-Modal artificial Neural Networks (MMNN) are a particular type of ANNs devised for learning from different sources of information, in general involving the use of an independent network for processing each modality and then a further network for integrating the gathered information and producing an output. For this work, in order to ensure a fair comparison between the RLMs and the neural networks, we trained the architecture devised in [255], which inherently works in our scenario. This network is composed of two individual networks (one for the fluxomic data and one for the transcriptomic data) whose outputs are then concatenated and further processed by another network. Therefore, unlike the other models present in this chapter, this one represents a case of late integration (see Section 1.5), since the two subnetworks are trained independently before being combined together. The design of such network architecture can be seen in Figure 3.1 (panel 3).

### 3.3.4 Training and testing pipeline

We split the dataset into a training set and a test set, with an 80:20 ratio. Then, we defined a subset of the training set as the validation set, we trained only on the training set, and we optimised the hyperparameters based on the performance of the models on the validation set. All methods and models, when applicable, were optimised applying extensive grid-search over the hyperparameters (see Table 3.2 for the hyperparameter search space of the neural networks). In case a grid-search would be too computationally expensive we applied a consistent number of random-search iterations. For the neural networks, the number of iterations exceeded 100. Finally, the best combination of hyperparameters was used to train the final model to make predictions on the unseen test set.

The explored machine learning models were evaluated over several metrics: the mean squared error

(MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{3.7}$$

where model predictions $\hat{y}_i$ are compared with observed growth rates $y_i$ across all the $n$ samples of the test set; the mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|\,; \tag{3.8}$$

the coefficient of determination ($R^2$)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.9}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We also computed for each method the standard deviation of the error distribution as a further metric:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}}, \tag{3.10}$$

where $e_i$ is the difference between the prediction and the ground truth and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$.

For the neural network models, the computations of the metrics were repeated 10 times each to ensure result consistency. Moreover, as a further robustness analysis, we conducted an RROC analysis [368] with all the algorithms (Figure 3.3(c)). All of these results are reported in the next section.

### 3.3.5 Feature relevance analysis

We used enrichment analysis, which is a statistical testing technique that computes the probability that a class of fluxes/genes is over-represented in a large set, to analyse and interpret our results (see Equation 2.4). For the fluxes, we performed hypergeometric tests using the MATLAB function *hygecdf*, and applied it to those fluxes to which the algorithms had attributed relevant weights (the threshold was chosen so as to reduce the number of fluxes to an easily interpretable amount). For the genes, we resorted to a different type of analysis since the lack of annotations for the transcriptomic data did not lead to meaningful results. The findings of these analyses are presented in Subsection 3.4.3.

We also examined the final models by inspecting directly the weights attributed to the input features, which represent the importance or contribution of the individual features to the model's performance. While this is straightforward with the RLMs, for the neural networks we developed a specific method in order to quantify the relevance that each feature had to the final prediction. To explain it, let us consider a neural network with a one-dimensional output and three hidden layers. Each node has a weight and a bias term, meaning that we can describe each layer in matrix notation with two matrices ($W$ and $B$, the matrices of the weights and the biases respectively). If we indicate

the input data as $X$ and the output as $o$, then it is possible to describe it mathematically in the following way:

$$\mathrm{o} = f(f(f(f(XW_1 + B_1)W_2 + B_2)W_3 + B_3)W_4 + B_o). \tag{3.11}$$

where $f$ is the non-linear activation function. Being almost all the activation functions currently used in research monotonic (included the ones used in the networks present in this chapter), and in view of the fact that only the relative importance of the features is of relevance for us, it is reasonable to ignore the functions and consider only the following expression

$$\mathrm{o} = (((XW_1 + B_1)W_2 + B_2)W_3 + B_3)W_4 + B_o, \tag{3.12}$$

from which, generalising, we can obtain that

$$\mathrm{o} = X \prod_{i=1}^{I} W_i + \sum_{j=1}^{I-1} B_j \prod_{k=j+1}^{I} W_k. \tag{3.13}$$

It is hence evident the fact that the weight determining the contribution of the input features to the model's output is just the product of the weights that each linked neuron possesses.

## 3.4 Results and Discussion

In this chapter, we aimed to determine whether GSMM-generated metabolic fluxes could be integrated with gene expression data through concatenation, by using RMLs and neural networks (Figure 3.1). As a case study, we focussed on predicting the growth rate of *S. cerevisiae* over a range of gene deletion strains. We used genome-scale metabolic models to extract metabolic information of yeast mutants in the exponential growth phase, employing transcriptomics information. We then evaluated how well different RLMs perform on a test set of 343 strains, and how accurately they recapitulate the findings already present in the literature.

### 3.4.1 Multi-omics prediction of cellular growth

We started from three state-of-the-art RLMs that were previously proposed for biological data analysis: Integrative Lasso with Penalty Factors (IPF-Lasso) [365], Group Lasso [367] and Principal Component Lasso (pcLasso) [366]. As described in Section 3.3, we then introduced Hybrid Group-IPF Lasso, which accounts both for different omic domains and intra-domain biological modules. Moreover, we considered the use of a modified regularisation term for IPF-Lasso and pcLasso (Section 3.3). Overall, we therefore tested the following RLMs: (i) IPF-Lasso, both L1 and L2, (ii) pcLasso, (iii) pc2Lasso, (iv) Group Lasso, (v) Hybrid Group-IPF Lasso. Together, they comprise different variants of group, view-specific and principal component regularisation (Figure 3.1). As a benchmark, we used artificial

Table 3.3: **Best hyperparameter values for the ANN on transcriptomic data**

| Hyperparameter | Value |
|---|---|
| batch size | 32 |
| epochs | 2400 |
| learning rate | $10^{-2}$ |
| no. neurons of first hidden layer | 3500 |
| no. neurons of second hidden layer | 4000 |
| optimiser | *RPROP* |
| dropout | 0.6 |
| loss | *Smooth_L*1 |

The number of neurons in the input layer (not a hyperparameter) was 6170, equal to the number of genes available in the dataset.

Table 3.4: **Best hyperparameter values for the ANN on fluxomic data**

| Hyperparameter | Value |
|---|---|
| batch size | 32 |
| epochs | 400 |
| learning rate | $10^{-5}$ |
| no. neurons of first hidden layer | 1200 |
| no. neurons of second hidden layer | 1800 |
| optimiser | *SGD* |
| dropout | 0.6 |
| loss | *Smooth_L*1 |

The number of neurons in the input layer (not a hyperparameter) was 459, obtained from the GSMM after removing the fluxes which were always zero.

neural networks and multi-modal artificial neural networks to better understand the advantages and drawbacks of using a less interpretable method with high predictive potential.

All the above methods (apart from the hybrid method) were tested over three datasets containing different types of information: (i) only fluxomic data; (ii) only transcriptomic data; (iii) fluxomic and transcriptomic data (in this case, the integration was accomplished through concatenation). This helped us understand better the contribution of the GSMM-generated data in terms of predictive performance and biological knowledge contributed to the models.

### 3.4.2 Comparison of multi-omics models of growth

Only the models with the best combination of hyperparameters were compared on the test set. In particular, we have reported the hyperparameters which were selected as the best combinations for the neural network models in Tables 3.3-3.4 (reproducibility for the other models is ensured by the original R functions that defined them).

Figure 3.3 and Table 3.5 provide a detailed overview of the results. It can be noted that the performance based only on reaction fluxes is considerably lower than the performance based on gene expression, consistently with previous results [255]. This is likely to indicate that fluxes, when used in isolation, have a smaller amount of relevant information for this task compared to transcriptomic data, thus they were compared only in a multi-view setting. In the next chapter we will see how

Table 3.5: **Multi-view results across all dataset-algorithm combinations**

| Data | Method | MSE ($\times 10^{-2}$) | MAE ($\times 10^{-2}$) | $R^2$ | $\sigma_e$ |
|---|---|---|---|---|---|
| | **Regularised Linear Models** | | | | |
| TRSC + FLUX | Group Lasso | 0.680 | 6.32 | 0.78 | 0.214 |
| | IPF-Lasso L1 | 0.577 | 5.76 | 0.81 | 0.212 |
| | IPF-Lasso L2 | **0.551** | **5.61** | **0.82** | **0.215** |
| | Hybrid Group | **0.570** | **5.75** | **0.81** | **0.213** |
| | pcLasso* | 0.812 | 6.70 | 0.73 | 0.206 |
| | pc2Lasso* | **0.702** | **6.29** | **0.77** | **0.209** |
| TRSC | Group Lasso | 0.558 | 5.65 | 0.82 | 0.219 |
| | IPF-Lasso L1 | 0.577 | 5.76 | 0.81 | 0.212 |
| | IPF-Lasso L2 | **0.544** | **5.61** | **0.82** | **0.216** |
| | pcLasso | 1.00 | 7.25 | 0.67 | 0.205 |
| | pc2Lasso | **0.837** | **6.68** | **0.72** | **0.207** |
| FLUX | Group Lasso | 1.74 | 9.70 | 0.43 | 0.206 |
| | IPF-Lasso L1 | 1.76 | 9.74 | 0.42 | 0.207 |
| | IPF-Lasso L2 | **1.76** | **9.75** | **0.42** | **0.207** |
| | pcLasso | 1.73 | 9.74 | 0.43 | 0.191 |
| | pc2Lasso | **1.72** | **9.72** | **0.43** | **0.191** |
| | **Artificial Neural Networks** | | | | |
| TRSC + FLUX | MMNN | 0.675 | 6.20 | 0.65 | 0.209 |
| | ANN | 0.640 | 6.02 | 0.70 | 0.214 |
| TRSC | MNNN | 8.48 | 6.98 | 0.72 | 0.182 |
| | ANN | 0.679 | 6.18 | 0.64 | 0.212 |
| FLUX | MMNN | 4.02 | 11.8 | -0.33 | 0.176 |
| | ANN | 1.70 | 9.23 | 0.13 | 0.205 |

Values in bold represent the scores of the methods proposed in this study. Asterisks indicate a statistically significant improvement for methods using TRSC and FLUX data over TRSC only, showing that some methods benefit more than others when fluxomic data are added to transcriptomic data as predictive features. The methods highlighted are the ones showing the best performance in each learning setting. The best performance is held by our modified version of IPF-Lasso with L2 penalty, which outperforms the other algorithms over almost all the comparison metrics.

different models and integration approaches can better exploit the metabolic information contained in the fluxomic data and use it to discriminate between phenotypic states depending on the patients' characteristics.

Amongst all the presented methods, only our proposed pc2Lasso managed to achieve an improvement in the performance when using more than one view, together with the original pcLasso, the MMNN and the ANN. Conversely, IPF-Lasso L1 fails to learn from the fluxes and the gene expression jointly. Specifically, its error scores remain unchanged when moving from one view to two, and a Wilcoxon signed-rank test run on the predicted and experimentally-measured growth rate distributions confirmed the overlap between their error distributions over the test set ($p = 0.19$). Moreover, a further confirmation of this is given by the weights that IPF-Lasso L1 attributes to the fluxes, which are all zeros.

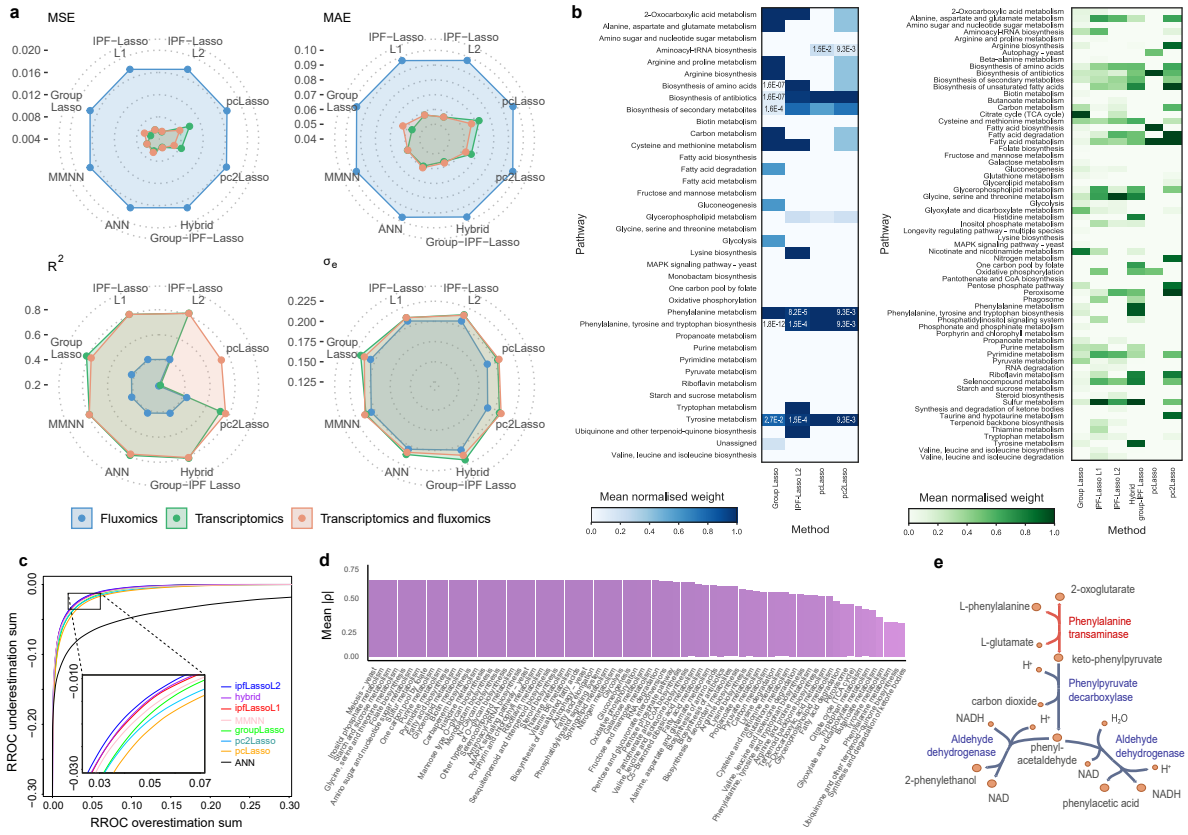



Figure 3.3: **Analyses results.** (a) Comparison of RLMs and MMNN across evaluation metrics and learning settings. The bigger the polygon drawn by the learning setting, the worse the results for MSE, MAE and $\sigma_e$, and the better for $R^2$. The fluxomic data alone do not perform well for all the metrics (except for $\sigma_e$). On the other hand, for some methods, combined learning with both transcriptomics and fluxomic data leads to better performance. (b) Average weight attributed to each of the related pathways according to the associated metabolic fluxes (left) and genes (right) for the regularised linear models. For better visualisation, we reported even the non-statistically significant pathways, and scaled the weights for each method separately. The statistically significant pathways are indicated by $p$-values (only for the metabolic fluxes). (c) RROC curves for the tested methods in the integration setting. Our IPF-Lasso and Group-IPF Lasso showed higher robustness than the other algorithms. (d) Mean absolute Pearson correlation along the pathways in the fluxomic dataset. The coefficients were computed by calculating the absolute values of the Pearson correlation between each metabolic flux and the growth rate, and then averaging them within each pathway. (e) Ehrlich pathway for the catabolism of phenylalanine. The reaction in red is amongst the ones selected by IPF-Lasso with L2-norm as a penalty. The main metabolites are represented by bigger circles.

Likewise, it is possible to gain some interesting insights by inspecting the weights (i.e. relative importance) that IPF-Lasso L2 gives the fluxes and the transcriptomic data, albeit the method does not show actual improvement. While the weights of the fluxes are all zeros, the weights of the genes are significantly different from the weights the algorithm attributes when trained only on transcriptomics, and an even smaller amount of them is selected. This could be interpreted as a particular indirect form of regularisation that reaction fluxes apply over the gene expression with this algorithm, which suggests that this multi-modal approach utilises profitably metabolic modelling to gain information that cannot be acquired from the transcriptomics alone. In Chapter 4 we will see how different models and integration techniques can still leverage the information provided by fluxomic data without displaying this regularising effect.

In general, the best-performing methods in the integration case (TRSC + FLUX), which adopt group and view-specific regularisation, do not display improved metrics over the TRSC case. On the other hand, methods employing principal component regularisation clearly display such improvement but remain with worse scores.

### 3.4.3 Interpretation of biological predictors

One of the purposes of adding a second view such as the metabolic fluxes was to improve the biological understanding and thus the interpretability of the input features, and consequently of the predicted output. The notion of interpretability we adopt here refers to the use of feature weights to establish which pathways/genes are important for yeast growth among the input features. From this perspective, we decided to look at the weights attributed to the metabolic fluxes by the algorithms and to conduct an enrichment analysis over the two different data types as a way to gauge the individual importance of the features.

Thanks to their transparent structure, RLMs can be interpreted immediately, as they directly assign a weight to each input feature. Since a typical characteristic of Lasso is the inner feature selection due to the fact that some input features are given zero as weight (which means that they are not used to make any prediction), all the RLMs share a similar property. Our analysis on the relevance of certain features takes thus into account solely the features that are not disregarded by the methods (i.e. with a non-zero weight). Figure 3.3b illustrates the outcome for the most common pathways that were found enriched for the RLMs. In the case of the genes, the pathways most present in the pool of the selected genes were considered.

Looking exclusively at the metabolic fluxes that were given the highest weights by each method, it is possible to cross-compare the algorithms. All the algorithms, except IPF-Lasso L1, Group Lasso and our hybrid Group-IPF Lasso, selected phenylalanine-involving reactions. Furthermore, all the al-

gorithms except Group Lasso selected tyrosine transaminase as a key reaction. It is widely known that in yeast these two compounds take part in the Ehrlich Pathway, which is directly related to fermentation. Moreover, different types of Lasso variations were capable of finding similar but not identical reactions, since while the two pcLasso versions found only one of the two PS decarboxylases reactions in the model, the two IPF-Lasso methods found the other one. Both these reactions have been found to support growth in *S. cerevisiae* [369]. Finally, phosphatidyl-L-serine and phosphatidylethanolamine were once again, like tyrosine transaminase, common to all but IPF-Lasso L1, Group Lasso and the hybrid method. The former is essential for cell growth [370], while the latter, under certain conditions, takes on crucial importance for yeast growth [371]. However, widespread differences were also found in terms of weight distribution across regularisation approaches.

To statistically evaluate such heterogeneity, we conducted pathway enrichment analyses on RLM weights, which indeed confirmed a varying use of biological information by individual regularisation strategies. Due to the diverse nature of the two types of data, the analyses were performed in different ways based on the dataset considered. In particular, we conducted a hypergeometric test to determine whether, among the fluxes deemed more relevant by the models, there were any metabolic pathways which were over-represented not by chance. In order to do this, we selected the most important fluxes for the predictions, i.e. the ones which were given the highest absolute weights by the models, and considered the associated pathways (as per the GSMM). Using Equation 2.4 and correcting for multiple hypothesis testing with the Benjamini-Hochberg procedure, we detected the pathways whose frequency, in the pool of relevant fluxes, was inexplicable by normal chance. As a result, in addition to the importance of phenylalanine, tyrosine and tryptophan, the enrichment highlighted the relevance of cysteine and methionine as previously known [372, 373], 2-oxocarboxylic acid and lysine when considering IPF-Lasso and aminoacyl-tRNA synthesis, arginine, alanine aspartate and glutamate when looking at the results from pcLasso and pc2Lasso (other results are reported in Table 3.6). Finally, we compared the fluxes and the related genes, i.e. the genes associated with the enzymes that catalyse each reaction selected by the algorithms, to evaluate whether there was a correspondence between them. Specifically, the genes associated with the reactions with the largest weights were considered and compared with the genes selected directly by the same method. The results showed that the genes associated with the selected reactions were not significantly present in the set of genes selected directly. This further strengthens our hypothesis that fluxes and genes carry qualitatively different information, potentially increasing the accuracy of multimodal methods compared to single-view ones.

In this chapter, we proposed and tested multimodal approaches with the intention of integrating information from metabolic models and experimentally obtained gene expression data. We showed that the metabolic information represented by model-derived flux rates is relevant for interpreting the predictions from machine learning models, and to better understand the interplay among genes, metabolism and growth in yeast. More specifically, we found that multi-omics data integration through principal component regularisation leads to predictive improvements in this setting, while other forms

Table 3.6: **Flux Enrichment Analyses for the regularised linear models**

| **Pathway** | **IPF-Lasso L1** | **IPF-Lasso L2** | **pc2Lasso** | **Group Lasso** |
|---|---|---|---|---|
| Phenylalanine, tyrosine and tryptophan biosynthesis | $1.33 \cdot 10^{-5}$ | $1.52 \cdot 10^{-4}$ | $9.30 \cdot 10^{-3}$ | $1.79 \cdot 10^{-12}$ |
| Phenylalanine metabolism | $1.79 \cdot 10^{-2}$ | $8.21 \cdot 10^{-8}$ | $9.30 \cdot 10^{-3}$ | |
| Tyrosine metabolism | $4.71 \cdot 10^{-2}$ | $1.52 \cdot 10^{-4}$ | $9.30 \cdot 10^{-3}$ | $2.74 \cdot 10^{-2}$ |
| Biosynthesis of amino acids | $9.68 \cdot 10^{-4}$ | | | $1.62 \cdot 10^{-7}$ |
| Biosynthesis of antibiotics | $3.90 \cdot 10^{-3}$ | | | $1.62 \cdot 10^{-7}$ |
| Biosynthesis of secondary metabolites | $3.90 \cdot 10^{-3}$ | | | $1.58 \cdot 10^{-4}$ |
| Cysteine and methionine metabolism | $1.44 \cdot 10^{-2}$ | | | |
| Aminoacyl-t RNA biosynthesis | | | $9.30 \cdot 10^{-3}$ | |
| 2-Oxocarboxylic acid metabolism | $1.45 \cdot 10^{-2}$ | | | |
| Lysine biosynthesis | $1.45 \cdot 10^{-2}$ | | | |

For each method we display the $p$-value associated to the pathway found (when present). As it can be noticed, phenylalanine- and tyrosine-related pathways are common to almost all the methods. All the $p$-values are below the defined threshold of 0.05. The results for pcLasso and the hybrid Group-IPF Lasso are not shown since the only enriched pathway for the former was the *Aminoacyl-t RNA biosynthesis*, with a $p$-value of $1.50 \cdot 10^{-2}$, while the latter was enriched only in *Valine, leucine and isoleucine biosynthesis* with a $p$-value of $2.06 \cdot 10^{-2}$.

of regularisation appear less effective in such task. While the metabolic fluxes were calculated through pFBA, it must be noted that alternative methods could potentially be used to compute flux rates, which may further improve predictive results. This is the case of Chapter 4, in which we use Flux Variability Analysis to compute the metabolic fluxes [94], but also of Chapter 2, where we adopted TRFBA to inject regulatory knowledge into our GSMM [324].

We found that regularised linear models can be a better choice than neural networks (even after extensive fine-tuning of the hyperparameters), as they have comparable performance (RLMs perform slightly better than neural networks, but with no statistically significant difference at significance threshold = 0.05) but require much shorter training times and allow for an immediate and natural interpretation of their result (unlike ANNs, which are black boxes that require more complex interpretation strategies often based on unintuitive assumptions). This suggests that powerful methods such as neural networks cannot be regarded as off-the-shelf methods to which to resort for any task indiscriminately, and other simpler methods should also be considered. As a further confirmation for this, we will show in the next chapter how more classical machine learning algorithms can perform at least as well as some deep learning models.

