Università degli Studi di Torino
Dipartimento di Oncologia

Doctoral School in Life and Health Sciences

PhD Thesis

Computational models as a tool to decipher cancer metabolic reprogramming

Candidato:
Niccoló Totis

Relatori:
Francesca Cordero
Marco Beccuti

Correlatore:
Gianfranco Balbo

Anno Accademico 2017–2018

CONTENTS

ABSTRACT

In any living organism, metabolic processes provide cells with cofactors that store energy reducing power, and molecules that are needed as precursors for the synthesis of all structural and functional components. At the intracellular level, metabolic processes are represented by an intricate network of thousands biochemical reactions. The huge number of reactions, the fast time scale of their activity and the high volatility of metabolites hugely complicate experimental procedures, thus rising their costs, and ultimately limiting the data availability. Owing to this complexity, the behavior of metabolic systems is inherently hard to study with traditional experimental and analytical tools, a situation that calls for new approaches. The importance of mathematical models in systems biology has grown in recent years, as their ability to understand, predict and manipulate biological systems has been successfully demonstrated in many cases. A wide range of useful modeling techniques has been already developed and published. The choice of the most appropriate among them should depend on the model goals, and on the experimental data and biological knowledge available for the phenomena under study.

The three-years work described in this thesis started with the review of the biggest challenges that modeling metabolism currently poses. Constraint-based models(CBMs) and kinetic models represent the two main classes of approaches used so far to model metabolism. While the former assumes a goal-directed nature of intracellular metabolism to predict reaction fluxes at the steady state, the latter exploits available mechanistic information and aims to describe the integrated dynamic response of the system to changes in the environment. Focusing more specifically on kinetic models composed of Ordinary Differential Equations (ODEs), we developed new computational methods that try to address their main limitations. As for many biological systems, metabolic responses are highly conditioned by many regulatory processes. Metabolic regulation can be schematically divided into two layers: the layer of fast adjustments in metabolite concentration and reaction fluxes, and the level slower modifications of enzyme expression. These regulatory processes are highly complex and have been poorly studied so far. Commonly, kinetic models of metabolism are built with bottom-up procedures, in which information on the structure and the kinetic parameters of each individual reaction is collected from different sources in the literature. While lots of detailed information can be found for central metabolic pathways like glycolysis, the same is not true for less studied reactions. Moreover, the available data is often produced by experiments performed on different cell types and in standard *in-vitro* conditions, where, for instance, effects of pH, temperature, unknown allosteric regulators, and kinetic differences in isoenzymes are always neglected. Hence, indetermination and uncertainty

constantly arise as major issues of the modeling phase. In order to clarify network interactions and kinetic parameters, approaches of reverse engineering and parameter estimation have been adopted in many situations. The performances of these techniques, however, strongly depend on the availability of high quality data, which often represents a limiting factor. In our work, we tried to move away from these approaches and to consider the problem from a different perspective. Acknowledging data scarcity, we chose to exploit the empirical knowledge of experimentalists to guide some modeling assumptions. After we define the system of metabolic reactions we are interested in, indeed, we choose to divide them into two categories: determined reactions, for which both the structure and the kinetic parameters are considered to be known and not affected by uncertainty, and undetermined reactions, for which we consider to have just partial information. Secondly, we assume an experimentalist can help us to define a specific objective function of the metabolic system. While this is conceptually similar to objective functions used in CBM, here we use it for optimization processes that are executed recurrently over the course of the simulation. These optimizations are meant to calculate, for each undetermined reaction, some time varying coefficients that replace the missing information and allow to reproduce the dynamic behavior of the system.

In a second work thread, we focused on baseline enzyme concentrations, and how their expression in different tissues impacts model behavior. Aiming to increase the specificity of our models for a tissue of interest, we developed a new method that exploits high-throughput biological data. In specific, a kinetic model which is fully parametrized for a reference condition, together with gene expression data from the same condition, can be used to generate a kinetic model specific for a new condition of interest, if gene expression data for this condition are available. These data are thus manipulated to adjust kinetic parameters related to enzyme concentrations and can effectively improve model performance for the new condition.

Besides intracellular metabolism, the main research interests of our group are related to cancer. One topic become important in recent years is tumor metabolic heterogeneity. Growing evidences support the idea that subpopulations of cancer cells with different metabolic traits coexist in the same tumor mass. This motivated us to move our attention to the study of metabolic events that happen at a mesoscopic scale and to consider inter-cellular metabolic interactions. With this aim, we worked on multi-scale models of cancer metabolism that link tumor growth to intracellular metabolic events. This represents also the main focus of our ongoing research efforts.