## 3.5 Conclusion and future directions

In this chapter, we investigated the potential of existing and novel multimodal regularised linear models in predicting *Saccharomyces cerevisiae* growth using experimental and metabolic model-derived multi-omic data. Our experiments included state-of-the-art regularisation methods such as group-based, view-specific and principal component regularisations. These were applied to a combination of genome-wide gene expression data and model-generated metabolic flux rates. We found that, in this setting, linear interpretable methods such as variations of Lasso can be preferable to artificial neural networks even on a relatively large dataset as, being the performance equivalent, they are much faster to train and much easier to interpret.

There are several limitations to this study which should be addressed. First, when optimising the neural networks, we conducted a random search over eight hyperparameters for over 100 iterations. This meant that we trained (and tested on the validation dataset) over 100 different models. This approach, necessary in this case given the high complexity of the models, is very prone to a type of overfitting called "optimisation bias", which could have caused the training pipeline to produce suboptimal, underperforming models. Indeed, by repeatedly testing on the validation dataset and comparing so many models, we would have risked to overfit it, thus hampering the models' ability to generalise on unseen, new data. Another issue consists of the impossibility of running additional statistical analyses on some biological terms of the genes selected by the models (which instead we did for the reactions). This was due to lack of annotations, which forced us to limit our enrichment analyses to the pathways deriving from the metabolic fluxes only, and to adopt less mathematically robust analyses for the genes. We also observed that some accurate state-of-the-art regularisation methods conceived for data integration fail in achieving accuracy improvements in our multi-omics setting, and that they inconsistently point to different sets of relevant biological variables. These findings highlight the need for new, more powerful solutions that can exploit the cross-modal information, in addition to the information held in the individual modalities. In the next chapter, we will go on to present a more advanced integrative pipeline which does not use raw feature concatenation to leverage the information contained in the different omics.

## 3.6 Related work, funding and final remarks

The work presented in this chapter has been published in Bioinformatics [282], and the paper was written in collaboration with multiple co-authors. As described in the introduction to this chapter, I performed all the analyses relating to the generation of the metabolic features, I devised, implemented, trained and tested the models, and I conducted the analysis and interpretation of the results as well.

This work was supported by the UKRI Research England's THYME project and by a Children's Liver Disease Foundation Research Grant.

# Chapter 4

# Integrating transcriptomic and fluxomic data with a late integration strategy for liver cancer diagnosis

## 4.1 Introduction

In the last chapter, we continued on our journey by investigating whether the novel information contained in genome-scale metabolic fluxes could be used in conjunction with gene expression data in a precision medicine setting, and we discovered that this depends heavily on the type of machine learning model used. In this final chapter, we conclude our quest by trying to answer these last few questions:

- How else can we exploit the information present in GSSM-generated fluxes together with the information held by gene expression data? Is it possible to obtain better integration results without sacrificing interpretability?
- When we have more than two omics, how do different omics combinations interplay with the data samples? For predictive purposes, are there situations where certain combinations are better than others?

As in the previous chapters, we will use a case study to examine them, but this time, unlike in the last instance, the presented task will finally be a precision medicine application with human data samples (and not single-celled organisms). The contributions of this chapter are the following: we show that more complex integrative approaches for GSMM-generated metabolic fluxes, gene expression data and clinical data are possible, and demonstrate that different combinations of omics allow for the detection of different genes, reactions and metabolic pathways associated to the model's prediction. Finally, we also show that certain omics combinations are more suitable for prediction than others, depending on the patients' clinical characteristics.

The presented case study is a typical binary classification task, in which we are trying to predict whether a patient has cancer or not starting from their metabolic fluxes, gene expression data and clinical features. The entirety of the study was conducted by me. This means that I performed all the analyses relating to the generation of the metabolic features, I designed, implemented and tested the entire machine learning pipeline, and I conducted the analysis and interpretation of the results as well.

## 4.2 Background

Hepatoblastoma is the most frequent epithelial liver tumour in infancy and childhood, with over 90% of cases diagnosed earlier than 4 years of age. This tumour is characterised by a high recurrence rate and metastatic aggressiveness, especially below this threshold age [374], which makes it paramount to be able to obtain an accurate prediction early on in the onset of the disease. Additionally, its incidence is increasing in several developed countries. The recent development of molecular methods allowed extending the general subtype classification of primary childhood liver cancers, including hepatoblastoma [375, 376], whose heterogeneity complicates the diagnosis of the disease. Moreover, clinical studies suggest that biomolecular mechanisms are associated with diverse prognostic outcomes and chemotherapy responses. Very recently, a few studies have started to explore the biological variability underlying hepatoblastoma, focusing on genomic biomarkers [377]. Likewise, machine learning has been adopted in the study of hepatoblastoma with encouraging results [378, 379, 380]. However, the role of metabolic rewiring – which is one of the main hallmarks of tumour cells [381] - has not been studied so far in hepatoblastoma. As a result, there is a general lack of robust biomarkers for this disease [382].

Following the path delineated by the previous case studies, in this chapter we investigate how different omics (and their combinations), interplaying with the patient's characteristics, affect the accuracy of a machine learning-based diagnosis by using a systems biology framework (already adopted with success in cancer research [383, 384, 385]) in conjunction with machine learning. Even though transcriptomics cannot be easily outperformed by other omic data [386], we have shown in the previous

chapters that GSMM-generated metabolic fluxes do contain different information that could be used to improve models' predictions. We also examine metabolic markers for hepatoblastoma in the hope that this will guide future research in the field. In particular, we study how experimentally measured gene expression plays a role in diagnosing hepatoblastoma when paired with both synthetic *in silico* metabolic data and clinical data such as gender and age of the patient.

Starting from a set of transcriptomic profiles, we use genome-scale metabolic modelling to estimate the associated metabolic activity across pathways in a precision medicine fashion. We then use support vector machines [387] as a predictor to identify hidden patterns that discriminate between phenotypic groups, and compare the performance of the different omics and their combinations, achieved by integrating the omics via Partial Least Squares Discriminant Analysis (PLSDA), in four alternative scenarios. For each scenario, we examine and present potential biomarkers, validating them against the existing literature. We report how specific omics combinations can be beneficial to the diagnosis of hepatoblastoma in different patients, and that the predictive power of each combination varies with their age, gender and clinical status.

## 4.3 Materials and Methods

### 4.3.1 Data gathering and homogenisation

We gathered relevant transcriptomic data from liver samples of children diagnosed with hepatoblastoma and for control subjects within the same age range sets [377, 388]. We selected three datasets whose gene expression profiles and clinical information have been retrieved from the Gene Expression Omnibus portal (`www.ncbi.nlm.nih.gov/gds`) under the accession numbers GSE75271, GSE131329 and from the BioStudies ArrayExpress portal (`https://www.ebi.ac.uk/biostudies/arrayexpress`) under the accession code E-MEXP-1851. The selection of these datasets considered the experimental platform utilised and, given the need for numerous samples to train a machine learning model, we prioritised the platform with the most abundant publicly-available data, which was in our case the Affymetrix microarray (Affymetrix Human Genome U133 Plus 2.0, Affymetrix Human Gene 1.0 ST and Affymetrix HG-U133A 2.0 GeneChipTM respectively). The gathered data comprise a total of 151 subjects including 128 hepatoblastoma patients and 23 controls (see Figure 4.1B). The average age is 2.6 years and there are 84 and 67 male and female subjects, respectively. The transcriptomic profiles cover 12,712 genes.

The pre-processing of data is fundamental to their meaningful analysis, free of technical biases [389, 390, 391]. In addition, their joint analysis required ensuring uniformity and batch effect removal, which we performed through ComBat [171]. ComBat is an empirical bayes-based pipeline which

consists of a gene-wise standardisation (via a gene-wise least squares regression) of the data followed by the estimation of the additive and multiplicative batch effect parameters through empirical priors (meaning that the priors are directly estimated from unbiased, i.e. standardised, data, and not assumed before the observation). Once the parameters have been estimated, the data can be corrected by dividing for the multiplicative batch parameter and subtracting by the additive one. Figure 4.2A shows the effect of homogenisation across the three datasets. Apart from removing the batch effect, we also cleaned the clinical data to guarantee consistency of labels and values across the three datasets. In particular, the age was rescaled so that it would represent, in each of the three datasets, the age of the patient in months, while for the gender and clinical status information common labels were chosen to make the patients immediately comparable across the datasets. Due to heterogeneous clinical formats, however, some information remained sparse, such as race, tumour stage and clinical course (Figure 4.1B), and could not be used. For this reason, the analysis of clinical data within the machine learning models adopted was limited to age, gender and clinical status information.

### 4.3.2 Patient-specific metabolic modelling of hepatoblastoma

To obtain metabolic information tailored to patient-specific metabolism, we adopted a GSMM approach. The base requirement is a mathematical representation of all the known biochemical reactions and transmembrane transporters present in an organism. Previous work has been done with GSMM to mechanistically characterise various human disorders, including liver diseases [392] and a range of cancer types [268, 393, 394, 395]. As we have already seen, GSMMs can be integrated with omics data to obtain context-specific models, representing the metabolic status across various conditions or tissues [264, 396]. Notably, tissue- and cell-specific metabolic models have been successfully used to identify, and successively validate, specific drug targets that inhibit cancer proliferation but do not affect normal cell proliferation [397, 398]. Through the mathematical representation of metabolic networks, GSMM can provide mechanistic insights regarding how hepatoblastoma works, with both the biochemical detail and completeness to interpret large molecular datasets.

**Transcriptomics data integration** In our experiments, the human metabolic reconstruction Recon2.2 [70] was used in order to estimate the metabolic activity associated with transcriptional cues in tumour and control liver. Following a precision medicine approach, we derived a different metabolic model for each patient [87, 56]. In doing so, we mapped the gene expression levels of the patients onto the metabolic network, thus determining the metabolic conditions from which to infer the reactions activity for each individual. Specifically, this process uses gene-reaction relationships encoded within Recon2.2 and generates sample-specific constraints that describe the maximal and minimal activity

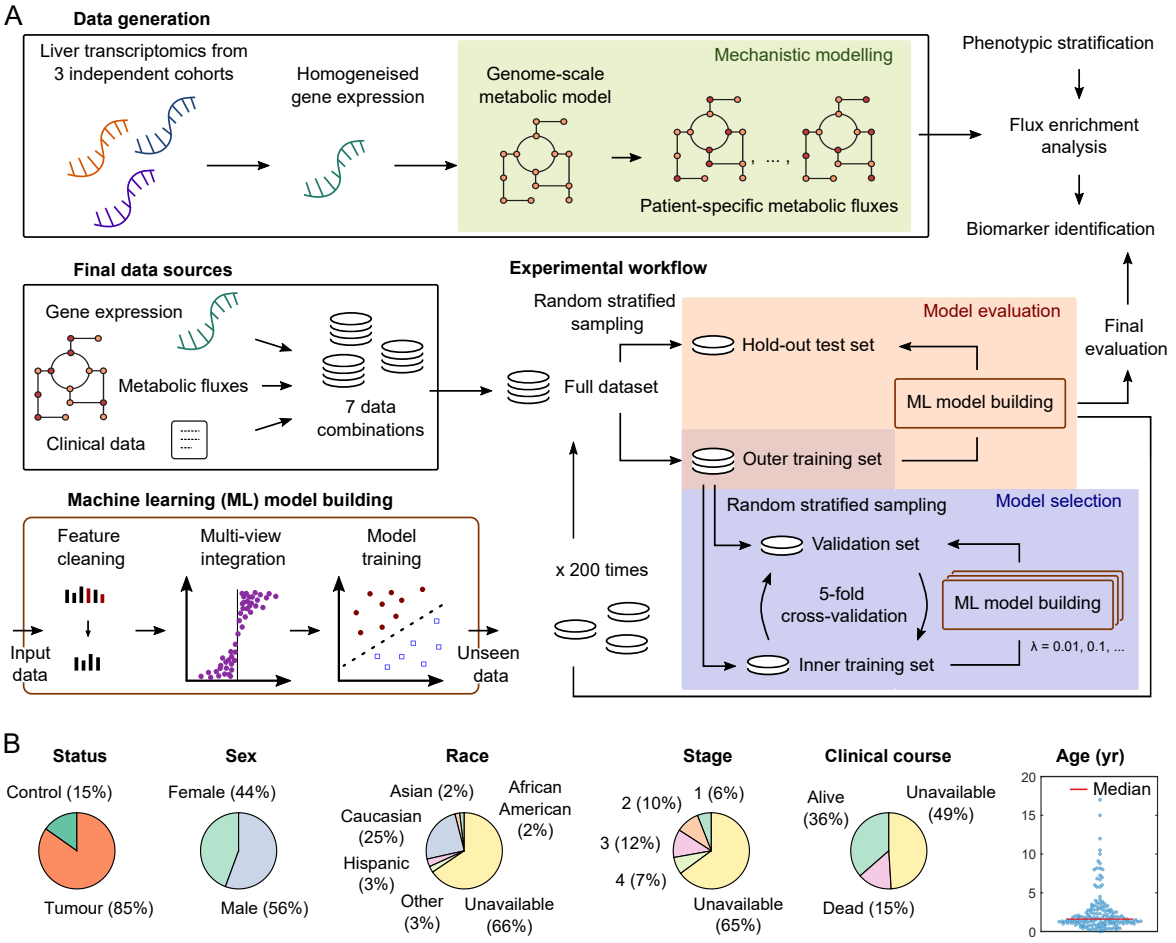



Figure 4.1: **Experimental pipeline and clinical data analysis.** A. Multi-omics and machine learning pipeline adopted in this chapter. Starting from liver gene expression profiles for hepatoblastoma patients and control subjects, we computed the variability of metabolic fluxes via FVA. For each of the combinations of transcriptomic, metabolic, and clinical data, we then performed a random stratified sampling to obtain a hold-out test set and an outer training set for machine learning model evaluation. Starting from this training set, we conducted a 5-fold cross-validation across hyperparameter values, and then evaluated the best model on the hold-out test set. Within each round of cross-validation, we also performed feature standardisation and cleaning and omics integration when necessary, in order to avoid any data leakage (brown box). We repeated the entire procedure 200 times to ensure the robustness of the results and re-ran the entire pipeline with a randomised dataset, whose phenotypes had been randomly permuted, so as to verify that the learned models correctly identified biologically meaningful patterns. B. Clinical data for the combined dataset used in the study. Race, tumour stage and clinical course had widespread missing entries, due to the original datasets having different information available, and thus were removed from most analyses.

that can be sustained by a given transcriptional state:

$$\begin{aligned} \mathbf{v}_{ub} &\leftarrow \mathbf{v}_{ub}\,[1+\gamma|\log\Theta|]^{\text{sign}(\Theta-1)} \\ \mathbf{v}_{lb} &\leftarrow \mathbf{v}_{lb}\,[1+\gamma|\log\Theta|]^{\text{sign}(\Theta-1)} \end{aligned} \tag{4.1}$$

where $\mathbf{v}_{ub}$ and $\mathbf{v}_{lb}$ represent the upper and lower bounds of the metabolic fluxes respectively, while $\Theta$ represents the gene expression level of the gene sets present in the genome-scale metabolic model, and $\gamma = 2$ (see Equations 1.2). The intuition behind this formulation is the following: in order to prevent the flux balance algorithm being influenced by extremely high values in the gene expression, thus generating unrealistic flux distributions, we adopted a logarithmic mapping to curb down the most extreme measurements in our transcriptomic data. The use of a logarithmic mapping is also consistent with the dynamics of the protein translation rate, which is almost linearly correlated to mRNA abundance for relatively small values but quickly becomes approximately constant as the abundance becomes high [399, 400]. $\gamma$ is a multiplicative factor representing the reliability of $\Theta$ as an indicator of the activity level of the associated reaction. Finally, the sign function has the purpose of taking into account the magnitude of $\Theta$ in the multiplication: if $\Theta$ is small, the reaction bounds are divided by the quantity in the square brackets, otherwise they are multiplied by it (thus correctly influencing the activity of the reaction).

Furthermore, we imposed additional experimental constraints (see Table 4.1) directly onto the genome-scale metabolic model, which are orthogonal to those given by gene expression. To this end, we performed a literature search on liver metabolism, collecting experimentally supported bounds to metabolic exchanges in the liver. In particular, we followed previous work on hepatocyte modelling [392], correcting for the modelling convention according to which exchange reactions that assume uptakes are represented by negative lower bounds. These secretion and uptake rates were taken from previous measurements [401], which investigated the changes in intracellular pathway fluxes of primary rat hepatocytes in response to low-insulin preconditioning and amino acid supplementation. Among the involved reactions, we set uptake bounds for glucose, glutamate and glutamine.

We performed these steps for all the 151 samples in our dataset, in parallel, thus obtaining 151 context-specific metabolic models, each associated with a specific individual.

**Flux variability analysis** To quantify the genome-scale metabolic state associated with collected transcriptomic profiles, we adopted flux variability analysis, which provides complete maximal (and minimal) cell metabolic capabilities across the biochemical network [94]. FVA operates by sequential maximisation and minimisation of each reaction activity to explore the boundaries of the feasible flux space. This algorithm yields a profile of maximal and minimal reaction rates (fluxes) for every biochemical reaction in Recon2.2, which collectively constitute a fluxomic profile. However, unlike the transcript levels, these metabolic fluxes do not belong to a single metabolic state, rather they represent the metabolic capabilities and limits of the individual's metabolic network, because the reactions are

Table 4.1: **Experimental values used to constrain the model**

| **Reaction** | **Reaction Lower Bound** |
|---|---|
| D-Glucose exchange | 2.025 |
| L-histidine exchange | −0.04425 |
| L-Isoleucine exchange | −0.0585 |
| L-Leucine exchange | −0.0825 |
| L-Lysine exchange | −0.2325 |
| L-Methionine exchange | −0.12 |
| L-phenylalanine exchange | −0.202774 |
| L-Threonine exchange | −0.12 |
| L-Tryptophan exchange | −0.0075 |
| L-Valine exchange | −0.04125 |
| H2O exchange | 25.3228 |
| O2 exchange | −28.05 |
| CO2 exchange | 21.7219 |
| L-alanine exchange | −0.02325 |
| L-asparagine exchange | −0.00135 |
| L-glutamine exchange | −2.325 |
| L-Tyrosine exchange | −0.05775 |
| L-cysteine exchange | −0.0555 |
| L-Arginine exchange | −0.2175 |
| Glycine exchange | −0.2625 |
| L-Proline exchange | 0.02925 |
| L-serine exchange | −0.1425 |
| L-Aspartate exchange | 0.00825 |
| L-Glutamate exchange | 0.15 |
| Ammonia exchange | −0.165 |
| Sulphate exchange | 0.16121 |
| Proton exchange | −0.42825 |
| Glycerol exchange | −6.675 |
| Ornithine exchange | 0.125 |
| Acetoacetate exchange | 0.1275 |
| (R)-3-Hydroxybutanoate transport via H+ symport | 0.05775 |
| L-Lactate exchange | −0.063 |
| Urea exchange | 3.375 |

The values were corrected by changing their sign, according to the convention by which lower bounds for exchange reactions are negative when the reaction admits uptakes.

maximised and minimised in turn. The reason why we computed a fluxomic profile for each individual is, like in the previous chapter, that we assume it should be possible to distinguish between healthy and cancer cells by looking at their metabolism. The optimisation problem was the following:

$$\begin{aligned} &\max\ (\min)\ v_i \\ &\text{subject to} \quad \mathbf{c}^\top \mathbf{v} = f_{max}, \\ &\qquad\qquad\ \mathbf{S}\,\mathbf{v} = 0, \\ &\qquad\qquad\ \mathbf{v}_{lb} \leq \mathbf{v} \leq \mathbf{v}_{ub}, \quad \text{for } i = 1, 2, \ldots, n, \end{aligned} \tag{4.2}$$

where $\mathbf{S}$ is the stoichiometric matrix that defines the chemical reactions present in the metabolic model, $\mathbf{c}$ is a vector for characterising the objective function $f$ (the biomass in our case) starting from $\mathbf{v}$, $f_{max}$ is the maximum value of $f$, and $\mathbf{v}_{lb}, \mathbf{v}_{ub}$ are the lower and upper bounds, respectively, of the metabolic reactions, as per Eq 4.1.

As an alternative to FVA, we also used, as in Chapter 3, pFBA [292], following recent advances on objective functions in mammalian metabolic modelling [85]. This approach, however, involves the adoption of specific cellular objectives that in this case did not provide sufficiently diversified metabolic profiles across all samples, which prompted us to employ FVA as it provides more unbiased estimates of metabolic variation across individuals. This could be explained by the fact that cancer cells present complex behaviour which may not be easily modelled with a single optimisation objective [402]. Exploratory results with this FBA variant (in which we used as objective function the maximisation of the biomass, to simulate the uncontrolled growth of cancer cells) are reported in the following sections together with the results for the FVA-generated metabolic fluxes.

The COBRA Toolbox [334] was used with the Gurobi solver to compute the metabolic fluxes in MATLAB R2021b.

### 4.3.3 Biomarker identification framework

The study was divided into two parts: we first analysed the metabolism of the patients with respect to the possible stratifications in the population, and then applied machine learning techniques to determine possible biomarkers and how different omics could affect the precision of diagnosis of hepatoblastoma. We decided to follow this two-fold approach (flux-based metabolic analysis first and machine learning-led knowledge discovery after) as this is the most promising for the delivery of robust biomarker insights. Conversely, enrichment by itself does not guarantee predictive power nor does it help prioritise candidate biomarkers for future studies [403].

**Flux enrichment analysis** To determine whether the over-represented pathways associated with the resulting metabolic reactions in the pool were overly present "by chance" or because there exist

some real biological mechanisms linked to the reactions, we decided to run a Flux Enrichment Analysis, which is a statistical testing technique that tests for the statistical relevance of biological pathways associated with a pool of reactions (see Equation 2.4 for a mathematical definition).

Before applying FEA, we removed all the reactions which had an absolute flux lower than 1e-7, considering them non-active, to account for the tolerance of the FVA solver. All the other reactions were instead included in the analysis. FEA was conducted on all the samples, in a stratified and non-stratified way, by using hypergeometric tests, and the Benjamini-Hochberg correction was used to take into account the multiple hypothesis testing scenario. We set 0.05 as a threshold for the $p$-value to determine whether the presence of an over-represented pathway was statistically significant or not. Whenever specific covariate information was not available for a sample, we discarded the sample for that stratification and conducted the analysis on the remaining data.

We followed this approach because we were interested in assessing whether different groups of individuals (healthy/ill, male/female, etc ...) showed changes in metabolic activity highly concentrated in specific pathways. The different cohorts were based on the available covariates and were organised as follows: tumour – control; male – female; older – younger than 4.5 years; older – younger than 3 years; alive – dead. The choice regarding the thresholds for the age was driven by the need for a deeper granularity in the analysis within the range [3, 4.5], which is considered to be critical to the diagnosis of the disease [374].

**Machine learning-led biomarker discovery** Support vector machines are machine learning models that can be trained to distinguish samples belonging to different groups, such as patients and control individuals [387]. Here, we trained and applied SVM models to identify predictive variables that best discriminate between phenotypic groups (tumour and control). Once identified, these variables could thus be regarded as biomarkers. The objective function for the training of our SVMs was the following:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \lambda \sum_i \max\Big(0, 1 - y_i(\mathbf{w}^\top x_i + b)\Big), \tag{4.3}$$

where $\lambda$ is a regularisation hyperparameter to optimise, $\mathbf{w}$ and $b$, respectively, weights and bias of the model, and $(x_i, y_i)$ the pair (features, class) of the $i$-th sample. In addition to SVMs, we also tested another machine learning algorithm, Random Forest (RF) [404], and a Neural Network (NN), a deep learning approach that usually achieves state-of-the-art performance in many modern artificial intelligence tasks (we used a similar architecture to the ANN from Chapter 3, but with only one layer). The performance of the three models was compared and we found out that the SVM model performed better than the other models or equally well in all the studied scenarios. Further information can be found in the next section. In the rest of the chapter, however, we decided to focus on the results of the SVMs only also because the SVM algorithm is computationally inexpensive if compared, for instance, with the NNs, and therefore of more practical use for real applications.

We adopted PLSDA as integrative approach in order to mitigate the problems deriving from the high dimensionality of the data combined with the small number of samples. In particular, the omics (transcriptomic and fluxomic) were projected onto two-dimensional spaces (one dimension per phenotypic trait; each omic was projected onto an independent space) in the explored integrative settings, explained below.

Our general training-evaluation pipeline, as reported in Figure 4.1A, was the following: starting from the complete sample set (151 samples), we performed a random stratified sampling of 10 samples (5 patients and 5 controls) to put aside as a test set. The remaining samples were used as training data for an SVM model with a linear kernel, which we then employed to predict the phenotypic group for the 10 hold-out samples. This train-test process was repeated on random data partitions 200 times in order to ensure the robustness of the results, given such a small test set. This was done in two ways: by looking at the performance distribution of our models (described by 200 points), and by summing the weight each feature was given in each of the 200 runs. This is equivalent to computing the average weight per feature, as we have considered the weights not in absolute terms, but in relation to each other (a feature is more important than another if it has higher absolute weight). The exact number of iterations was a result of a trial-and-error procedure, through which we determined that a lower number of repetitions would increase the standard deviation of the performance distributions (thus making our results less robust), while a higher number would simply increase the duration of the experiments, with negligible gains in terms of results robustness. Given the over-representation of tumour samples (see Figure 4.1B), at each iteration we employed random under-sampling of tumour samples and over-sampling of control samples in order to obtain 30 samples for both groups (60 samples in total). We did this after the generation of the hold-out test sets, to avoid any data leakage that could affect the robustness of our pipeline. In other words, we randomly sampled, in a stratified fashion, 30+30 samples out of the 141 samples which did not belong to the test set. During the model building stage, we also performed feature standardisation and hyperparameter optimisation of $\lambda$ through grid search. This, together with the under- and over-sampling of the 60 samples described above, was conducted within a 5-fold cross-validation framework on the remaining 141 samples, thus controlling for overfitting. The optimisation procedure for $\lambda$ was selected for its robustness, but alternative approaches are possible. For instance, several meta-heuristics have been developed recently based on animal group behaviour and particle dynamics [405, 406]. Such algorithms have previously been applied in combination with metabolic modelling [265], and it has been shown that they can be beneficial when optimising hyperparameters of SVM [407]. In this work, however, we opted for a more standard procedure that was applicable to all the investigated machine learning models.

Moreover, we performed feature selection by removing all the constant variables and the ones which were identical to others (in the case of fluxomic data, for instance, reactions in a pathway with a locally linear topology could share the same value at all times). The feature selection procedure was itself performed within the cross-validation framework, in order to avert any overly optimistic performance

evaluation of the SVM models during the hyperparameter optimisation.

We conducted these experiments in 6 different scenarios, with the aim of investigating how different combinations of omic data would influence the predictive power of the SVM models and their sensitivity to different biological entities (genes, fluxes, pathways): (i) use of transcriptomic data only; (ii) use of fluxomic data only; (iii) use of transcriptomic and clinical data (age, gender); (iv) use of fluxomic and clinical data; (v) use of transcriptomic and fluxomic data; (vi) use of transcriptomic, fluxomic and clinical data. To the above scenarios, we added also a final setting in which we trained the SVM models only with the clinical data, in order to eradicate any possible bias caused by the collection of the data (*sampling bias*).

To verify that the learned models correctly identified biologically meaningful patterns, we tested (through the same evaluation process) SVM models built starting from a permuted version of the dataset [408]. Specifically, we performed an additional 200 test iterations while randomly reassigning phenotypic labels to each sample prior to conducting the cross-validation, as previously suggested [409]. We did this for each of the 6+1 scenarios described above for completeness of the analysis.

Since we wanted to investigate how the discriminative power and sensitivity to biological mechanisms would change with different omics integrations, we decided to analyse the weights assigned by the SVMs to each variable during training, with the rationale that a higher weight in absolute value corresponds to a higher relevance. For the integrative experiments, the weights were computed by projecting the weights attributed to the latent dimensions back onto the original feature space. Moreover, in order to have a broader picture of the main metabolic pathways detected in the four integrative scenarios, we conducted FEA in each of them. For each scenario, we selected only the fluxes whose weight was in the 99.5th percentile. 0.05 was set as the threshold value for significance.

All the analyses were conducted in python, and the SVM, RF, NN and PLSDA algorithms were implemented with the library scikit-learn [410].

## 4.4 Results and Discussion

The scope of this study was to investigate how different omics and their combinations may contribute to a computer-aided diagnosis of hepatoblastoma both in terms of accuracy and understanding of the biological mechanisms underlying the disease. In this framework, we focused on the use of individuals' transcriptomes and model-generated fluxomic profiles in order to capture the metabolic alterations associated with the disease, in a precision medicine fashion. These omics readouts were integrated and used to build predictive models through the machine learning pipeline displayed in Figure 4.1A.

### 4.4.1 Genome-scale model characterisation of hepatoblastoma metabolism

Following a condition-specific modelling approach, we estimated the metabolic activity differences associated with varying transcriptional patterns across individuals. A genome-scale stoichiometric model of human metabolism was used as a platform for gene expression profiles obtained from three independent cohorts of individuals. As a result, we obtained maximal and minimal rates achievable through each biochemical reaction present in the model under the given transcriptional states. Figure 4.2B shows a principal component analysis of transcriptomic and fluxomic (maximal fluxes only) profiles. In both cases, hepatoblastoma patients and healthy controls display an almost linear separation. From a machine learning standpoint, this suggested that patient phenotypic classification could be achieved with high accuracy even with a limited number of samples. Indeed, being PCA a linear transformation, the fact that the transformed data were linearly separable entailed that the original data were linearly separable as well, which considerably simplified the solution of the task as simpler models could be used. On the other hand, PCA revealed no obvious global relationship between subject age and multi-omics variation. An alternative graphical representation of Figure 4.2B, in which age is replaced by gender, can be found in Figure 4.3 instead.

The metabolic flux variation can be decomposed into metabolic capabilities across the pathways in the tumour and control groups described above, illustrated in Figure 4.2C. From the figure, generated from the maximal fluxes, it is possible to observe a widespread reduced activity in several pathways associated with hepatoblastoma, such as in the central metabolism, nucleotide salvage and interconversion. However, up-regulation was found in glutathione and CoA metabolism.

We then used FEA to obtain a picture of the most relevant metabolic pathways for groups of individuals defined based on their health status, sex, age, and clinical course. When doing so, FEA showed several statistically significant differences among the chosen cohorts (alive - dead, younger - older than 3 years, younger - older than 4.5 years, male-female, tumour - control). FEA computed over the maximal reaction fluxes generated by FVA returned several statistically significant differences. When considering the forward reaction direction, all the enrichments had in common many relevant pathways, since the reaction rates were generally higher, which meant many more active reactions in the metabolism (Figure 4.2D). The generally enriched pathways were the ones associated with extracellular transport and nucleotide interconversion (as in the cases above), ubiquinone synthesis and keratan and cholesterol metabolism. The only exception to this was the alive-dead contrast, which did not present the reactions associated with the keratan sulphate synthesis. In all three cases, no significant differences were found across age groups (older – younger than 4.5 years and older – younger than 3 years), which probably indicates that within this age range there are no specific metabolic changes. When considering the backward direction of reversible reactions, all the enrichments had in common the reactions associated with extracellular transport and nucleotide interconversion as the most relevant, while the citric acid cycle and the nuclear transport reactions were not critical for the

tumour-control comparison, unlike the other stratifications. Moreover, the two age stratifications and the alive-dead contrast showed as important reactions the ones related to the metabolism of valine, leucine and isoleucine. The FEA conducted on the pFBA-generated metabolic fluxes displayed as statistically relevant for all the stratifications the reactions associated with nucleotide interconversion and glutamate metabolism. Furthermore, extracellular transport reactions were the most relevant, again for all the stratifications. All the enrichments but the one associated with the tumour–control stratification showed the importance of the reactions composing the citric acid cycle as well. Given that there was very little differentiation between the enrichments for the pFBA fluxes compared to the FVA-generated ones, as already stated, we decided to focus on the FVA-generated fluxes in the rest of the analysis.

### 4.4.2 Biochemical marker identification

The analyses above could identify changes in metabolic activity associated with a range of subject sub-cohorts. To understand which changes can be more strictly linked to carcinogenesis, we adopted machine learning techniques. Since the maximal fluxes presented more diversified metabolic profiles, we decided to focus only on them for the rest of the study. Figure 4.12A shows the classification results obtained for all the 6+1 omics combinations studied, including gene expression, metabolic fluxes, and clinical information. On average, SVM models achieved a mean accuracy and MCC of around 0.9, with the exception of models trained only on clinical data. In contrast, SVM models obtained from permuted versions of the dataset on average proved no better than a random model for all the omics integrations, with a mean accuracy close to 0.6 and a mean MCC around 0.1, and with a standard deviation much larger than the one of the models trained on the original dataset. This indicates that the original models captured meaningful patterns underlying the clinical state of the subjects, as expected. Since in this study we adopted, in addition to the SVM models, other two machine and deep learning models, namely Random Forests and Neural Networks, we compared their performances for each omics combination by using Wilcoxon signed-rank test. This statistical test was used to determine whether the accuracies and MCCs of the models belonged to an identical distribution or not. In particular, only the RF models presented performance distributions that were determined as different from the SVMs' ones, and exclusively when the models were trained with transcriptomic data only. In this case, the RF models performed slightly worse than the SVM models (see Figures 4.4-4.5).

As it can be seen from Figures 4.4-4.5, the performance of the different models does not change significantly between omics combinations: this is likely due to the apparent linear separability of the data which can be deduced from Figure 4.2B. However, as we will see in the next Subsection, different omics combinations can actually affect diagnostical performance on different groups of people, based on their clinical characteristics (namely, age, gender and health status). In order to gain some insights

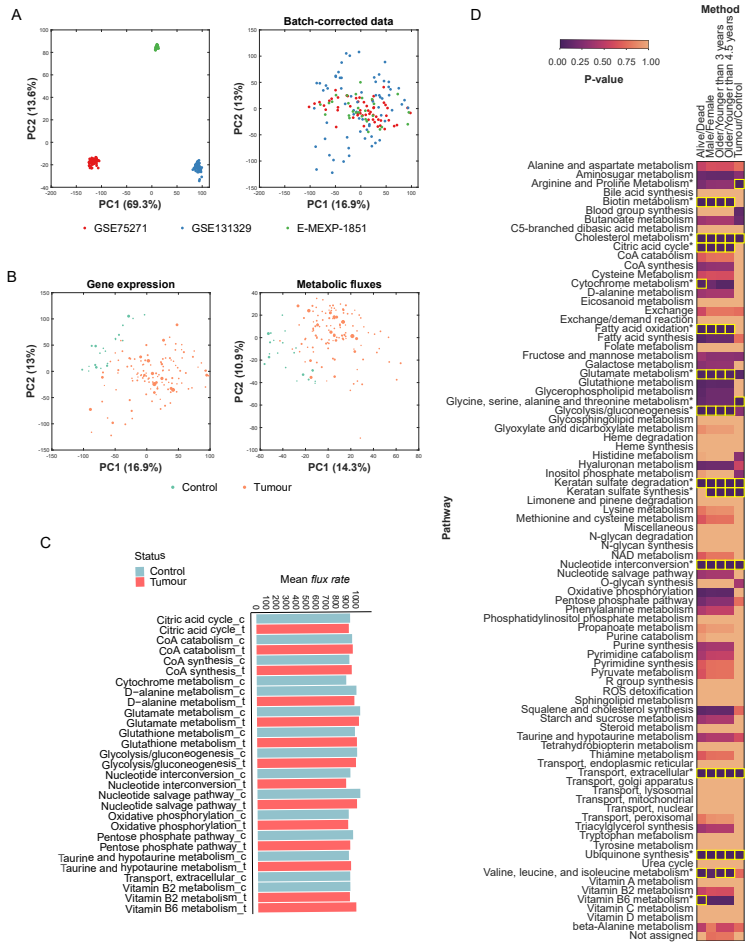



Figure 4.2: **GSMM characterisation of hepatoblastoma metabolism.** A. Principal component visualisation of the three transcriptomic datasets considered. Upon batch correction through ComBat, the datasets correctly overlap, indicating that confounding experiment-specific variation has been reduced. B. Principal component visualisation of the aggregated cohort in terms of transcriptomic and fluxomic state, displaying the main phenotypic groups. The two groups appear circumscribed to well-defined areas of the principal component space for both omics across subjects, indicating that they describe distinct characteristics in the two groups. In contrast, no clear trend can be observed in terms of subject age, here represented by the circle size. An alternative representation of these graphs is given in Figure 4.3. C. Average flux in each pathway across patients and controls, obtained through FVA. Pathways associated with glutathione and CoA metabolism were found up-regulated, while the ones linked to nucleotide salvage, D-alanine metabolism, and central metabolism were down-regulated. D. Flux enrichment analysis over the pathways in the genome-scale metabolic reconstruction for the flux rates from the FVA (maximal fluxes). Pathways with * and yellow contour are statistically significantly enriched in at least one stratification. In particular, extracellular transport, nucleotide interconversion, ubiquinone synthesis and keratan and cholesterol metabolism are the pathways enriched in all the stratifications. No significant difference was detected between the two age-based stratification enrichments.

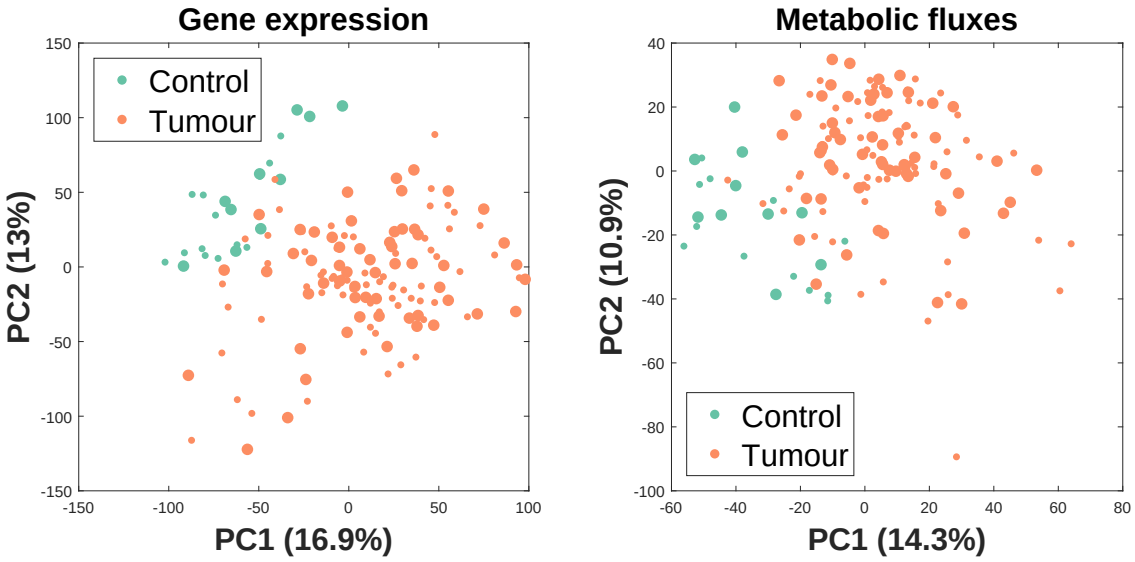



Figure 4.3: **Principal component visualisation of the aggregated cohort in terms of transcriptomic and fluxomic state, displaying the main phenotypic groups.** The two groups appear circumscribed to well-defined areas of the principal component space for both omics across subjects, indicating that they describe distinct characteristics in the two groups. In contrast, no clear trend can be observed in terms of subject gender, here represented by the circle size (small circles=male, big circles=female).

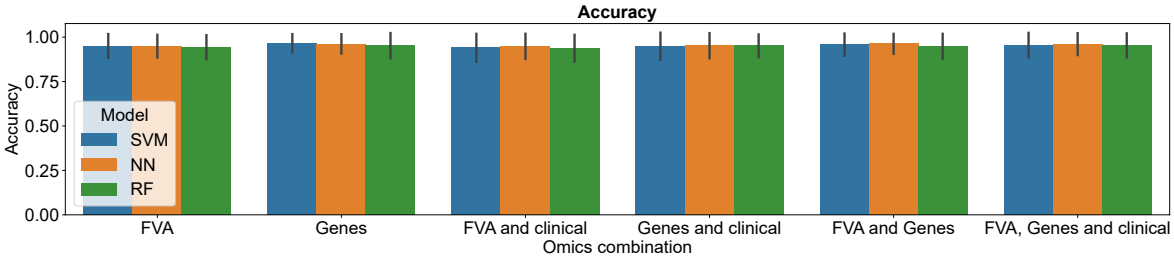



Figure 4.4: **Classification accuracy for Support Vector Machine, Random Forest and Neural Network models.** The three model types performed comparably well (in general, no statistically significant difference could be detected) in all omics combinations.

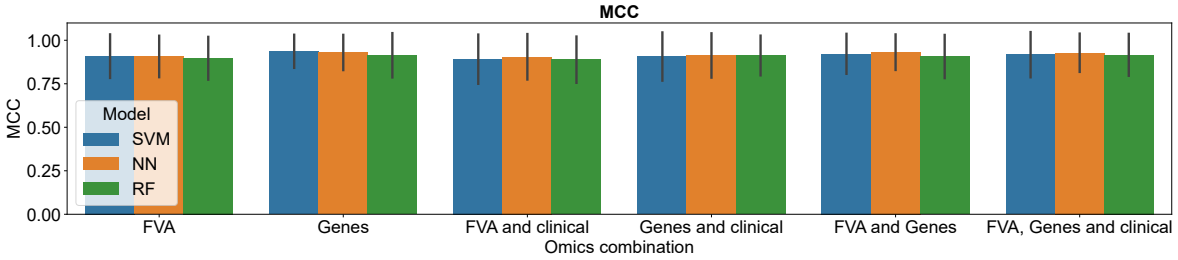



Figure 4.5: **Matthew correlation coefficient for Support Vector Machine, Random Forest and Neural Network models.** The three model types perform comparably well (in general, no statistically significant difference could be detected) in all omics combinations.

regarding potential biomarkers for hepatoblastoma, we analysed the weights $\mathbf{w}$ from Eq 4.3 given to the input features by the SVM models as a proxy for feature importance. Unlike in the previous chapter, we did not notice any regularisation-like phenomenon caused by the integration of fluxomic with transcriptomic data (see Figures 4.6-4.8). However, the integration did redistribute the weights across the features in a way which varied depending on the omics used.

Figures 4.12D-E show the total weight (i.e. relative importance) each gene and reaction were given across the 200 iterations conducted. To facilitate comparison between the scenarios, the weights were quantile-normalised. The plot reveals that the most critical genes and reactions for subject classification vary depending on the data sources employed. In particular, the integrations highlighted as more relevant genes EPCAM, FRRS1L and ZBED8, whereas the base scenario with the SVMs trained only with gene expression had determined as more important the genes TMPRSS15, NPVF and HHLA2 (not relevant in the integrative experiments). It is interesting to note that none of these genes is associated with the reactions deemed relevant by the SVM models. EPCAM is a gene classified as tumour antigen in the database UniProt [411] while FRRS1L, more specifically, concerns the regulation of the glutamate receptor signalling pathway. The role of glutamate metabolism in hepatocytes is well-known and established [412]. Gene ZBED8 is instead a gene for which not much information has been collected yet, which suggests it might be involved in the metabolism of hepatoblastoma in an indirect way. Among the genes which were instead detected solely in the single-omic scenarios, TMPRSS15 is responsible for the activation of pancreatic proteolytic proenzymes, while NPVF is a neuropeptide and HHLA2 participates in the proliferation of T cells and regulation of cytokine production in lieu, which have a prominent role in inhibiting (but sometimes even stimulating) growth of cancer cells [413, 414]. Among these, EPCAM, TMPRSS15 and HHLA2 can be found in blood samples [415], as also reported in The Human Protein Atlas database (https://www.proteinatlas.org) [416], whereas EPCAM and HHLA2 can be also found in urine samples [417, 418].

When considering which reactions were deemed important by the SVM models, the results high-

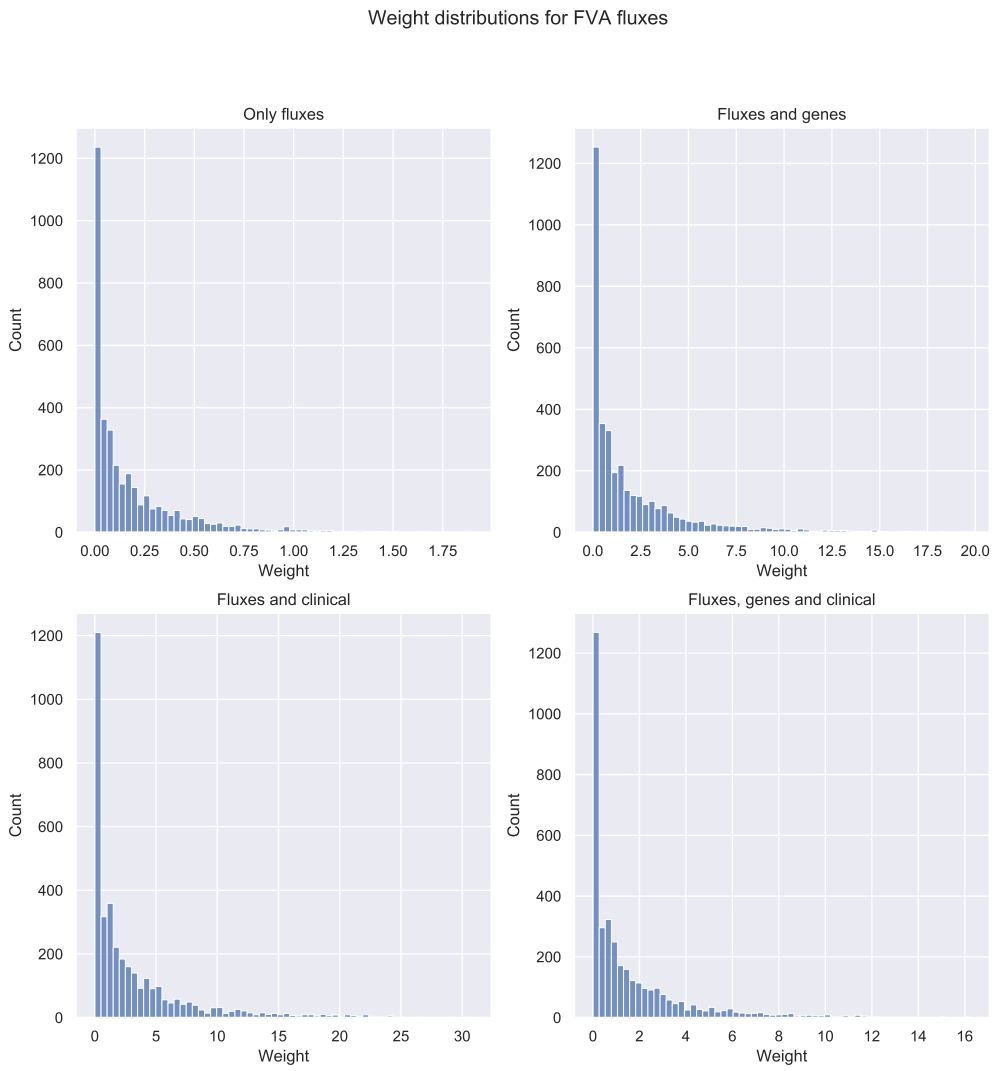



Figure 4.6: **Total weight distributions for metabolic reaction fluxes computed via FVA in the four different experimental settings.** The shape of the distribution does not change in the various scenarios.

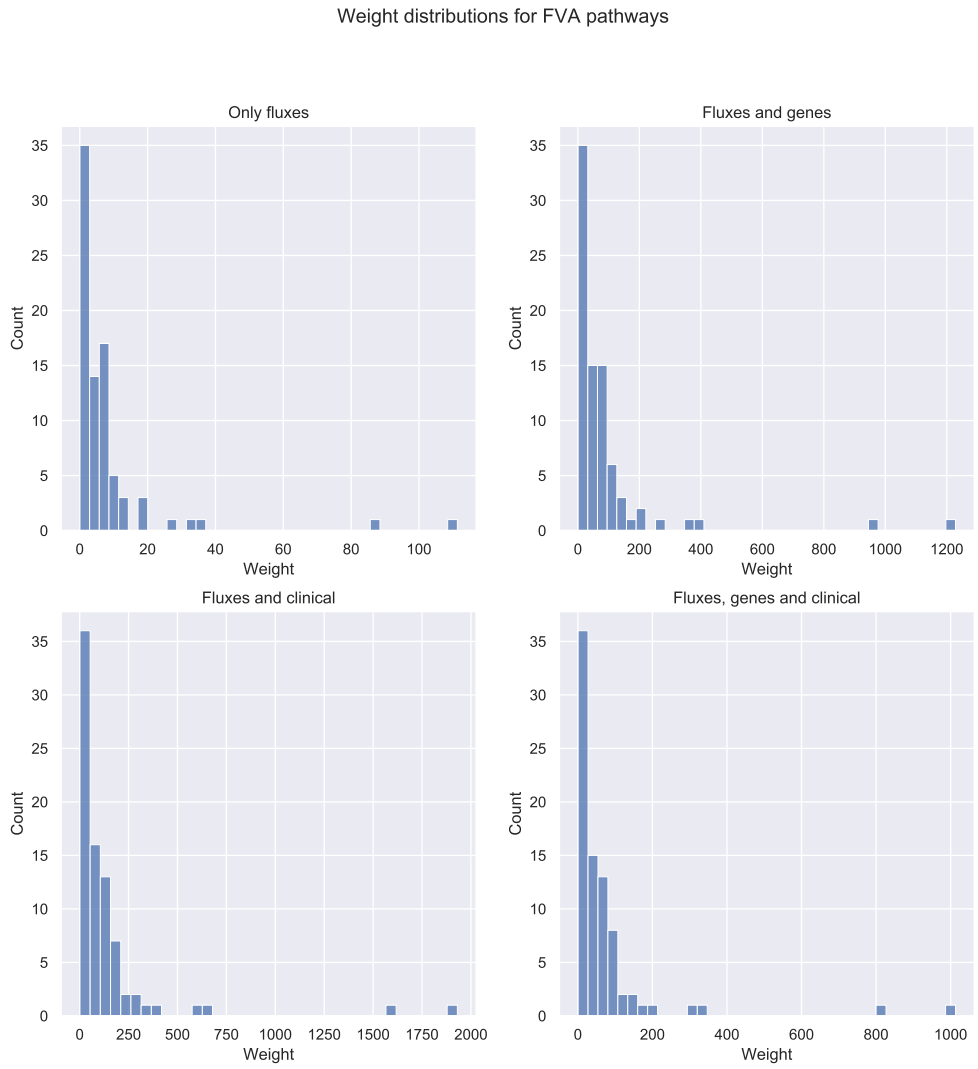



Figure 4.7: **Total weight distributions for metabolic pathways.** The weights were computed starting from the weights given to the reactions associated to each pathway by the SVMs trained with reaction fluxes generated via FVA. It is easy to notice that the shape of the distribution has not changed in the various settings.

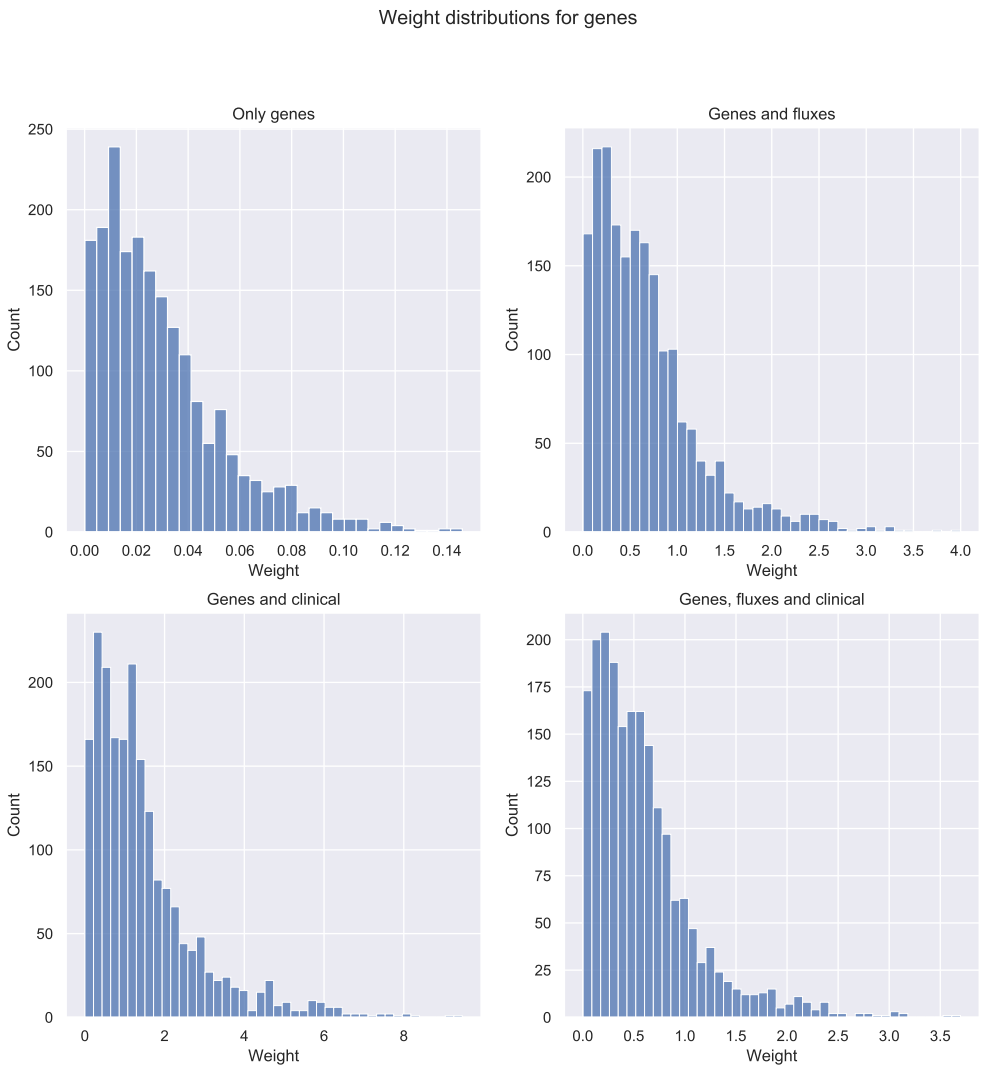



Figure 4.8: **Total weight distributions for the gene expression data in the four integration settings.** The shape of the distribution does not significantly change in the various scenarios.

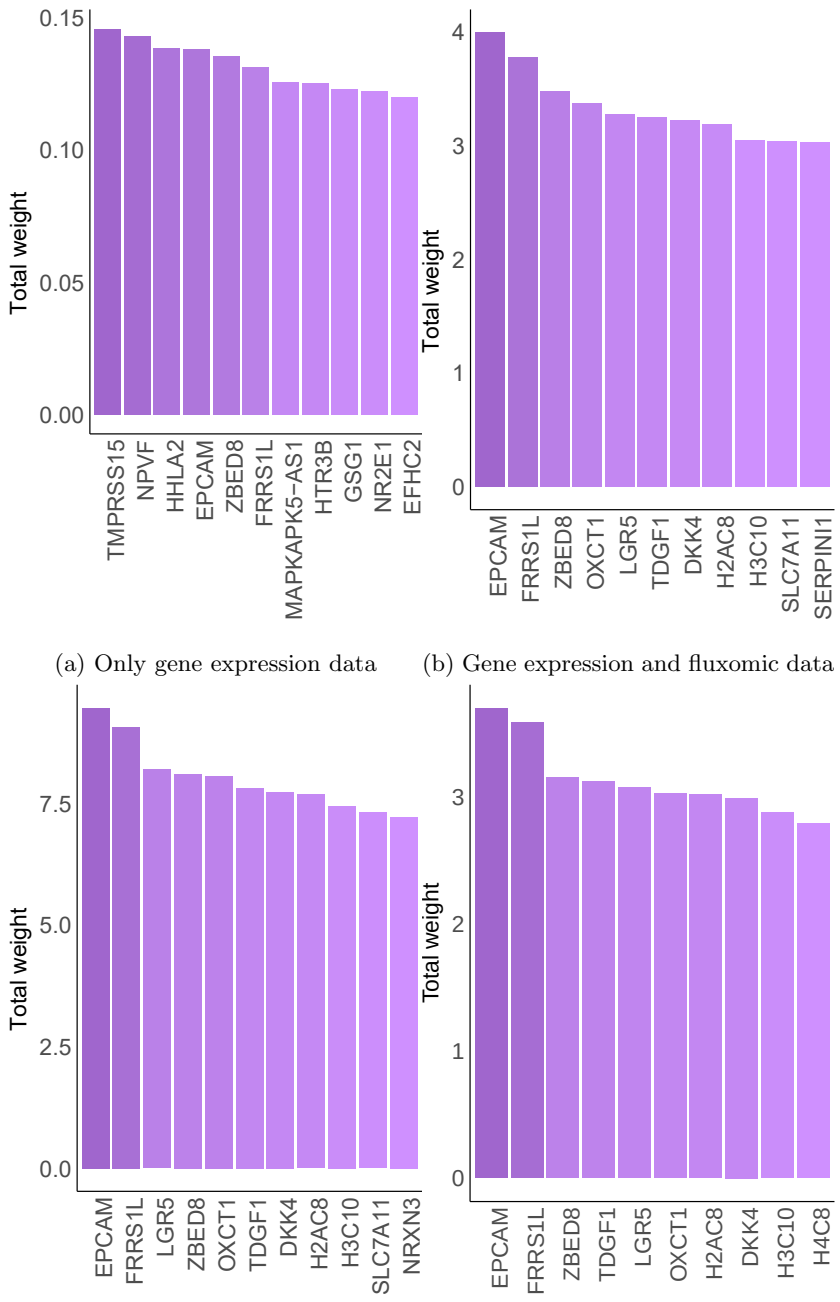
lighted that reactions DASCBR (dehydroascorbate reductase, which participates in glutamate and ascorbate metabolism), RNMK (ribosylnicotinamide kinase, which is involved in the metabolism of nicotinamide adenine dinucleotide, a potential target for treating cancer [419]), AASAD3m (L-aminoadipate-semialdehyde dehydrogenase, which participates in the production of lysine, whose acetylation is responsible for cancer development [420, 421, 422]) and EX_lys_L(e) (lysine exchange) were present in all the experimental scenarios.

When exploiting only the information contained in fluxomic data, the following reactions were identified as useful for the diagnosis of hepatoblastoma: LNS14DM (lanosterol 14-alpha-demethylase, which has shown to be able to decrease the proliferation of cancer cells [423]), G6PDA (glucosamine-6-phosphate deaminase), NAHCO3_HCLt (bicarbonate transport, that may be used in a diagnostic setting [424] but controversially in therapy [425, 426]), THYMDtm (thymidine transport) and MI1PP (myo-inositol 1-phosphatase, regulating myo-inositol which can be used in cancer treatment [427]).

Finally, the reactions that were useful for the prediction only in the integrative settings were CLCFTRte (CFTR chloride transport), FAOXC220200x and FAOXC180x (beta-oxidation of long-chain fatty acid). An alternative graphical representation of these plots, which allows for easier comparison between integrative scenarios, is provided in Figures 4.9-4.10.

When using FEA, we noticed that the enriched pathways were more stable across the four settings than the genes. In particular, we found that Vitamin B6 metabolism and Cholesterol metabolism were observed only when gene expression data were integrated with fluxomic data (regardless of the presence of clinical data in the integration), while the Pentose phosphate pathway was enriched only in the single-omic setting and when transcriptomics was integrated with fluxomics (but not in the presence of clinical data). Triacylglycerol synthesis was instead absent only when integrating the metabolic fluxes with clinical data, as opposed to Inositol phosphate metabolism, Amino sugar metabolism, Glutathione metabolism, Citric acid cycle, Glycerophospholipid metabolism, Lysine metabolism, NAD metabolism and Oxidative phosphorylation, which were found to be enriched in all scenarios. Overall, the integration of transcriptomic data with fluxomic data contributes to a greater number of enriched pathways. These results are summarised in Figure 4.12B, and others showing the models' weights distributions are presented in Figures 4.6-4.8. In order to test for the robustness of our biomarker identification pipeline, we also considered the median of the weight distribution for each feature, instead of the total weight. This led to similar results to the ones presented above (in this scenario, taking the sum of the weights is qualitatively equivalent to taking the mean, since all the weights would need to be divided by 200 iterations, thus leaving their relative "importance ranking" unchanged), with the following differences being found.

Among the reactions that the models considered important for the prediction, AASAD3m was not present in the single-omic scenario, while RNMK was not relevant when the metabolic fluxes were



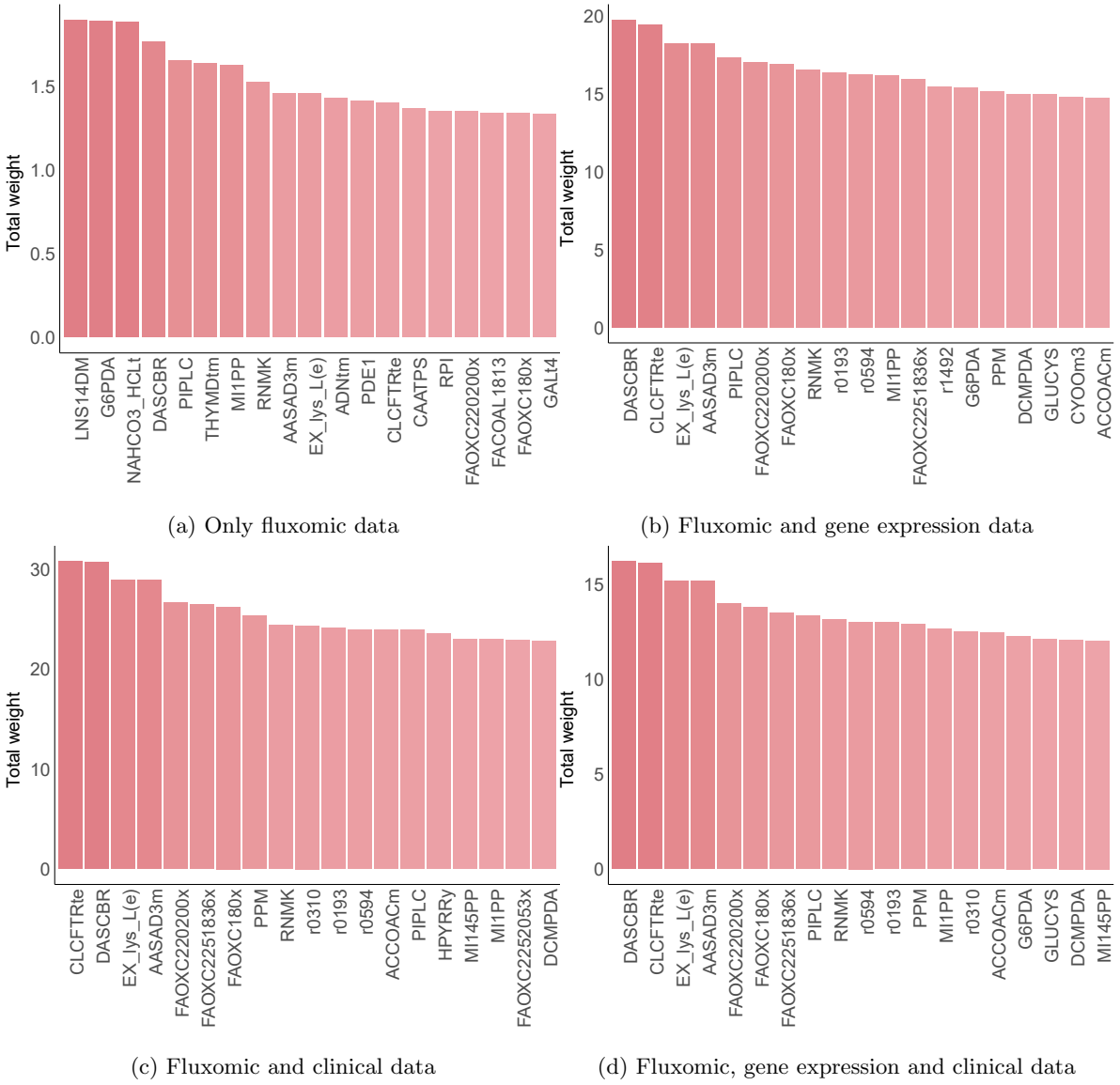

(a) Only gene expression data

(b) Gene expression and fluxomic data

(c) Gene expression and clinical data

(d) Gene expression, fluxomic and clinical data

Figure 4.9: **Total weight bar plots for the weights attributed to the genes in the four integrative scenarios.** The bar plots display only the genes with weight in the 99.5th percentile.



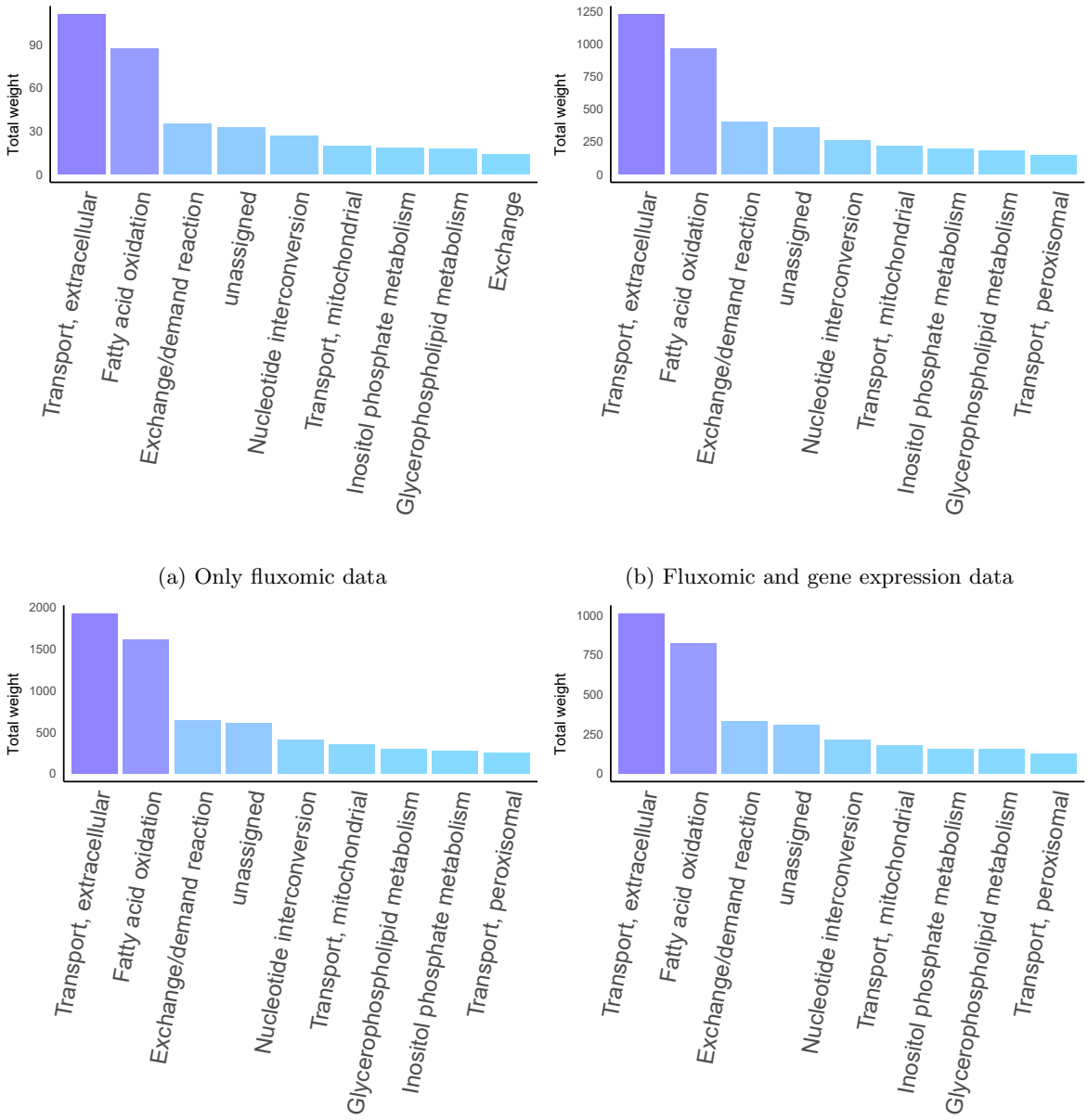

(a) Only fluxomic data

(b) Fluxomic and gene expression data

(c) Fluxomic and clinical data

(d) Fluxomic, gene expression and clinical data

Figure 4.10: **Total weight bar plots for the weights attributed to the metabolic reactions in the four integrative scenarios.** The bar plots display only the reactions with weight in the 99.5th percentile.




(a) Only fluxomic data

(b) Fluxomic and gene expression data

(c) Fluxomic and clinical data

(d) Fluxomic, gene expression and clinical data

Figure 4.11: **Total weight bar plots for the weights attributed to the metabolic pathways in the four integrative scenarios.** The bar plots display only the pathways with weight in the 90th percentile. The most relevant pathways are consistent across the four settings, however, the integrations reveal the importance of peroxisomal transport, not shown in the single-omic scenario.

integrated with the genes. FAOXC2251836x replaced FAOXC180x in the integrative settings, while ADNtm was found to be significant only in the metabolic fluxes without any other omic.

Regarding the genes and pathways, the results were identical to the ones reported above, whereas for the weight-informed FEA the following results were found: NAD metabolism, Amino sugar metabolism, Glutathione metabolism, Citric acid cycle, Lysine metabolism, Oxidative phosphorylation were enriched in all four scenarios; Glycerophospholipid metabolism was observed only in the integrative settings, unlike Triacylglycerol synthesis which was present only in the single-omic scenario. Finally, Inositol phosphate metabolism was enriched in all four settings except when integrating metabolic fluxes and clinical data.

### 4.4.3 Relation between clinical data and diagnosis accuracy

We then asked if the constructed SVM models could be used to gain insights that can more accurately diagnose hepatoblastoma in a precision medicine fashion. We therefore analysed more in depth the trained models, in order to find directly applicable heuristics for guiding their use. In particular, we looked at how age, gender and health status could affect the predictive performance of different combinations of omics.

In Figure 4.12C, the accuracy distribution of the SVM models for different omics combinations is reported. It can be noticed that, across all omics combinations, female patients (in purple) tend to obtain a more accurate diagnosis (with the worst performance, in the case of a healthy subject, being above 94%, achieved by using only transcriptomic and clinical data) on average. On the other hand, in presence of a male, ill patient, the only use of transcriptomics will provide the best diagnostic performance. Similarly, the figure shows the desired property of our training and evaluation pipeline, namely the ability to discriminate with higher accuracy patients suffering from the tumour (in red). Notably, when a patient has hepatoblastoma, the best predictive performance is achieved by integrating both transcriptomic and fluxomic data. Conversely, in the case of healthy control subjects, transcriptomic and fluxomic data separately represent the best two options for a correct computer-aided diagnosis. In both stratifications, as expected, the use of merely clinical data corresponds instead to trying to guess the phenotype of the individual, since there is no relation between age, gender and health status. This simple analysis allowed us to double-check that there were no spurious associations in the data due to their collection. Even though the information regarding clinical status cannot be exploited in a diagnostic setting, it is always possible to make use of other clinical information such as age and gender when choosing which omics combination to adopt for the diagnosis.

In particular, we investigated the performance of the SVM models with a more fine-grained detail to find less visible patterns in the performance distribution. Interestingly, we found that patients of

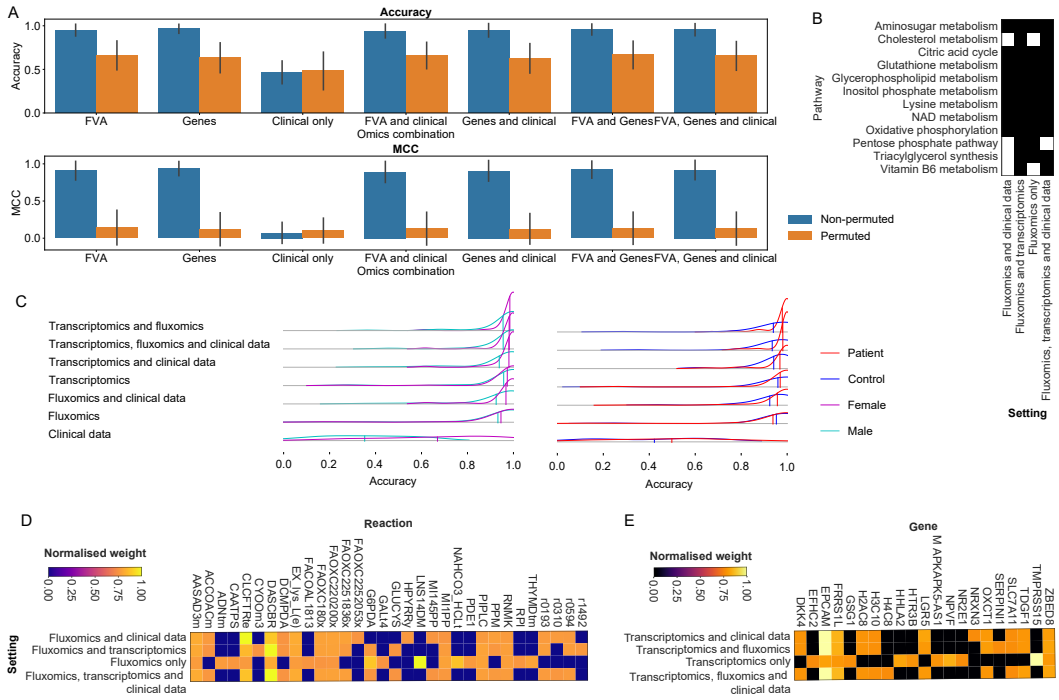



Figure 4.12: **Results of multi-omic integration.** A. Classification accuracy (top) and Matthews Correlation Coefficient (bottom) for SVM models that recognise tumour and control samples. Blue and orange bars respectively represent the performance of SVM models built using the original datasets and the same datasets with random sample labelling to phenotypic groups. The results with the original labels significantly outperform those with permuted labels, which approximate the performance of a random classifier. This proves further that our models are capable of learning from transcriptomic and fluxomic data the most relevant features which can be then used for biological interpretation. B. Statistically significantly enriched pathways from flux enrichment analysis across the four experimental settings. A black entry means that the pathway is significantly enriched in the cohort. When combining fluxomic and transcriptomic data the enrichment returned more statistically significant pathways than in the other settings. C. Distribution of the accuracy of the SVM models trained, according to clinical data. Different groups of individuals can find beneficial a machine learning-aided diagnosis with different omics combinations. For female patients (in purple), all omics combinations tend to obtain a more accurate diagnosis. Moreover, if the patient has hepatoblastoma (in red), the best predictive performance can be achieved by integrating both transcriptomic and fluxomic data, while in the case of healthy control subjects, these two omics must be used separately to obtain a more accurate computer-aided diagnosis. Means are represented by vertical lines. D-E. Total weight attributed to reactions (D) and genes (E) in the four integrative scenarios. The weight distributions were first quantile-transformed so that they could be comparable, and the weights were then normalised in [0, 1]. An alternative representation of graphs D-E is shown in Figures 4.9-4.10.

age = 0.7 were much less likely to receive a correct diagnosis than patients of different ages ($p$-value ¡ 0.001), while for patients of age = 7 the integrations of omics performed better than the use of single omics in general ($p$-value = 0.057). Finally, as a further addition to Figure 4.12C, we found that omics integrations achieved overall better accuracy than single omics ($p$-value = 0.057) when the patient was female and had hepatoblastoma.

In general, both the characteristics of the dataset and biological factors might contribute to the above patterns. The highest accuracy observed for hepatoblastoma patients could be due to their over-representation in the dataset, which grants a more complete distribution for this class of subjects. Similarly, the omics integration might result more effective for the patients for this reason. In contrast, the higher discriminatory power found for the female patients cannot be explained in this way, given that no clear connection was seen between gender and health status (Figure 4.3). Besides, differences between genders in terms of omics accuracy could be underlain by specificities in developmental programming of growth and metabolism, which present sex differences not only in normal development but also in disease [428], and are linked to specific risk factors in childhood cancers [429]. Thus, critical aspects of metabolic rewiring in female subjects could be better captured by the GSMMs here developed, leading to better accuracy.

## 4.5 Conclusions and future directions

In this chapter, we started from the results obtained in the previous two chapters to design and implement an interpretable multi-omic integrative pipeline. We investigated molecular biomarkers and metabolic mechanisms in a precision medicine fashion, in order to shed light on the onset of hepatoblastoma, potentially helping diagnose more accurately this disease in young patients. An aspect of this involved also the investigation of the sensitivity that such a method could have with respect to the characteristics of such patients, namely age, gender and the observed phenotypic trait. Moreover, we examined how different combinations of data can interplay with these and highlight different aspects of the metabolism of the patient. In particular, important genes as revealed by the integrations, are linked to cancer metabolism and hallmarks. Starting from gene expression profiles, we generated metabolic fluxes representing the metabolic state of the patients and, with the addition of clinical data, integrated all this information within a machine learning pipeline to determine whether an integrative approach could lead to an improvement in the diagnostic performance of our models.

We investigated genes, reactions and metabolic pathways by resorting to a feature importance approach in quest of potential biomarkers to guide future research in hepatoblastoma. We demonstrated that different omics combinations can achieve optimal predictive performance for different patients according to the patients' clinical data, even though the individual omics used can have a signifi-

cantly skewed performance distribution [386], and that machine learning models can be endowed with different sensitivity to distinct biological entities based on the omics combinations they are trained on. Finally, we extracted novel mechanistic biomarkers whose study could be relevant in the research regarding the mechanisms underlying hepatoblastoma.

Overall, the results presented in this chapter suggest that using systems biology approaches in conjunction with machine learning methods can provide valuable insights into the biological mechanisms of rare cancer conditions, for which omics data and biological knowledge are not as widely available as for the most commonly studied diseases (such as breast or lung cancer).

One of the limitations of our study is the practical impossibility, with current technology, to directly measure metabolic fluxes in human patients [86]. Here (and in the previous chapters), we mitigated this by adopting genome-scale metabolic models, but these require some experimentally measured information such as transcriptomic data. Another potential limitation of our approach is the challenging direct applicability in some cases. Specifically, we have determined when to use which combination of omics, but this information is exploitable only in a limited number of cases, such as when the conditions determining the omics combination to use are based on clinical information such as gender or age (and, within this, only when we are within certain ranges). Yet, this work can serve as a guide for further research in hepatoblastoma, and the biomarkers found could potentially lead to the development of new diagnostic or therapeutic tools. Our approach has the advantage of elucidating how molecular entities (and their importance in the development of the disease) can be related to hepatoblastoma, with a granularity that is based on the patient's clinical information.

In the future, new studies could focus on improving the omics integrative approaches, as well as investigating other omics data in this setting, such as proteomics. For larger datasets, alternative optimisation methods based on heuristics could be investigated and adopted to improve the speed and quality of the training phase for the studied machine learning models [405, 406].

## 4.6 Related work, funding and final remarks

The work presented in this chapter has been published in Computers in Biology and Medicine [430], and the paper was written in collaboration with multiple co-authors. As described in the introduction to this chapter, I performed all the analyses relating to the generation of the metabolic features, I designed, implemented and tested the entire machine learning pipeline, and I conducted the analysis and interpretation of the results as well.

This work was supported by a Research Award from the Children's Liver Disease Foundation, grant number SG/2019/06/03, and a Network Development Award from The Alan Turing Institute, grant

number TNDC2-100022.

We would like to thank Dr Jane Hartley, Consultant Paediatric Hepatologist at the Birmingham Women's and Children's NHS Foundation Trust, for inspiring discussions and advice regarding this work.

# Chapter 5

# An interesting future direction

In this work, we have discussed how machine learning methods can be leveraged to exploit the mechanistic information contained in genome-scale metabolic models by integrating them with multi-omic data in a precision medicine scenario. We have shown how different approaches are possible (and sometimes necessary) depending on the data and the task at issue, and that there is still leeway for improvement and new developments, which hints at a promising future for precision medicine.

In this short chapter, we speculate about what future developments the field could take, and present a novel framework which aims at taking advantage of GSMMs even further. In particular, so far topological information from these networks has been rarely used compared to more common analytical approaches such as FBA [431, 432], albeit it has shown promising results [433]. With the underlying hypothesis that metabolic topological information can complement the mechanistic biological information directly present in GSMMs [434], we propose a framework whose aim is to combine these two aspects of GSMMs in an end-to-end fashion, and offer a possible pipeline for hypothesis evaluation and experiment design.

In the following sections we will describe each step of the pipeline, that is graphically represented in Figure 5.1. To fix the ideas, we will take as an example a classification task in which we have to determine whether the patient has Alzheimer's Disease or not, it being understood that the computational framework we propose here can be adopted for any task that can involve the use of GSMMs in the way we have presented so far. In order to be as comprehensive as possible, and lacking experimental results, we will limit ourselves to describing the potential approaches that could be taken to exploit this aspect of GSMMs, and to listing the challenges that are more likely to occur.

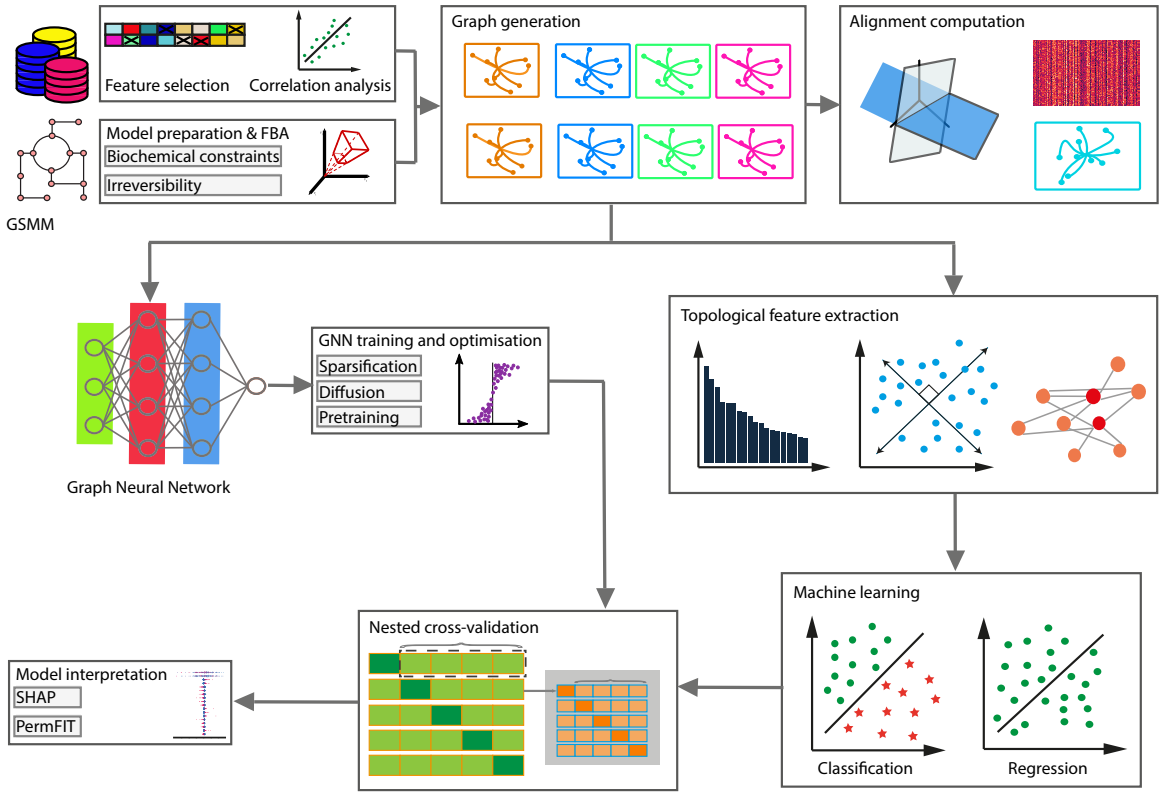



Figure 5.1: **General framework for the investigation of topological features in genome-scale metabolic models.** Starting from transcriptomic data,a context-specific GSMM is chosen and used to generate the metabolic fluxes, upon conversion of the model and application of consistent metabolic constraints. From the metabolic fluxes, metabolic graphs are generated and the alignment between them and transcriptomic data is computed to test for the suitability of the graph implementation. Topological features can then be analysed in two ways: directly, through the extraction from the graphs and their use as input in classical machine learning models; or indirectly, by using the graphs as input to Graph Neural Networks instead. In the latter case, training of the network can potentially be simplified with network pretraining, graph sparsification or by exploiting knowledge of graph diffusion dynamics. Model training is conducted within a cross-validation framework, such as nested cross-validation, to guarantee robustness. Finally, results are biologically interpreted with the use of techniques such as SHAP or PermFIT.

## 5.1 Data and Model preparation

Data preprocessing, as presented in Subsection 1.3.1, is fundamental in every machine learning task. When aggregating data from different sources, extra care will need to be taken to reduce the batch effect (as we did in Chapter 4) and other discrepancies in the data. Of equal importance is the choice of the model(s). Even though, in principle, the GSMMs used in the previous chapters or mentioned in Subsection 1.2.1 can be adopted, for a complex task such as the modelling of Alzheimer's Disease a more precise brain-specific model (such as the one from [435]) could be used. However, using a cell-specific model is not assurance of adherence to results, since metabolic constraints, as explained in Subsection 1.2.1, can greatly influence the metabolic outcome of a simulation. For this reason, when adopting such models, it will be paramount to make sure that such metabolic parameters are in agreement with the true experimental values measured for the species under study, which implies that these will need to be adapted in case their values are known only for other species. Same care needs to be taken when applying constraints that are exclusively in accordance with certain phenotypic traits (diseased/healthy states), as the metabolic fluxes generated, in case of wrong parameters, will not be coherent with the metabolic picture of that particular phenotypic state.

## 5.2 Fluxomic data generation

Ideally, the pipeline could be used with any variation of FBA or sampling technique. If using FVA, one could also consider making the GSMM irreversible, i.e. converting it into a format in which the reversible reactions are uncoupled and replaced by two reactions with opposite direction, as this could help distinguish and elucidate better the contribution of the reversible reaction in both directions.

## 5.3 Metabolic graphs generation

After obtaining the metabolic fluxes, the next step in our proposed pipeline consists of generating metabolic graphs. Our suggestion would be to adopt the Mass Flow Graph (MFG) paradigm from [436], and create one MFG for each patient starting from the metabolic profile computed with the patient-specific metabolic models. An MFG is a graph whose nodes are the reactions of the metabolic network and whose edges represent the flow of metabolites produced by one reaction and consumed by the other (normalised by the total consumption flow of each metabolite). In particular, the weight between reaction A and reaction B (considering directionality) is computed as the probability that any randomly chosen metabolite is produced by A and consumed by B. We suggest to adopt this data structure because it allows to include reaction directionality, which is essential for a coherent and

complete understanding of metabolism [436]. Before feeding these graphs to a Graph Neural Network model, however, we recommend to filter the data upstream and remove all the edges whose weight is lower than $10^{-4}$, in order to eliminate all the edges that represent weak interactions between reaction pairs. This should mitigate overfitting and help the model differentiate the graphs more easily. As a further possibility, one could also choose to enrich these graphs by explicitly adding the outward and inward flux for each reaction as node features.

MFGs are not the only possible graph representation for GSMM-generated metabolic fluxes. An alternative representation could be obtained, for instance, by using the Continuous k-Nearest Neighbor (CkNN) algorithm [437], which has shown to generate graphs that can improve classification performance of Graph Convolutional Networks, compared to other algorithms or tabular feature matrices [438]. CkNN is a neighbourhood method, i.e. a method that generates edges between samples (in this case, reactions), based on their distance in the sample space (meaning that it uses the raw features). In particular, CkNN considers two samples to be neighbours only if their distance (determined by a predefined similarity measure) is smaller than the geometric mean between the distances of the two samples with their farthest k-th neighbours, with k being a hyperparameter chosen by the user. In fact, CkNN generates an edge between two samples depending not only on their relative distance/similarity, but also on the density of the sample region to which they belong, in order to account for the different scales of proximity that there can be between samples.

## 5.4 Graph-feature alignment

Even though the choice of MFGs is well motivated from the point of view of biological interpretability, another alternative representation such as the geometric one provided by CkNN could be equally relevant. A possible way to understand which representation is better in a scenario with GNNs, is to consider the alignment between the generated graph and the raw features (in our suggested framework, the metabolic fluxes obtained from the GSMM), as this has proven to be a good indicator of predictive performance with GCNs [439]. Even though CkNN-generated graphs have demonstrated to increase the performance of these models [438], GNNs have not been tried on MFGs yet, therefore the measurement of alignment could help choose the most promising approach. Moreover, the measure of alignment defined in [439] could be used to make other interesting comparisons, such as the one between the gene expression data and the flux-based graphs, which could potentially inform the process of generation of the metabolic fluxes and thus guide the researcher towards the FBA variant and parameter values that lead to the highest degree of alignment between the transcriptomic data and the metabolic graphs.

## 5.5 Analysis of topological features

After the conversion of the metabolic fluxes into MFGs (or CkNN-generated geometric graphs) the next step in the pipeline would be to analyse the topological features of the metabolic networks. This can be done in two ways: either directly (by using them as input in classical machine learning models), or indirectly (with a GNN).

### 5.5.1 Direct analysis of topological features

A possible way to analyse the topological features of GSMM-derived metabolic networks consists of first deriving these features directly and then using machine learning models to extract as much predictive information as possible. This can be done, for example, with HCGA, which computes a great variety of topological features and then uses proven methods to analyse the resulting tabular data matrix [440]. HCGA (Highly Comparative Graph Analysis) is an open-source python library that automatically extracts thousands of topological features from graphs (spectral properties, centrality measures, communities, etc ...) that can later be used in supervised settings while retaining interpretability of the generated features. Given that too many features could lead to overfitting (because of the emergence of spurious patterns, especially in small datasets), the library allows the user to select what type of features to extract depending on the computation time required for each and on the level of statistical complexity, and to choose whether to compute the features on the entire graph or on its largest connected component only. The choice of which features to compute in this case is purely technical, i.e. dependent on the available hardware and time.

Alternatively, the topological features used by Machicao *et al.* could be generated [441]. They tested two groups of features in a classification scenario, using traditional machine learning algorithms: the first group consisted in the top 5 principal components of the available gene expression dataset; the second one instead was composed by classical topological measures such as average degree and average hierarchical degree, average geodesic path length and assortativity, computed on a graph derived from a genome-scale metabolic network. Unlike the work presented in this thesis, however, the authors did not use FBA approaches, instead they first defined and computed a Reaction Activity Score (RAS), which determined the level of activity of each reaction, and then built a weighted matrix based on these levels and the metabolites shared between the reactions. For reaction $r$ and sample $s$, the RAS was computed starting from the available gene expression data according to the two following formulae:

$$\begin{aligned} RAS^{\wedge}_{r,s} &= \min(E_{g,s} \mid g \in G_r) \\ RAS^{\vee}_{r,s} &= \sum_{g \in I_r} E_{g,s}, \end{aligned} \tag{5.1}$$

where $E_{g,s}$ is the expression level of gene $g$ in sample $s$, $G_r$ is the set of genes necessary for reaction $r$ to occur, and $I_r$ is the set of genes encoding the isozymes of $r$. These equations are very similar to Equations 1.2, with the difference being that in this case the RAS is the activity level of the reaction and does not influence its bounds directly, and that in case of isozymes we sum the gene expression instead of taking the maximum value across the gene set. This alternative approach is interesting and could be investigated in parallel with FBA-generated fluxes, but does not take into account the total flow of metabolites that originates from the first principle of "balance" on which FBA is built. However, for robustness purposes it could be interesting to compare how reactions distributions change between the two methods.

The use of a machine learning model with the extracted features could then be employed as a way to demonstrate the importance of the topology of these networks, since if the performance of the model is good this entails that the topological features contain relevant information. The choice of the model is dependent on the type of features extracted, on their amount compared to the number of available graphs, and on the task to perform, as in any machine learning task.

### 5.5.2 Indirect analysis of topological features through GNN

To determine the importance of the topological features in genome-scale metabolic networks, it is also possible to extract and investigate them indirectly, by means of a Graph Neural Network. This involves using graphs directly as input, and exploiting the topological features indirectly through a model. In our fictional example, which is a classification task to determine patients' health status (presence of Alzheimer's Disease or not), the machine learning task to solve would be a classification at the graph level, as opposed to classification at the node level (where one would classify the single nodes of the graphs, which in the proposed approach are the reactions) or at the edge level (usually consisting in predicting whether two nodes are linked or not).

A GNN is a neural network that preserves graph symmetries (permutation invariances) [442]. Zhou *et al.* have described the general design pipeline of GNNs in [443]. The first step is usually the generation of the graphs (the MFGs, in our suggested pipeline). A subsequent step is the characterisation of these graphs, which, in the case of MFGs, is straightforward given that MFGs are directed and homogeneous graphs (whereas the CkNN algorithm generates undirected graphs). The structure of GNNs, being them neural networks, can be defined in terms of modules/components, each depending on the type and characteristics of the graphs to use. In particular, there are three types of modules: (i) propagation modules; (ii) sampling modules; (iii) pooling modules. (i) are modules that define how the information is aggregated across the entire graph starting from the individual nodes. There are several types of propagation modules, among which the most common are the spectral ones and the attentional ones, both being convolutional approaches [443]. The former operate with the spectral

representation of the graph, which is a graph invariant, while the latter uses an attention mechanism when aggregating the node features across the graph. However, when using spectral convolutions, several problems arise. In particular, they are computationally expensive, especially when dealing with large graphs, as it may be in our field. Moreover, when used on unseen graphs that have a significantly different structure (and therefore eigenvalues) from the training ones, they do not scale very well in terms of learned filters [444]. On the contrary, attention-fuelled graph neural networks can be successfully used on large, noisy graphs (like MFGs generated from GSSM metabolic fluxes), because the function used to weigh the aggregated node features is learned directly from the data and not explicitly defined before training, which enables the model to better direct the flow of information from the various regions of the graph to the node being updated [443]. (ii) are modules useful when the graphs are large and present many nodes heavily connected with each other, which is normally not the case for metabolic networks, given that reactions usually involve few metabolites each. However, reducing the size of the network via node/edge sampling can still be useful in that it speeds up the computation of the model and even allows for certain bigger, more complete networks to be used, which otherwise would be impossible because of the long computational time required. Finally, (iii) are modules necessary for the aggregation of the node features at the graph level [444]. An example of pooling aggregation operator is SAGPool, which is an attention-based approach [445]. Once all these modules have been assembled, the GNN is ready for training. However, in general GNNs do not take explicitly into account the characteristics of the global topology of the network, and in certain circumstances cannot even learn them [446]. For this reason, Wang *et al.* designed BiFusion, a graph neural network tailored for the use with bipartite graphs (as genome-scale metabolic networks are), which alternatively could be used for the analysis directly on the metabolic network[447].

Given that the training of complex neural networks is always difficult to conduct successfully, several approaches have been developed, based either on the modification of the general training procedure or on transformations of the input graphs. In particular, spectral sparsification of graphs as proposed in [448] has been shown to robustly improve classification performance in GNNs, the reason being that the preserved Laplacian quadratic form is strongly associated to graph partitioning and community detection [438]. An approach consisting in explicitly considering the way information flows through the graph during aggregation was instead proposed in [449]. They integrated the graph diffusion dynamics into a GNN by replacing the graph stochastic matrix with the diffusion stochastic matrix in the model formulation. In both cases, the classification accuracy of the GNN saw an improvement. Finally, a third approach that could be adopted consists of using both node-level and graph-level pretraining together (in this order), and then training and fine-tuning the neural network with some linear models on top of the graph representation layers in an end-to-end fashion [450]. This way, the GNN can learn local and global representations at the same time, thus improving significantly its classification performance.

## 5.6 Model validation and interpretation

As already described in Subsection 1.5.1, to obtain a robust and unbiased performance estimation of the model, a cross-validation framework needs to be used. In particular, if one wants to select and optimise a model architecture from the ones suggested in the previous section, nested cross-validation is the recommended procedure to adopt [451]. This variant of cross-validation has been used with success to accurately evaluate and compare models' performance, whether they be GNNs or more classical machine learning models [441, 452].

Finally, in order to understand how relevant the topological features of metabolic networks are and foster further investigation into their role and significance, interpretability approaches such as SHAP [453] or PermFIT [454] could be adopted.

# Chapter 6

# Conclusion

In this dissertation, we have tried to provide a peek into the fast-moving research field of Systems biology, focusing in particular on genome-scale metabolic models and their applications in the area of precision medicine. In Chapter 1, we have tried to lay the foundation for the rest of the thesis, and introduced the data and models typically used in the field. We have also highlighted the possible problems of this research area, in particular in the interplay between the data and the machine learning models in a multimodal scenario. Even though this treatise can be considered anything but complete, we are confident that it should provide a good introduction to the methodological approaches we have examined in the subsequent chapters.

Starting from Chapter 2, we have investigated the application of GSMM-generated data, i.e. metabolic flux rates, in various machine learning scenarios. In this chapter we have shown how *in silico*-generated metabolic data contain different information than transcriptomic data, and that this information can be leveraged by machine learning models to outperform models which use transcriptomic data, when trying to predict gene regulatory associations. In particular, this was the case only when using the metabolic data within the proposed transfer learning scenario, which suggested that successfully extracting and exploiting this information is possible only under certain conditions/within specific frameworks.

Indeed this was confirmed in Chapter 3, when trying to predict growth rate for yeast *Saccharomyces cerevisiae* strains. In this chapter, we continued the investigation started in the previous one, by trying to understand whether the integration of fluxomic and gene expression data could improve models' performance. In particular, we compared several regularised linear models and neural network architectures in a regression setting, reaching the remarkable conclusion that the former can perform comparably well as the latter in these biological scenarios. However, we noticed how the goodness of

the outcome depends on the model adopted for the prediction, which is coherent with previous results [455].

In Chapter 4 we continued our research journey by investigating a more complex integrative approach in a classification setting. Instead of simply concatenating the omics, here we used Partial Least Squares Discriminant Analysis to account for the high dimensionality of the data and the small dataset size. What we found was that different omics combinations perform differently on patients of distinct clinical characteristics, meaning that a physician could, by simply looking at the patient's clinical data, decide which type of analysis to perform in order to maximise the probability of a correct diagnosis. The late integration strategy adopted in this case study proved effective, however other late integration strategies should be explored in the future, in order to probe the boundaries of this approach.

Finally, in Chapter 5 we speculated about a future research direction that could enrich the usage of GSMMs in machine learning applications for precision medicine. In this chapter we suggested a pipeline to exploit the topological information which is built into genome-scale metabolic models. We advocated for Graph Neural Networks as the go-to model architecture to adopt in order to utilise this type of information, and envisage that even though this direction has not been explored yet in this form, approaches taking advantage of metabolic topological information will advance to the forefront of the field.

Genome-scale metabolic models, with their inherent mechanistic transparency, are the ideal tool to develop a healthcare system tailored to the patients, but cannot be used by themselves because of the limited predictive power of the generated metabolic data. In conjunction with machine learning techniques, however, and integrated with other omics, their scope can be expanded, and as we have tried to convey in this work, this innovative approach is likely to keep its promise of better health.

# References


[1] Landeck L, Kneip C, Reischl J, Asadullah K. Biomarkers and personalized medicine: current status and further perspectives with special focus on dermatology. Experimental Dermatology. 2016;25(5):333-9.

[2] Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell. 2014;157(3):740-52.

[3] Partridge L, Deelen J, Slagboom PE. Facing up to the global challenges of ageing. Nature. 2018;561(7721):45-56.

[4] Marć M, Bartosiewicz A, Burzyńska J, Chmiel Z, Januszewicz P. A nursing shortage–a prospect of global and local policies. International nursing review. 2019;66(1):9-16.

[5] Ochi S, Tsubokura M, Kato S, Iwamoto S, Ogata S, Morita T, et al. Hospital staff shortage after the 2011 triple disaster in Fukushima, Japan-an earthquake, tsunamis, and nuclear power plant accident: a case of the Soso District. PloS one. 2016;11(10):e0164952.

[6] Smittenaar C, Petersen K, Stewart K, Moitt N. Cancer incidence and mortality projections in the UK until 2035. British journal of cancer. 2016;115(9):1147-55.

[7] Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, et al. Infectious disease in an era of global change. Nature Reviews Microbiology. 2022;20(4):193-205.

[8] Kench A, Janeja VP, Yesha Y, Rishe N, Grasso MA, Niskar A. Clinico-genomic data analytics for precision diagnosis and disease management. In: Healthcare Informatics (ICHI), 2015 International Conference on. IEEE; 2015. p. 263-71.

[9] Ho D, Quake SR, McCabe ER, Chng WJ, Chow EK, Ding X, et al. Enabling technologies for personalized and precision medicine. Trends in biotechnology. 2020;38(5):497-518.

[10] Eyassu F, Angione C. Modelling pyruvate dehydrogenase under hypoxia and its role in cancer metabolism. Royal Society open science. 2017;4(10):170360.

[11] Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. Cell metabolism. 2016;23(1):27-47.

[12] Zieba A, Grannas K, Söderberg O, Gullberg M, Nilsson M, Landegren U. Molecular tools for companion diagnostics. New biotechnology. 2012;29(6):634-40.

[13] Pacheco MP, Bintener T, Sauter T. Towards the network-based prediction of repurposed drugs using patient-specific metabolic models. EBioMedicine. 2019;43:26-7.

[14] Zarrinpar A, Lee DK, Silva A, Datta N, Kee T, Eriksen C, et al. Individualizing liver transplant immunosuppression using a phenotypic personalized medicine platform. Science translational medicine. 2016;8(333):333ra49-9.

[15] Blasiak A, Khong J, Kee T. CURATE. AI: optimizing personalized medicine with artificial intelligence. SLAS TECHNOLOGY: Translating Life Sciences Innovation. 2020;25(2):95-105.

[16] Lee DK, Chang VY, Kee T, Ho CM, Ho D. Optimizing combination therapy for acute lymphoblastic leukemia using a phenotypic personalized medicine digital health platform: Retrospective optimization individualizes patient regimens to maximize efficacy and safety. SLAS TECHNOLOGY: Translating Life Sciences Innovation. 2017;22(3):276-88.

[17] Zarrinpar A, Kim UB, Boominathan V. Phenotypic Response and Personalized Medicine in Liver Cancer and Transplantation: Approaches to Complex Systems. Advanced Therapeutics. 2020;3(4):1900167.

[18] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Transactions on Biomedical Engineering. 2017;64(2):263-73.

[19] Wang F, Zhang P, Dudley J. Healthcare Data Mining with Matrix Models. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 2137-8.

[20] Jatrniko W, Arsa DMS, Wisesa H, Jati G, Ma'sum MA. A review of big data analytics in the biomedical field. In: Big Data and Information Security (IWBIS), International Workshop on. IEEE; 2016. p. 31-41.

[21] Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearbook of medical informatics. 2014;9(1):14.

[22] Ceri S, Kaitoua A, Masseroli M, Pinoli P, Venco F. Data management for heterogeneous genomic datasets. IEEE/ACM transactions on computational biology and bioinformatics. 2016;14(6):1251-64.

[23] Montanari P, Bartolini I, Ciaccia P, Patella M, Ceri S, Masseroli M. Pattern similarity search in genomic sequences. IEEE Transactions on Knowledge and Data Engineering. 2016;28(11):3053-67.

[24] Behjati S, Tarpey PS. What is next generation sequencing? Archives of Disease in Childhood-Education and Practice. 2013;98(6):236-8.

[25] Shi Y, Kim S. Towards Information Analysis for Big Data. In: Control and Automation (CA), 2014 7th Conference on. IEEE; 2014. p. 3-5.

[26] Gupta A. Big data analysis using computational intelligence and Hadoop: a study. In: Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on. IEEE; 2015. p. 1397-401.

[27] Horgan RP, Kenny LC. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician & Gynaecologist. 2011;13(3):189-95.

[28] Saad M, He S, Thorstad W, Gay H, Barnett D, Zhao Y, et al. Learning-based Cancer Treatment Outcome Prognosis using Multimodal Biomarkers. IEEE Transactions on Radiation and Plasma Medical Sciences. 2021.

[29] Phillips KA, Trosman JR, Kelley RK, Pletcher MJ, Douglas MP, Weldon CB. Genomic sequencing: assessing the health care system, policy, and big-data implications. Health affairs. 2014;33(7):1246-53.

[30] O'Sullivan S, Jeanquartier F, Jean-Quartier C, Holzinger A, Shiebler D, Moon P, et al. Developments in AI and Machine Learning for Neuroimaging. Artificial Intelligence and Machine Learning for Digital Pathology. 2020:307-20.

[31] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563-77.

[32] Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Computational and Structural Biotechnology Journal. 2022;20:333-42.

[33] Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Scientific Reports. 2021;11(1):1-12.

[34] Schwarz E, Alnæs D, Andreassen OA, Cao H, Chen J, Degenhardt F, et al. Identifying multimodal signatures underlying the somatic comorbidity of psychosis: the COMMITMENT roadmap. Molecular Psychiatry. 2021;26(3):722-4.

[35] Shui L, Ren H, Yang X, Li J, Chen Z, Yi C, et al. The era of radiogenomics in precision medicine: an emerging approach to support diagnosis, treatment decisions, and prognostication in oncology. Frontiers in Oncology. 2021:3195.

[36] He S, Soraghan JJ, O'Reilly BF, Xing D. Quantitative analysis of facial paralysis using local binary patterns in biomedical videos. IEEE Transactions on Biomedical Engineering. 2009;56(7):1864-70.

[37] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020;580(7802):252-6.

[38] Neumann B, Walter T, Hériché JK, Bulkescher J, Erfle H, Conrad C, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature. 2010;464(7289):721-7.

[39] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 2019;6(1):54.

[40] Davidson EM, Poon MT, Casey A, Grivas A, Duma D, Dong H, et al. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. BMC medical imaging. 2021;21(1):1-13.

[41] Wood DA, Kafiabadi S, Al Busaidi A, Guilhem E, Lynch J, Townend M, et al. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer; 2020. p. 254-65.

[42] Barrett CL, Kim TY, Kim HU, Palsson BØ, Lee SY. Systems biology as a foundation for genome-scale synthetic biology. Current opinion in biotechnology. 2006;17(5):488-92.

[43] Mardinoglu A, Nielsen J. The Impact of Systems Medicine on Human Health and Disease. Frontiers in physiology. 2016;7:552.

[44] Palsson BØ. Systems biology: simulation of dynamic network states. Cambridge University Press; 2011.

[45] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Systems Biology. 2014;8(Suppl 2):I1.

[46] Ivanov O, van der Schaft A, Weissing FJ. Steady states and stability in metabolic networks without regulation. Journal of theoretical biology. 2016;401:78-93.

[47] Cairns RA, Harris IS, Mak TW. Regulation of cancer cell metabolism. Nature Reviews Cancer. 2011;11(2):85.

[48] Edwards LM. Metabolic systems biology: a brief primer. The Journal of physiology. 2017;595(9):2849-55.

[49] Palsson B. Systems biology. Cambridge University Press; 2015.

[50] Angione C. Human Systems Biology and Metabolic Modelling: A Review – From Disease Metabolism to Precision Medicine. BioMed Research International. 2019;2019(8304260).

[51] Vijayakumar S, Conway M, Lió P, Angione C. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. Briefings in bioinformatics. 2018;19(6):1218-35.

[52] Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology. 2012;10(4):291-305.

[53] Rawls KD, Dougherty BV, Blais EM, Stancliffe E, Kolling GL, Vinnakota K, et al. A simplified metabolic network reconstruction to promote understanding and development of flux balance analysis tools. Computers in Biology and Medicine. 2019;105:64-71.

[54] Biedendieck R, Borgmeier C, Bunk B, Stammen S, Scherling C, Meinhardt F, et al. Systems biology of recombinant protein production using Bacillus megaterium. In: Methods in enzymology. vol. 500. Elsevier; 2011. p. 165-95.

[55] Ryu JY, Kim HU, Lee SY. Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism. Proceedings of the National Academy of Sciences. 2017;114(45):E9740-9.

[56] Angione C. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. Bioinformatics. 2018;34(3):494-501.

[57] Ebrahim A, Brunk E, Tan J, O'brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. Nature communications. 2016;7(1):1-9.

[58] Fondi M, Liò P. Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. Microbiological research. 2015;171:52-64.

[59] Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, et al. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS computational biology. 2009;5(8):e1000489.

[60] Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. PLoS computational biology. 2019;15(7):e1007084.

[61] Domenzain I, Sánchez B, Anton M, Kerkhoven EJ, Millán-Oropeza A, Henry C, et al. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nature communications. 2022;13.

[62] Robaina Estévez S, Nikoloski Z. Generalized framework for context-specific metabolic model extraction methods. Frontiers in plant science. 2014;5:491.

[63] Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC systems biology. 2011;5(1):180.

[64] Puniya BL, Zhao Z, Helikar T. MADRID: a pipeline for MetAbolic Drug Repurposing IDentification. arXiv preprint arXiv:201102103. 2020.

[65] Töpfer N, Kleessen S, Nikoloski Z. Integration of metabolomics data into metabolic networks. Frontiers in plant science. 2015;6:49.

[66] Jensen K, Gudmundsson S, Herrgård MJ. Enhancing metabolic models with genome-scale experimental data. In: Systems biology. Springer; 2018. p. 337-50.

[67] Joyce AR, Palsson BØ. The model organism as a system: integrating'omics' data sets. Nature reviews Molecular cell biology. 2006;7(3):198.

[68] Aurich MK, Fleming RM, Thiele I. Metabotools: A comprehensive toolbox for analysis of genome-scale metabolic models. Frontiers in physiology. 2016;7:327.

[69] Bordbar A, Palsson BO. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. Journal of internal medicine. 2012;271(2):131-41.

[70] Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics. 2016;12(7):1-7.

[71] Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BØ. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC systems biology. 2010;4(1):140.

[72] Smallbone K. Striking a balance with Recon 2.1. arXiv preprint arXiv:13115696. 2013.

[73] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences. 2007;104(6):1777-82.

[74] Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature biotechnology. 2018;36(3):272-81.

[75] Robinson JL, Kocabaş P, Wang H, Cholley PE, Cook D, Nilsson A, et al. An atlas of human metabolism. Science signaling. 2020;13(624):eaaz1482.

[76] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.

[77] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic acids research. 2005;33(suppl_1):D54-8.

[78] Khodaee S, Asgari Y, Totonchi M, Karimi-Jafari MH. iMM1865: A new reconstruction of mouse genome-scale metabolic model. Scientific Reports. 2020;10(1):6177.

[79] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature biotechnology. 2010;28(3):245.

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

[81] Di Filippo M, Colombo R, Damiani C, Pescini D, Gaglio D, Vanoni M, et al. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. Computational biology and chemistry. 2016;62:60-9.

[82] Vivek-Ananth R, Samal A. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. Biosystems. 2016;147:1-10.

[83] Yilmaz LS, Walhout AJ. Metabolic network modeling with model organisms. Current opinion in chemical biology. 2017;36:32-9.

[84] Lewis NE, Abdel-Haleem AM. The evolution of genome-scale models of cancer metabolism. Frontiers in physiology. 2013;4:237.

[85] Schinn SM, Morrison C, Wei W, Zhang L, Lewis NE. Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells. Metabolic Engineering. 2021;66:21-30.

[86] Niedenführ S, Wiechert W, Nöh K. How to measure metabolic fluxes: a taxonomic guide for 13C fluxomics. Current opinion in biotechnology. 2015;34:82-90.

[87] Angione C, Lió P. Predictive analytics of environmental adaptability in multi-omic network models. Scientific reports. 2015;5:15147.

[88] Fernandes S, Robitaille J, Bastin G, Jolicoeur M, Wouwer AV. Dynamic metabolic flux analysis of underdetermined and overdetermined metabolic networks. IFAC-PapersOnLine. 2016;49(26):318-23.

[89] Rügen M, Bockmayr A, Steuer R. Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional FBA. Scientific reports. 2015;5:15247.

[90] Lularevic M, Racher AJ, Jaques C, Kiparissides A. Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. Biotechnology and bioengineering. 2019;116(9):2339-52.

[91] Ataman M, Hatzimanikatis V. Heading in the right direction: thermodynamics-based network analysis and pathway engineering. Current Opinion in Biotechnology. 2015;36:176-82.

[92] Willemsen AM, Hendrickx DM, Hoefsloot HC, Hendriks MM, Wahl SA, Teusink B, et al. MetDFBA: incorporating time-resolved metabolomics measurements into dynamic flux balance analysis. Molecular BioSystems. 2015;11(1):137-45.

[93] Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson OE, Palsson BO. Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Nature Communications. 2017;7:46249.

[94] Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic engineering. 2003;5(4):264-76.

[95] Scott WT, Smid EJ, Block DE, Notebaart RA. Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts. Microbial cell factories. 2021;20(1):1-15.

[96] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics. 2022;23(1):bbab454.

[97] Liu Q, Hu P. Extendable and explainable deep learning for pan-cancer radiogenomics research. Current opinion in chemical biology. 2022;66:102111.

[98] Purohit V, Wagner A, Yosef N, Kuchroo VK. Systems-based approaches to study immunometabolism. Cellular & Molecular Immunology. 2022:1-12.

[99] Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific reports. 2021;11(1):1-13.

[100] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research. 2018;24(6):1248-59.

[101] Molnar C. Interpretable machine learning. Lulu. com; 2020.

[102] Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. AI magazine. 2019;40(2):44-58.

[103] Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion. 2022;77:29-52.

[104] Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, et al. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Research. 2017;27(4):601-12.

[105] Riglar DT, Richmond DL, Potvin-Trottier L, Verdegaal AA, Naydich AD, Bakshi S, et al. Bacterial variability in the mammalian gut captured by a single-cell synthetic oscillator. Nature communications. 2019;10(1):4665.

[106] Woroszyło M, Ciecholewska-Juśko D, Junka A, Pruss A, Kwiatkowski P, Wardach M, et al. The impact of intraspecies variability on growth rate and cellular metabolic activity of bacteria exposed to rotating magnetic field. Pathogens. 2021;10(11):1427.

[107] Tonner PD, Darnell CL, Engelhardt BE, Schmid AK. Detecting differential growth of microbial populations with Gaussian process regression. Genome research. 2017;27(2):320-33.

[108] Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biology. 2021;22:1-22.

[109] Schulz C, Kumelj T, Karlsen E, Almaas E. Genome-scale metabolic modelling when changes in environmental conditions affect biomass composition. PLoS Computational Biology. 2021;17(5):e1008528.

[110] Mandakovic D, Cintolesi Á, Maldonado J, Mendoza SN, Aïte M, Gaete A, et al. Genome-scale metabolic models of Microbacterium species isolated from a high altitude desert environment. Scientific Reports. 2020;10(1):5560.

[111] Covert MW, Palsson BO. Constraints-based models: regulation of gene expression reduces the steady-state solution space. Journal of theoretical biology. 2003;221(3):309-25.

[112] den Besten HM, Aryani DC, Metselaar KI, Zwietering MH. Microbial variability in growth and heat resistance of a pathogen and a spoiler: all variabilities are equal but some are more equal than others. International Journal of Food Microbiology. 2017;240:24-31.

[113] van Rosmalen RP, Smith R, Dos Santos VM, Fleck C, Suarez-Diez M. Model reduction of genome-scale metabolic models as a basis for targeted kinetic models. Metabolic Engineering. 2021;64:74-84.

[114] Mishra S, Wang Z, Volk MJ, Zhao H. Design and application of a kinetic model of lipid metabolism in Saccharomyces cerevisiae. Metabolic Engineering. 2023;75:12-8.

[115] Boshagh F, Rostami K, van Niel EW. Application of kinetic models in dark fermentative hydrogen production–A critical review. International Journal of Hydrogen Energy. 2022.

[116] González-Ayala J, Calvo-Hernández A, Santillán M. Thermodynamic performance of coupled enzymatic reactions: A chemical kinetics model for analyzing cotransporters, ion pumps, and ATP synthases. Biophysical Chemistry. 2023;293:106932.

[117] Mu Y, Wang G, Yu HQ. Kinetic modeling of batch hydrogen production process by mixed anaerobic cultures. Bioresource Technology. 2006;97(11):1302-7.

[118] Schmiester L, Schälte Y, Fröhlich F, Hasenauer J, Weindl D. Efficient parameterization of large-scale dynamic models based on relative measurements. Bioinformatics. 2020;36(2):594-602.

[119] Smallbone K, Simeonidis E, Swainston N, Mendes P. Towards a genome-scale kinetic model of cellular metabolism. BMC systems biology. 2010;4(1):1-9.

[120] Ali Eshtewy N, Scholz L. Model reduction for kinetic models of biological systems. Symmetry. 2020;12(5):863.

[121] Zhang Y, Rajapakse JC. Machine learning in bioinformatics. vol. 4. John Wiley & Sons; 2009.

[122] Leung MK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE. 2016;104(1):176-97.

[123] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular systems biology. 2016;12(7):878.

[124] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Briefings in bioinformatics. 2017;18(5):851-69.

[125] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321.

[126] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface. 2018;15(141):20170387.

[127] Kitchin R. The data revolution: Big data, open data, data infrastructures and their consequences. SAGE Publishing; 2014.

[128] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

[129] Zeng ISL, Lumley T. Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science). Bioinformatics and Biology Insights. 2018;12:1177932218759292.

[130] Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. Neurocomputing. 2017;237:350-61.

[131] Cai Y, Gu H, Kenney T. Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization. Microbiome. 2017;5(1):110.

[132] Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering. 2014;40(1):16-28.

[133] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(5):849-911.

[134] Iuliano A, Occhipinti A, Angelini C, De Feis I, Liò P. Combining pathway identification and breast cancer survival prediction via screening-network methods. Frontiers in genetics. 2018;9:206.

[135] Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. The Journal of Machine Learning Research. 2009;10:2013-38.

[136] Fan J, Song R, et al. Sure independence screening in generalized linear models with NP-dimensionality. The Annals of Statistics. 2010;38(6):3567-604.

[137] Guo NL, Wan YW. Network-based identification of biomarkers coexpressed with multiple pathways. Cancer informatics. 2014;13:CIN-S14054.

[138] Grissa D, Pétéra M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E. Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. Frontiers in molecular biosciences. 2016;3:30.

[139] Kohavi R, John GH, et al. Wrappers for feature subset selection. Artificial intelligence. 1997;97(1-2):273-324.

[140] Goldberg DE. Genetic algorithms in search. Optimization, and MachineLearning. 1989.

[141] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

[142] Nandi S, Subramanian A, Sarkar RR. An integrative machine learning strategy for improved prediction of essential genes in Escherichia coli metabolism using flux-coupled features. Molecular BioSystems. 2017;13(8):1584-96.

[143] Maulud D, Abdulazeez AM. A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends. 2020;1(4):140-7.

[144] Alin A. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(3):370-4.

[145] Ray S. A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE; 2019. p. 35-9.

[146] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on neural networks. 2005;16(3):645-78.

[147] Occhipinti A, Eyassu F, Rahman TJ, Rahman PK, Angione C. In silico engineering of Pseudomonas metabolism reveals new biomarkers for increased biosurfactant production. PeerJ. 2018;6:e6046.

[148] Brunk E, George KW, Alonso-Gutierrez J, Thompson M, Baidoo E, Wang G, et al. Characterizing strain variation in engineered E. coli using a multi-omics-based workflow. Cell systems. 2016;2(5):335-46.

[149] Madala NE, Piater LA, Steenkamp PA, Dubery IA. Multivariate statistical models of metabolomic data reveals different metabolite distribution patterns in isonitrosoacetophenone-elicited Nicotiana tabacum and Sorghum bicolor cells. SpringerPlus. 2014;3(1):254.

[150] Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comparative and functional genomics. 2001;2(3):155-68.

[151] Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circulation research. 2017;121(9):1092-101.

[152] Berikol GB, Yildiz O, Özcan İT. Diagnosis of acute coronary syndrome with a support vector machine. Journal of medical systems. 2016;40(4):84.

[153] Mazurowski MA, Zhang J, Grimm LJ, Yoon SC, Silber JI. Radiogenomic analysis of breast cancer: luminal B molecular subtype is associated with enhancement dynamics at MR imaging. Radiology. 2014;273(2):365-72.

[154] de Jong TV, Moshkin YM, Guryev V. Gene expression variability: the other dimension in transcriptome analysis. Physiological genomics. 2019;51(5):145-58.

[155] Fair BJ, Blake LE, Sarkar A, Pavlovic BJ, Cuevas C, Gilad Y. Gene expression variability in human and chimpanzee populations share common determinants. Elife. 2020;9:e59929.

[156] Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. PLoS computational biology. 2010;6(8):e1000910.

[157] Nguyen A, Yoshida M, Goodarzi H, Tavazoie SF. Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. Nature communications. 2016;7(1):11246.

[158] Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, Oczko-Wojciechowska M, et al. Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer research. 2005;65(4):1587-97.

[159] Shoag J, Barbieri CE. Clinical variability and molecular heterogeneity in prostate cancer. Asian journal of andrology. 2016;18(4):543.

[160] Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546(7658):431-5.

[161] Wiggins GA, Black MA, Dunbier A, Morley-Bunker AE, kConFab Investigators, Pearson JF, et al. Increased gene expression variability in BRCA1-associated and basal-like breast tumours. Breast cancer research and treatment. 2021;189:363-75.

[162] Roberts AG, Catchpoole DR, Kennedy PJ. Identification of differentially distributed gene expression and distinct sets of cancer-related genes identified by changes in mean and variability. NAR Genomics and Bioinformatics. 2022;4(1):lqab124.

[163] Klepstad P, Fladvad T, Skorpen F, Bjordal K, Caraceni A, Dale O, et al. Influence from genetic variability on opioid use for cancer pain: a European genetic association study of 2294 cancer pain patients. Pain. 2011;152(5):1139-45.

[164] Grün D. Revealing dynamics of gene expression variability in cell state space. Nature methods. 2020;17(1):45-9.

[165] Ran D, Daye ZJ. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. Nucleic acids research. 2017;45(13):e127-7.

[166] Foreman R, Wollman R. Mammalian gene expression variability is explained by underlying cell state. Molecular systems biology. 2020;16(2):e9146.

[167] Mantsoki A, Devailly G, Joshi A. Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. Computational biology and chemistry. 2016;63:52-61.

[168] Ho JW, Stefani M, Dos Remedios CG, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics. 2008;24(13):i390-8.

[169] Kegerreis B, Catalina MD, Bachali P, Geraci NS, Labonte AC, Zeng C, et al. Machine learning approaches to predict lupus disease activity from gene expression data. Scientific reports. 2019;9(1):9617.

[170] Prabhakaran S, Azizi E, Carr A, Pe'er D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In: International conference on machine learning. PMLR; 2016. p. 1070-9.

[171] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

[172] McCulloch WS, Pitts WH. A Logical Calculus of the Ideas Immanent in Nervous Activity. Embodiments of Mind. 2016:19-38.

[173] Rosenbaltt F. The perceptron–a perciving and recognizing automation. Cornell Aeronautical Laboratory. 1957.

[174] Minsky ML, Papert SA. Perceptrons: expanded edition. MIT press; 1988.

[175] Ivakhnenko A. Cybernetics and forecasting techniques;.

[176] Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing. 2017;234:11-26.

[177] Bouwmans T, Javed S, Sultana M, Jung SK. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks. 2019;117:8-66.

[178] Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience. 2018;2018.

[179] Nisbet R, Elder J, Miner G. Basic Algorithms for Data Mining: A Brief Overview. Handbook of Statistical Analysis and Data Mining Applications. 2009:121-50.

[180] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. https://www.deeplearningbook.org.

[181] Calamuneri A, Donato L, Scimone C, Costa A, D'Angelo R, Sidoti A. On machine learning in biomedicine. Life Safety and Security. 2017;5(12):96-9.

[182] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems. 1989;2(4):303-14.

[183] Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM transactions on computational biology and bioinformatics. 2014;12(4):928-37.

[184] Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501-9.

[185] Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446-54.

[186] Chen R, Yang L, Goodison S, Sun Y. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics. 2020;36(5):1476-83.

[187] Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FC, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420).

[188] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013.

[189] Xu J, Wu P, Chen Y, Meng Q, Dawood H, Dawood H. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC bioinformatics. 2019;20(1):1-11.

[190] Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC bioinformatics. 2020;21:1-20.

[191] Simidjievski N, Bodnar C, Tariq I, Scherer P, Andres Terre H, Shams Z, et al. Variational autoencoders for cancer data integration: design principles and computational practice. Frontiers in genetics. 2019;10:1205.

[192] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541-51.

[193] Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest J, et al. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. arXiv preprint arXiv:200910549. 2020.

[194] Li C, Sun H, Liu Z, Wang M, Zheng H, Wang S. Learning Cross-Modal Deep Representations for Multi-Modal MR Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2019. p. 57-65.

[195] Zhou T, Ruan S, Canu S. A review: Deep learning for medical image segmentation using multi-modality fusion. Array. 2019;3:100004.

[196] Karthiga R, Narasimhan K. Automated diagnosis of breast cancer from ultrasound images using diverse ML techniques. Multimedia Tools and Applications. 2022:1-25.

[197] Maqsood S, Damaševičius R, Maskeliūnas R. TTCNN: A Breast Cancer Detection and Classification towards Computer-Aided Diagnosis Using Digital Mammography in Early Stages. Applied Sciences. 2022;12(7):3273.

[198] Ramalingam A, Aurchana P, Dhanalakshmi P, Vivekananadan K, Venkatachalapathy V. Analysis of Oral Squamous Cell Carcinoma into Various Stages using Pre-Trained Convolutional Neural Networks. In: IOP Conference Series: Materials Science and Engineering. vol. 993. IOP Publishing; 2020. p. 012058.

[199] Ayalew YA, Fante KA, Mohammed MA. Modified U-Net for liver cancer segmentation from computed tomography images with a new class balancing method. BMC Biomedical Engineering. 2021;3(1):1-13.

[200] Punn NS, Agarwal S. Modality specific U-Net variants for biomedical image segmentation: a survey. Artificial Intelligence Review. 2022:1-45.

[201] Li S, Han H, Sui D, Hao A, Qin H. A novel radiogenomics framework for genomic and image feature correlation using deep learning. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 899-906.

[202] Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nature Methods. 2018.

[203] Guo W, Xu Y, Feng X. DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing. arXiv preprint arXiv:170503094. 2017.

[204] Mohmad Yousoff SN, Baharin A, Abdullah A. Differential Search Algorithm in Deep Neural Network for the Predictive Analysis of Xylitol Production in Escherichia Coli. In: Mohamed Ali MS, Wahid H, Mohd Subha NA, Sahlan S, Md Yunus MA, Wahap AR, editors. Modeling, Design and Simulation of Systems. Singapore: Springer Singapore; 2017. p. 53-67.

[205] Clauwaert J, Menschaert G, Waegeman W. DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic acids research. 2019;47(6):e36-6.

[206] Sun S. A survey of multi-view machine learning. Neural Computing and Applications. 2013;23(7-8):2031-8.

[207] Wang Xl, Li Jy, Liu Y, Wang Yf, Zhao Ds. Building localized bioinformatics platform based on Galaxy and high performance computing cluster. In: Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on. IEEE; 2013. p. 712-6.

[208] Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A. Disaggregating asthma: Big investigation versus big data. Journal of Allergy and Clinical Immunology. 2017;139(2):400-7.

[209] Adossa N, Khan S, Rytkönen KT, Elo LL. Computational strategies for single-cell multi-omics integration. Computational and Structural Biotechnology Journal. 2021.

[210] Serra A, Fratello M, Fortino V, Raiconi G, Tagliaferri R, Greco D. MVDA: a multi-view genomic data integration methodology. BMC bioinformatics. 2015;16(1):261.

[211] Gardiner LJ, Carrieri AP, Bingham K, Macluskie G, Bunton D, McNeil M, et al. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. PloS one. 2022;17(2):e0263248.

[212] Moon S, Lee H. MOMA: A Multi-Task Attention Learning Algorithm for Multi-Omics Data Interpretation and Classification. Bioinformatics. 2022.

[213] Wang J, Lu CH, Kong XZ, Dai LY, Yuan S, Zhang X. Multi-view manifold regularized compact low-rank representation for cancer samples clustering on multi-omics data. BMC bioinformatics. 2022;22(12):1-22.

[214] Li X, Ma J, Leng L, Han M, Li M, He F, et al. MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. Frontiers in Genetics. 2022:127.

[215] Hu Y, Zhao L, Li Z, Dong X, Xu T, Zhao Y. Classifying the multi-omics data of gastric cancer using a deep feature selection method. Expert Systems with Applications. 2022:116813.

[216] Lee JW, Su Y, Baloni P, Chen D, Pavlovitch-Bedzyk AJ, Yuan D, et al. Integrated analysis of plasma and single immune cells uncovers metabolic changes in individuals with COVID-19. Nature biotechnology. 2022;40(1):110-20.

[217] Xiao Y, Ma D, Yang YS, Yang F, Ding JH, Gong Y, et al. Comprehensive metabolomics expands precision medicine for triple-negative breast cancer. Cell Research. 2022:1-14.

[218] Lewis JE, Kemp ML. Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. Nature communications. 2021;12(1):1-14.

[219] Li B, Wang T, Nabavi S. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; 2021. p. 1-9.

[220] Albaradei S, Napolitano F, Thafar MA, Gojobori T, Essack M, Gao X. MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. Computational and Structural Biotechnology Journal. 2021.

[221] Abdelaziz M, Wang T, Elazab A. Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. Journal of Biomedical Informatics. 2021;121:103863.

[222] Xu C, Liu D, Zhang L, Xu Z, He W, Jiang H, et al. AutoOmics: New multimodal approach for multi-omics research. Artificial Intelligence in the Life Sciences. 2021;1:100012.

[223] Zhang X, Wang J, Lu J, Su L, Wang C, Huang Y, et al. Robust Prognostic Subtyping of Muscle-Invasive Bladder Cancer Revealed by Deep Learning-Based Multi-Omics Data Integration. Frontiers in Oncology. 2021;11.

[224] Dutta P, Patra AP, Saha S. DeePROG: Deep Attention-based Model for Diseased Gene Prognosis by Fusing Multi-omics Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021.

[225] Wang W, Zhang X, Dai DQ. DeFusion: a denoised network regularization framework for multi-omics integration. Briefings in Bioinformatics. 2021;22(5):bbab057.

[226] Zhang X, Xing Y, Sun K, Guo Y. OmiEmbed: a unified multi-task deep learning framework for multi-omics data. Cancers. 2021;13(12):3047.

[227] Uzunangelov V, Wong CK, Stuart JM. Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. PLoS computational biology. 2021;17(4):e1008878.

[228] Zhang M, Young GS, Chen H, Li J, Qin L, McFaline-Figueroa JR, et al. Deep-learning detection of cancer metastases to the brain on MRI. Journal of Magnetic Resonance Imaging. 2020;52(4):1227-36.

[229] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. Moronet: multi-omics integration via graph convolutional networks for biomedical data classification. bioRxiv. 2020.

[230] Zhou T, Thung KH, Zhu X, Shen D. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Human brain mapping. 2019;40(3):1001-16.

[231] Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). BMC genomics. 2019;20(11):1-11.

[232] Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In: Pacific Symposium on Biocomputing 2020. World Scientific; 2019. p. 355-66.

[233] Badic B, Hatt M, Durand S, Le Jossic-Corcos C, Simon B, Visvikis D, et al. Radiogenomics-based cancer prognosis in colorectal cancer. Scientific reports. 2019;9(1):1-7.

[234] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in genetics. 2019;10:166.

[235] Yu KH, Berry GJ, Rubin DL, Re C, Altman RB, Snyder M. Association of omics features with histopathology patterns in lung adenocarcinoma. Cell systems. 2017;5(6):620-7.

[236] Zhu X, Yao J, Xiao G, Xie Y, Rodriguez-Canales J, Parra ER, et al. Imaging-genetic data mapping for clinical outcome prediction via supervised conditional gaussian graphical model. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2016. p. 455-9.

[237] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE. 2015;103(9):1449-77.

[238] Maglanoc LA, Kaufmann T, Jonassen R, Hilland E, Beck D, Landrø NI, et al. Multimodal fusion of structural and functional brain imaging in depression using linked independent component analysis. Human Brain Mapping. 2020;41(1):241-55.

[239] Sertbas M, Ulgen KO. Unlocking human brain metabolism by genome-scale and multiomics metabolic models: relevance for neurology research, health, and disease. OMICS: A Journal of Integrative Biology. 2018;22(7):455-67.

[240] Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances. 2021;49:107739.

[241] Borhani N, Ghaisari J, Abedi M, Kamali M, Gheisari Y. A deep learning approach to predict inter-omics interactions in multi-layer networks. BMC bioinformatics. 2022;23(1):1-17.

[242] Lejeune FX, Ichou F, Camenen E, Colsch B, Mauger F, Peltier C, et al. A Multimodal Omics Exploration of the Motor and Non-Motor Symptoms of Parkinson's Disease. International Journal of Translational Medicine. 2022;2(1):97-112.

[243] Nguyen ND, Huang J, Wang D. A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data. Nature Computational Science. 2022;2(1):38-46.

[244] Anžel A, Heider D, Hattab G. MOVIS: A multi-omics software solution for multi-modal time-series clustering, embedding, and visualizing tasks. Computational and Structural Biotechnology Journal. 2022.

[245] Bredikhin D, Kats I, Stegle O. Muon: multimodal omics analysis framework. Genome Biology. 2022;23(1):1-12.

[246] Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology. 2022:1-4.

[247] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications. 2021;12(1):1-13.

[248] Foran DJ, Chen W, Yang L. Automated image interpretation and computer-assisted diagnostics. Stud Health Technol Inform. 2013;185:77-108.

[249] Tixier F, Cheze-Le-Rest C, Schick U, Simon B, Dufour X, Key S, et al. Transcriptomics in cancer revealed by Positron Emission Tomography radiomics. Scientific reports. 2020;10(1):1-11.

[250] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics. 2022;23(2):bbab569.

[251] Yue T, Jia X, Petrosino J, Sun L, Fan Z, Fine J, et al. Computational integration of nanoscale physical biomarkers and cognitive assessments for Alzheimer's disease diagnosis and prognosis. Science advances. 2017;3(7):e1700669.

[252] Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. Radiology. 2012;264(2):387-96.

[253] Lin W, Gao Q, Du M, Chen W, Tong T. Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. Computers in Biology and Medicine. 2021;134:104478.

[254] Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920-33.

[255] Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proceedings of the National Academy of Sciences. 2020;117(31):18869-79.

[256] Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature communications. 2018;9(1):1-10.

[257] Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO. A biochemically-interpretable machine learning classifier for microbial GWAS. Nature communications. 2020;11(1):1-11.

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

[259] Chien J, Larsen P. Predicting the Plant Root-Associated Ecological Niche of 21 Pseudomonas Species Using Machine Learning and Metabolic Modeling. arXiv preprint arXiv:170103220. 2017.

[260] Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E. Metabolic Network Prediction of Drug Side Effects. Cell Systems. 2016;2(3):209-13.

[261] Ajjolli Nagaraja A, Fontaine N, Delsaut M, Charton P, Damour C, Offmann B, et al. Flux prediction using artificial neural network (ANN) for the upper part of glycolysis. PloS one. 2019;14(5):e0216178.

[262] Wu SG, Wang Y, Jiang W, Oyetunde T, Yao R, Zhang X, et al. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS computational biology. 2016;12(4):e1004838.

[263] Roy S, Radivojevic T, Forrer M, Marti JM, Jonnalagadda V, Backman T, et al. Multiomics data collection, visualization, and utilization for guiding metabolic engineering. Frontiers in bioengineering and biotechnology. 2021;9:45.

[264] Vijayakumar S, Conway M, Lió P, Angione C. Optimization of multi-omic genome-scale models: Methodologies, hands-on tutorial, and perspectives. In: Metabolic Network Reconstruction and Modeling. Springer; 2018. p. 389-408.

[265] Angione C, Costanza J, Carapezza G, Lió P, Nicosia G. Multi-target analysis and design of mitochondrial metabolism. PloS one. 2015;10(9):e0133825.

[266] Choon YW, Mohamad MS, Deris S, Chong CK, Omatu S, Corchado JM. Gene knockout identification using an extension of bees hill flux balance analysis. BioMed research international. 2015;2015.

[267] Daud KM, Mohamad MS, Zakaria Z, Hassan R, Shah ZA, Deris S, et al. A non-dominated sorting Differential Search Algorithm Flux Balance Analysis (ndsDSAFBA) for in silico multi-objective optimization in identifying reactions knockout. Computers in biology and medicine. 2019;113:103390.

[268] Occhipinti A, Hamadi Y, Kugler H, Wintersteiger C, Yordanov B, Angione C. Discovering Essential Multiple Gene Effects through Large Scale Optimization: an Application to Human Cancer Metabolism. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020.

[269] Cavill R, Keun HC, Holmes E, Lindon JC, Nicholson JK, Ebbels TM. Genetic algorithms for simultaneous variable and sample selection in metabonomics. Bioinformatics. 2009;25(1):112-8.

[270] Yaneske E, Angione C. The poly-omics of ageing through individual-based metabolic modelling. BMC bioinformatics. 2018;19(14):83-96.

[271] Jalili M, Scharm M, Wolkenhauer O, Damaghi M, Salehzadeh-Yazdi A. Exploring the Metabolic Heterogeneity of Cancers: A Benchmark Study of Context-Specific Models. Journal of Personalized Medicine. 2021;11(6):496.

[272] Vijayakumar S, Rahman PKMSM, Angione C. A hybrid flux balance analysis and machine learning pipeline elucidates the metabolic response of cyanobacteria to different growth conditions. iScience. 2020:101818.

[273] Vijayakumar S, Angione C. Protocol for hybrid flux balance, statistical, and machine learning analysis of multi-omic data from the cyanobacterium Synechococcus sp. PCC 7002. STAR protocols. 2021;2(4):100837.

[274] Nielsen J. Systems biology of metabolism: a driver for developing personalized and precision medicine. Cell metabolism. 2017;25(3):572-9.

[275] Gilpin W, Huang Y, Forger DB. Learning dynamics from large biological datasets: machine learning meets systems biology. Current Opinion in Systems Biology. 2020.

[276] Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, et al. A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell. 2019;177(6):1649-61.

[277] Zhang J, Petersen SD, Radivojevic T, Ramirez A, Pérez-Manríquez A, Abeliuk E, et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. Nature communications. 2020;11(1):1-13.

[278] Occhipinti A, Angione C. A Computational Model of Cancer Metabolism for Personalised Medicine. In: Building Bridges in Medical Science 2021. Cambridge Medical Journal; 2021. .

[279] Ben Guebila M, Thiele I. Predicting gastrointestinal drug effects using contextualized metabolic models. PLoS computational biology. 2019;15(6):e1007100.

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

[281] Sahu A, Blätke MA, Szymański JJ, Töpfer N. Advances in Flux Balance Analysis by Integrating Machine Learning and Mechanism-based Models. Computational and Structural Biotechnology Journal. 2021.

[282] Magazzù G, Zampieri G, Angione C. Multimodal regularised linear models with flux balance analysis for mechanistic integration of omics data. Bioinformatics. 2021.

[283] Samal SS, Radulescu O, Weber A, Fröhlich H. Linking metabolic network features to phenotypes using sparse group lasso. Bioinformatics. 2017;33(21):3445-53.

[284] Cuperlovic-Culf M. Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. Metabolites. 2018;8(1):4.

[285] Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural computation. 1996;8(7):1341-90.

[286] Costa FF. Big data in biomedicine. Drug discovery today. 2014;19(4):433-40.

[287] Kuo MH, Sahama T, Kushniruk AW, Borycki EM, Grunwell DK. Health big data analytics: current perspectives, challenges and potential solutions. International Journal of Big Data Intelligence. 2014;1(1-2):114-26.

[288] Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. Computers & Industrial Engineering. 2020:106854.

[289] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research. 2020:1-19.

[290] Grama M, Musat M, Muñoz-González L, Passerat-Palmbach J, Rueckert D, Alansary A. Robust aggregation for adaptive privacy preserving federated learning in healthcare. arXiv preprint arXiv:200908294. 2020.

[291] Erdrich P, Steuer R, Klamt S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC systems biology. 2015;9(1):1-12.

[292] Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology. 2010;6(1):390.

[293] Yurkovich JT, Palsson BO. Quantitative-omic data empowers bottom-up systems biology. Current opinion in biotechnology. 2018;51:130-6.

[294] Santiago-Rodriguez TM, Hollister EB. Multi 'omic data integration: A review of concepts, considerations, and approaches. In: Seminars in Perinatology. vol. 45. Elsevier; 2021. p. 151456.

[295] Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: From computational needs to data mining and sharing. Frontiers in Genetics. 2020:1598.

[296] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research. 2018;46(20):10546-62.

[297] Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One. 2018;13(8):e0202344.

[298] Iuliano A, Occhipinti A, Angelini C, De Feis I, Liò P. Cosmonet: An r package for survival analysis using screening-network methods. Mathematics. 2021;9(24):3262.

[299] Martínez Arbas S, Busi SB, Queirós P, De Nies L, Herold M, May P, et al. Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies. Frontiers in Genetics. 2021;12:858.

[300] López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the integration of omics and non-omics data. Genes. 2019;10(3):238.

[301] Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, et al. Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. PLoS genetics. 2015;11(12):e1005689.

[302] Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nature chemical biology. 2010;6(11):787-9.

[303] Yang MQ, Weissman SM, Yang W, Zhang J, Canaann A, Guan R. MISC: missing imputation for single-cell RNA sequencing data. BMC systems biology. 2018;12(7):55-63.

[304] Zhou X, Chai H, Zhao H, Luo CH, Yang Y. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network. GigaScience. 2020;9(7):giaa076.

[305] Zhu Q, Li H, Ye H, Zhang Z, Wang R, Fan Z, et al. Incomplete multi-modal brain image fusion for epilepsy classification. Information Sciences. 2022;582:316-33.

[306] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.

[307] Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020;586(7829):E14-6.

[308] Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, et al. The AIMe registry for artificial intelligence in biomedical research. Nature methods. 2021;18(10):1128-31.

[309] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.

[310] Lamprecht AL, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Science. 2020;3(1):37-59.

[311] Vijayakumar S, Magazzù G, Moon P, Occhipinti A, Angione C. A Practical Guide to Integrating Multimodal Machine Learning and Metabolic Modeling. In: Computational Systems Biology in Medicine and Biotechnology. Springer; 2022. p. 87-122.

[312] Davidson EH. Emerging properties of animal gene regulatory networks. Nature. 2010;468(7326):911-20.

[313] Ye Y, Li SL, Wang SY. Construction and analysis of mRNA, miRNA, lncRNA, and TF regulatory networks reveal the key genes associated with prostate cancer. PloS one. 2018;13(8):e0198055.

[314] Gardner TS, Di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003;301(5629):102-5.

[315] Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008;9(10):770-80.

[316] Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology. 2005;4(1).

[317] Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005;21(1):71-9.

[318] Pio G, Ceci M, Prisciandaro F, Malerba D. Exploiting causality in gene network reconstruction based on graph embedding. Mach Learn. 2020;109(6):1231-79.

[319] Luo Q, Liu X, Yi D. Reconstructing Gene Networks from Microarray Time-Series Data via Granger Causality. In: Zhou J, editor. Complex Sciences. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 196-209.

[320] Mignone P, Pio G, D'Elia D, Ceci M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinform. 2020;36(5):1553-61.

[321] Yeang CH, Vingron M. A joint model of regulatory and metabolic networks. BMC bioinformatics. 2006;7(1):332.

[322] Yurkovich JT, Palsson BO. Solving puzzles with missing pieces: the power of systems biology. Proceedings of the IEEE. 2015;104(1):2-7.

[323] Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences. 2010;107(41):17845-50.

[324] Motamedian E, Mohammadi M, Shojaosadati SA, et al. TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. Bioinform. 2017;33(7):1057-63.

[325] Richelle A, Kellman BP, Wenzel AT, Chiang AW, Reagan T, Gutierrez JM, et al. What does your cell really do? Model-based assessment of mammalian cells metabolic functionalities using omics data. bioRxiv. 2020.

[326] Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC bioinformatics. 2007;8(S6):S9.

[327] Wang Z, Danziger SA, Heavner BD, Ma S, Smith JJ, Li S, et al. Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. PLoS computational biology. 2017;13(5):e1005489.

[328] Mignone P, Pio G, Džeroski S, Ceci M. Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks. Scientific Reports. 2020;10:22295.

[329] Marbach D, Costello JC, Küffner R, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012 Jul;9(8):796-804.

[330] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 04;4(2):249-64.

[331] Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic acids research. 2016:gkw1033.

[332] Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database. 2015;2015.

[333] Motamedian E, Taheri E, Bagheri F. Proliferation inhibition of cisplatin-resistant ovarian cancer cells using drugs screened by integrating a metabolic model and transcriptomic data. Cell proliferation. 2017;50(6):e12370.

[334] Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. Nature protocols. 2019;14(3):639-702.

[335] Levatic J, Ceci M, Kocev D, Dzeroski S. Semi-supervised classification trees. J Intell Inf Syst. 2017;49(3):461-86.

[336] Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006;34(Database-Issue):535-9.

[337] Kaufmann J, Asalone K, Corizzo R, Saldanha C, Bracht J, Japkowicz N. One-Class Ensembles for Rare Genomic Sequences Identification. In: International Conference on Discovery Science. Springer; 2020. p. 340-54.

[338] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Comput Appl Math. 1987;20:53 65.

[339] Levatic J, Kocev D, Ceci M, Dzeroski S. Semi-supervised trees for multi-target regression. Inf Sci. 2018;450:109-27. Available from: https://doi.org/10.1016/j.ins.2018.03.033.

[340] Corizzo R, Pio G, Ceci M, Malerba D. DENCAST: distributed density-based clustering for multi-target regression. J Big Data. 2019;6:43.

[341] Pio G, Serafino F, Malerba D, Ceci M. Multi-type clustering and classification from heterogeneous networks. Inf Sci. 2018;425:107-26.

[342] Serafino F, Pio G, Ceci M. Ensemble Learning for Multi-Type Classification in Heterogeneous Networks. IEEE Trans Knowl Data Eng. 2018;30(12):2326-39.

[343] Ji M, Sun Y, Danilevsky M, Han J, Gao J. Graph Regularized Transductive Classification on Heterogeneous Information Networks. In: Machine Learning and Knowledge Discovery in Databases. Springer Berlin; 2010. p. 570-86.

[344] Ma Y, Bai S, An S, Liu W, Liu A, Zhen X, et al. Transductive Relation-Propagation Network for Few-shot Learning. In: IJCAI 20; 2020. p. 804-10.

[345] Petković M, Džeroski S, Kocev D. Feature Ranking for Multi-target Regression with Tree Ensemble Methods. In: Proc. of DS 2017; 2017. p. 171-85.

[346] Chawla NV, Karakoulas G. Learning from labeled and unlabeled data: An empirical study across techniques and domains. Journal of Artificial Intelligence Research. 2005;23:331-66.

[347] Pio G, Mignone P, Magazzù G, Zampieri G, Ceci M, Angione C. Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction. Bioinformatics. 2022;38(2):487-93.

[348] Dikicioglu D, Pir P, Oliver SG. Predicting complex phenotype–genotype interactions to enable yeast engineering: Saccharomyces cerevisiae as a model organism and a cell factory. Biotechnology journal. 2013;8(9):1017-34.

[349] Lian J, Mishra S, Zhao H. Recent advances in metabolic engineering of Saccharomyces cerevisiae: New tools and their applications. Metabolic Engineering. 2018;50:85-108.

[350] Bao Z, HamediRad M, Xue P, Xiao H, Tasan I, Chao R, et al. Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. Nature biotechnology. 2018;36(6):505.

[351] Rantasalo A, Landowski CP, Kuivanen J, Korppoo A, Reuter L, Koivistoinen O, et al. A universal gene expression system for fungi. Nucleic acids research. 2018;46(18):e111-1.

[352] Gardner TS. Synthetic biology: from hype to impact. Trends in biotechnology. 2013;31(3):123-5.

[353] Airoldi EM, Huttenhower C, Gresham D, Lu C, Caudy AA, Dunham MJ, et al. Predicting cellular growth from gene expression signatures. PLoS Computational Biology. 2009;5(1):e1000257.

[354] Wytock TP, Motter AE. Predicting growth rate from gene expression. Proceedings of the National Academy of Sciences. 2019;116(2):367-72.

[355] Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics. 2016;19(2):325-40.

[356] Mignone P, Pio G, D'Elia D, Ceci M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. Bioinformatics. 2020;36(5):1553-61.

[357] Castillo S, Patil KR, Jouhten P. Yeast Genome-Scale Metabolic Models for Simulating Genotype–Phenotype Relations. Yeasts in Biotechnology and Human Health: Physiological Genomic Approaches. 2019:111.

[358] Pelechano V, Pérez-Ortín JE. There is a steady-state transcriptome in exponentially growing yeast cells. Yeast. 2010;27(7):413-22.

[359] Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics. 2005;21(9):2067-75.

[360] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic acids research. 2002;30(4):e15-5.

[361] O'Duibhir E, Lijnzaad P, Benschop JJ, Lenstra TL, van Leenen D, Koerkamp MJG, et al. Cell cycle population effects in perturbation studies. Molecular systems biology. 2014;10(6):732.

[362] Chowdhury R, Chowdhury A, Maranas CD. Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. Metabolites. 2015;5(4):536-70.

[363] Yeast Drop-out Mix Complete media; 2018. Accessed : 16/01/2018. https://www.usbio.net/media/D9515.

[364] Yeast Nitrogen Base (YNB) media; 2018. Accessed : 16/01/2018. https://www.usbio.net/media/Y2025.

[365] Boulesteix AL, De Bin R, Jiang X, Fuchs M. IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. Computational and mathematical methods in medicine. 2017;2017.

[366] Tay JK, Friedman J, Tibshirani R. Principal component-guided sparse regression. arXiv preprint arXiv:181004651. 2018.

[367] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(1):49-67.

[368] Hernández-Orallo J. ROC curves for regression. Pattern Recognition. 2013;46(12):3395-411.

[369] Griac P. Regulation of yeast phospholipid biosynthetic genes in phosphatidylserine decarboxylase mutants. Journal of bacteriology. 1997;179(18):5843-8.

[370] Kuge O, Nishijima M, Akamatsu Y. Phosphatidylserine biosynthesis in cultured Chinese hamster ovary cells. III. Genetic evidence for utilization of phosphatidylcholine and phosphatidylethanolamine as precursors. Journal of Biological Chemistry. 1986;261(13):5795-8.

[371] Kodaki T, Yamashita S. Characterization of the methyltransferases in the yeast phosphatidylethanolamine methylation pathway by selective gene disruption. European journal of biochemistry. 1989;185(2):243-51.

[372] Sutter BM, Wu X, Laxman S, Tu BP. Methionine inhibits autophagy and promotes growth by inducing the SAM-responsive methylation of PP2A. Cell. 2013;154(2):403-15.

[373] Yoshida S, Imoto J, Minato T, Oouchi R, Kamada Y, Tomita M, et al. A novel mechanism regulates H2S and SO2 production in Saccharomyces cerevisiae. Yeast. 2011;28(2):109-21.

[374] Feng J, Polychronidis G, Heger U, Frongia G, Mehrabi A, Hoffmann K. Incidence trends and survival prediction of hepatoblastoma in children: a population-based study. Cancer communications. 2019;39(1):1-9.

[375] Bharti S, Bharti JN, Sinha A, Yadav T. Common and rare histological variants of hepatoblastoma in children: a pathological diagnosis and review of the literature. Gastrointestinal tumors. 2021;8(2):41-6.

[376] Kubota N, Fujiwara N, Hoshida Y. Clinical and molecular prediction of hepatocellular carcinoma risk. Journal of Clinical Medicine. 2020;9(12):3843.

[377] Sumazin P, Chen Y, Treviño LR, Sarabia SF, Hampton OA, Patel K, et al. Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups. Hepatology. 2017;65(1):104-21.

[378] Naeem S, Ali A, Qadri S, Khan Mashwani W, Tairan N, Shah H, et al. Machine-Learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images. Applied Sciences. 2020;10(9):3134.

[379] de Senneville BD, Khoubai FZ, Bevilacqua M, Labedade A, Flosseau K, Chardot C, et al. Deciphering Tumour Tissue Organization by 3D Electron Microscopy and machine learning. bioRxiv. 2021.

[380] Khan RA, Luo Y, Wu FX. Machine learning based liver disease diagnosis: A systematic review. Neurocomputing. 2022;468:492-509.

[381] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

[382] Kremer N, Walther AE, Tiao GM. Management of hepatoblastoma: an update. Current opinion in pediatrics. 2014;26(3):362-9.

[383] Xie L, Evangelidis T, Xie L, Bourne PE. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS computational biology. 2011;7(4):e1002037.

[384] Rahman M, Islam T, Gov E, Turanli B, Gulfidan G, Shahjaman M, et al. Identification of prognostic biomarker signatures and candidate drugs in colorectal cancer: insights from systems biology analysis. Medicina. 2019;55(1):20.

[385] Lai X, Eberhardt M, Schmitz U, Vera J. Systems biology-based investigation of cooperating microRNAs as monotherapy or adjuvant therapy in cancer. Nucleic acids research. 2019;47(15):7753-66.

[386] Ray B, Henaff M, Ma S, Efstathiadis E, Peskin ER, Picone M, et al. Information content and analysis methods for multi-modal high-throughput biomedical data. Scientific reports. 2014;4(1):1-10.

[387] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.

[388] Cairo S, Armengol C, De Reyniès A, Wei Y, Thomas E, Renard CA, et al. Hepatic stem-like phenotype and interplay of Wnt/$\beta$-catenin and Myc signaling in aggressive childhood liver cancer. Cancer cell. 2008;14(6):471-84.

[389] Fu J, Zhang Y, Wang Y, Zhang H, Liu J, Tang J, et al. Optimization of metabolomic data processing using NOREVA. Nature Protocols. 2022;17(1):129-51.

[390] Tang J, Fu J, Wang Y, Li B, Li Y, Yang Q, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Briefings in Bioinformatics. 2020;21(2):621-36.

[391] Goh WWB, Yong CH, Wong L. Are batch effects still relevant in the age of big data? Trends in Biotechnology. 2022.

[392] Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications. 2014;5(1):1-11.

[393] Yizhak K, Gaude E, Le Dévédec S, Waldman YY, Stein GY, van de Water B, et al. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. Elife. 2014;3:e03641.

[394] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357(6352).

[395] Nilsson A, Nielsen J. Genome scale metabolic modeling of cancer. Metabolic engineering. 2017;43:103-12.

[396] Cho JS, Gu C, Han TH, Ryu JY, Lee SY. Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring. Current Opinion in Systems Biology. 2019;15:1-11.

[397] Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T. Predicting selective drug targets in cancer through metabolic networks. Molecular systems biology. 2011;7(1):501.

[398] Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J. Identification of anti-cancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology. 2014;10(3):721.

[399] Firczuk H, Kannambath S, Pahle J, Claydon A, Beynon R, Duncan J, et al. An in vivo control map for the eukaryotic mRNA translation machinery. Molecular systems biology. 2013;9(1):635.

[400] Paltanea M, Tabirca S, Scheiber E, Tangney M. Logarithmic growth in biological processes. In: 2010 12th International Conference on Computer Modelling and Simulation. IEEE; 2010. p. 116-21.

[401] Chan C, Berthiaume F, Lee K, Yarmush ML. Metabolic flux analysis of cultured hepatocytes exposed to plasma. Biotechnology and bioengineering. 2003;81(1):33-49.

[402] Dai Z, Yang S, Xu L, Hu H, Liao K, Wang J, et al. Identification of Cancer–associated metabolic vulnerabilities by modeling multi-objective optimality in metabolism. Cell Communication and Signaling. 2019;17(1):1-15.

[403] Fabris F, Palmer D, de Magalhães JP, Freitas AA. Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. Briefings in bioinformatics. 2020;21(3):803-14.

[404] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[405] Dinh PH. Combining gabor energy with equilibrium optimizer algorithm for multi-modality medical image fusion. Biomedical Signal Processing and Control. 2021;68:102696.

[406] Dinh PH. A novel approach based on grasshopper optimization algorithm for medical image fusion. Expert Systems with Applications. 2021;171:114576.

[407] Zhou J, Qiu Y, Zhu S, Armaghani DJ, Li C, Nguyen H, et al. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence. 2021;97:104015.

[408] Ojala M, Garriga GC. Permutation tests for studying classifier performance. Journal of Machine Learning Research. 2010;11(6).

[409] Valente G, Castellanos AL, Hausfeld L, De Martino F, Formisano E. Cross-validation and permutations in MVPA: validity of permutation strategies and power of cross-validation schemes. NeuroImage. 2021:118145.

[410] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

[411] Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2020 11;49(D1):D480-9. Available from: https://doi.org/10.1093/nar/gkaa1100.

[412] Brosnan ME, Brosnan JT. Hepatic glutamate metabolism: a tale of 2 hepatocytes. The American journal of clinical nutrition. 2009;90(3):857S-861S.

[413] Dunlop RJ, Campbell CW. Cytokines and advanced cancer. Journal of pain and symptom management. 2000;20(3):214-32.

[414] Vidal-Vanaclocha F, Amézaga C, Asumendi A, Kaplanski G, Dinarello CA. Interleukin-1 receptor blockade reduces the number and size of murine B16 melanoma hepatic metastases. Cancer research. 1994;54(10):2667-72.

[415] Shimonosono M, Arigami T, Yanagita S, Matsushita D, Uchikado Y, Kijima Y, et al. The association of human endogenous retrovirus-H long terminal repeat-associating protein 2 (HHLA2) expression with gastric cancer prognosis. Oncotarget. 2018;9(31):22069.

[416] Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single–cell type transcriptomics map of human tissues. Science Advances. 2021;7(31):eabh2169.

[417] Dai Y, Wang Y, Cao Y, Yu P, Zhang L, Liu Z, et al. A Multivariate Diagnostic Model Based on Urinary EpCAM-CD9-Positive Extracellular Vesicles for Prostate Cancer Diagnosis. Frontiers in oncology. 2021;11.

[418] Lin G, Ye H, Wang J, Chen S, Chen X, Zhang C. Immune checkpoint human endogenous retrovirus-H long terminal repeat-associating protein 2 is upregulated and independently predicts unfavorable prognosis in bladder urothelial carcinoma. Nephron. 2019;141(4):256-64.

[419] Pramono AA, Rather GM, Herman H, Lestari K, Bertino JR. NAD-and NADPH-contributing enzymes as therapeutic targets in cancer: an overview. Biomolecules. 2020;10(3):358.

[420] Zhao Q, Zhang Z, Li J, Xu F, Zhang B, Liu M, et al. Lysine acetylome study of human hepatocellular carcinoma tissues for biomarkers and therapeutic targets discovery. Frontiers in genetics. 2020;11.

[421] Vazquez Rodriguez G, Abrahamsson A, Turkina MV, Dabrosin C. Lysine in Combination With Estradiol Promote Dissemination of Estrogen Receptor Positive Breast Cancer via Upregulation of U2AF1 and RPN2 Proteins. Frontiers in oncology. 2020;10:2650.

[422] Zhang R, Noordam L, Ou X, Ma B, Li Y, Das P, et al. The biological process of lysine-tRNA charging is therapeutically targetable in liver cancer. Liver International. 2021;41(1):206-19.

[423] Hargrove TY, Friggeri L, Wawrzak Z, Sivakumaran S, Yazlovitskaya EM, Hiebert SW, et al. Human sterol 14$\alpha$-demethylase as a target for anticancer chemotherapy: towards structure-aided drug design1. Journal of lipid research. 2016;57(8):1552-63.

[424] Gorbatenko A, Olesen CW, Boedtkjer E, Pedersen SF. Regulation and roles of bicarbonate transport in cancer. Frontiers in physiology. 2014;5:130.

[425] Robey IF, Martin NK. Bicarbonate and dichloroacetate: evaluating pH altering therapies in a mouse model for metastatic breast cancer. BMC cancer. 2011;11(1):1-10.

[426] Yang OC, Loh SH. Acidic stress triggers sodium-coupled bicarbonate transport and promotes survival in A375 human melanoma cells. Scientific reports. 2019;9(1):1-12.

[427] Chhetri DR. Myo-Inositol and its derivatives: Their emerging role in the treatment of human diseases. Frontiers in pharmacology. 2019;10:1172.

[428] Dearden L, Bouret SG, Ozanne SE. Sex and gender differences in developmental programming of metabolism. Molecular metabolism. 2018;15:8-19.

[429] Williams LA, Sample J, McLaughlin CC, Mueller BA, Chow EJ, Carozza SE, et al. Sex differences in associations between birth characteristics and childhood cancers: a five-state registry-linkage study. Cancer Causes & Control. 2021;32(11):1289-98.

[430] Magazzù G, Zampieri G, Angione C. Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods. Computers in Biology and Medicine. 2022;151:106244.

[431] Martínez VS, Saa PA, Jooste J, Tiwari K, Quek LE, Nielsen LK. The topology of genome-scale metabolic reconstructions unravels independent modules and high network flexibility. PLOS Computational Biology. 2022;18(6):e1010203.

[432] Alam MT, Medema MH, Takano E, Breitling R. Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism. FEBS letters. 2011;585(14):2389-94.

[433] Väremo L, Nookaew I, Nielsen J. Novel insights into obesity and diabetes through genome-scale metabolic modeling. Frontiers in physiology. 2013;4:92.

[434] Deo RC, Hunter L, Lewis GD, Pare G, Vasan RS, Chasman D, et al. Interpreting metabolomic profiles using unbiased pathway models. PLoS computational biology. 2010;6(2):e1000692.

[435] Abdik E, Çakır T. Systematic investigation of mouse models of Parkinson's disease by transcriptome mapping on a brain-specific genome-scale metabolic network. Molecular Omics. 2021;17(4):492-502.

[436] Beguerisse-Díaz M, Bosque G, Oyarzún D, Picó J, Barahona M. Flux-dependent graphs for metabolic networks. NPJ systems biology and applications. 2018;4(1):1-14.

[437] Berry T, Sauer T. Consistent manifold representation for topological data analysis. arXiv preprint arXiv:160602353. 2016.

[438] Qian Y, Expert P, Panzarasa P, Barahona M. Geometric graphs from data to aid classification tasks with graph convolutional networks. Patterns. 2021;2(4):100237.

[439] Qian Y, Expert P, Rieu T, Panzarasa P, Barahona M. Quantifying the alignment of graph and features in deep learning. IEEE Transactions on Neural Networks and Learning Systems. 2021;33(4):1663-72.

[440] Peach RL, Arnaudon A, Schmidt JA, Palasciano HA, Bernier NR, Jelfs KE, et al. HCGA: Highly comparative graph analysis for network phenotyping. Patterns. 2021;2(4):100227.

[441] Machicao J, Craighero F, Maspero D, Angaroni F, Damiani C, Graudenzi A, et al. On the Use of Topological Features of Metabolic Networks for the Classification of Cancer Samples. Current Genomics. 2021;22(2):88-97.

[442] Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko AB. A gentle introduction to graph neural networks. Distill. 2021;6(9):e33.

[443] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. 2020;1:57-81.

[444] Daigavane A, Ravindran B, Aggarwal G. Understanding Convolutions on Graphs. Distill. 2021;6(9):e32.

[445] Lee J, Lee I, Kang J. Self-attention graph pooling. In: International conference on machine learning. PMLR; 2019. p. 3734-43.

[446] Garg V, Jegelka S, Jaakkola T. Generalization and representational limits of graph neural networks. In: International Conference on Machine Learning. PMLR; 2020. p. 3419-30.

[447] Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. Bioinformatics. 2020;36(Supplement_1):i525-33.

[448] Spielman DA, Srivastava N. Graph sparsification by effective resistances. In: Proceedings of the fortieth annual ACM symposium on Theory of computing; 2008. p. 563-8.

[449] Peach RL, Arnaudon A, Barahona M. Semi-supervised classification on graphs using explicit diffusion dynamics. arXiv preprint arXiv:190911117. 2019.

[450] Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks. arXiv preprint arXiv:190512265. 2019.

[451] Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:181112808. 2018.

[452] Errica F, Podda M, Bacciu D, Micheli A. A fair comparison of graph neural networks for graph classification. arXiv preprint arXiv:191209893. 2019.

[453] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[454] Mi X, Zou B, Zou F, Hu J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nature communications. 2021;12(1):1-12.

[455] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal. 2021;19:3735-46.