Across different sections of this thesis, we show the potential of computational approaches originally defined in the field of computer science to facilitate the modeling phase. Petri Nets are bipartite directed graphs, a high level graphical formalism which has been extensively used for various modeling purposes since its formalization. Here we show how their timed formula-

tion, Stochastic Petri Nets, and their colored variant, Stochastic Symmetric Nets, can ease modeling efforts, helping the modeler to represent metabolic reactions and characteristics of indetermination and heterogeneity present in the system.

1 INTRODUCTION

Metabolism can be considered a tightly regulated engine that provides essential cellular components such as energy equivalents, redox cofactors, biomass building blocks and precursors for chemical modifications of proteins or DNA. Metabolism is thus strongly connected with any other intracellular processes and has a great impact on cell growth.

The Warburg Effect

Scientific dissertations regarding cancer metabolism always start from the definition of the Warburg Effect [137]. In his pioneering work in the 1920s, the German scientist Otto Warburg showed that all of the cancer cells he investigated avidly ferment glucose, secrete lactate, and suppress oxidative phosphorylation, and that this behavior emerges even in the presence of oxygen. This contrasts with the Pasteur effect that normal cells display, which consists in the observation that O_2 is able to inhibit fermentation.
Warburg interpreted tumor lactate secretion as an indication that oxidative metabolism was damaged [263]. His assertion that "respiration of all cancer cells is damaged" has been long debated, since Warburg himself and his coworkers, as well as contemporary investigators, had experimental evidences that contradicted it. [87].
Today, we understand that the relative increase in glycolysis under aerobic conditions was mistakenly interpreted as evidence for damage to respiration. Several proofs that cancer cells have functional mitochondria have in fact been collected since then [38]. At the present moment, the topic of the relevance of fermentative versus oxidative pathways remains still open and debated, but it is widely believed that in most cancer cells cultures ATP synthesis is split roughly equally between glycolysis and oxidative phosphorylation, rather than the greater than 90% attributable to the latter in most cells. As the ATP yield from aerobic glycolysis amounts to 2 ATP per molecule of glucose, while it adds up to more than 30 ATP/glucose for oxidative phosphorylation, this implies that cancer cells have a high rate of glucose utilization. This feature is indeed daily exploited to localize tumors with PET scans, as the tracer 18F-deoxyglucose is strongly uptaken by the tumor [262, 39].

Similarities between cancer and highly proliferating normal cells

An additional consideration concerns the interesting similarity between metabolic traits in cancer cells and highly proliferating non-transformed cells. When these cells enter a high proliferative state, they tend to express glucose transporters and glycolytic enzymes out of proportion to the machinery required, displaying high fermentation of glucose with suppressed oxidative phosphorylation. As this phenomenon is directly linked to proliferation, some authors are prone to believe that the Warburg effect might reflect proliferation-associated changes in metabolism rather than a unique feature of malignancy [262].

The Crabtree effect

Another short-term reversible phenomenon that resembles the Warburg effect, the Crabtree effect , has been observed in cell cultures . In metabolically adapted cell lines that are grown in hypoxic/anaerobic conditions, when glucose concentration in the media surpasses a threshold value, the cells have a reduced need for oxidative phosphorylation by the TCA cycle and rapidly switch to glycolysis as their major source of energy. Some authors indeed proposed that this short-term reversible phenomenon might represent an advantage of cancer cells in vivo, as it would allow them to adapt their metabolism to the rather heterogeneous microenvironments in malignant solid overgrowths.

Differential substrate utilization

Also the idea that cancer cells rely only on glucose as a energy source has changed, in favor of a much more heterogeneous substrate utilization . Our view of repertoire of substrates that cancer cells are able to metabolize is still expanding. Importantly, these findings reinforce the thesis that cancer mitochondria are functional. After glucose, glutamine was the first carbon source to be recognized as a preferential nutrient source for cancer. In is in fact referred as a "glutamine addiction" to describe the observation that glutamine deprivation induces starvation and cell death *in-vitro*. Acetate and other fatty acids, lactate, branched chain amino acids, serine, and glycine represent additional fuels for many cancers [36, 263]. These complement glucose to sustain the core metabolic functions of cancer cells: energy formation, biomass assimilation, and redox control.

Despite this accumulation of significant experimental evidences, we need to be aware that metabolic processes are highly voluble, and depend on the specific context of observation. Hence,regarding the relevance of many of these experimental findings, it still has to be fully understood how many of those that were obtained *in-vitro* can be used to draw conclusions on cancer cells actual *in-vivo* behavior.

At present the research on cancer metabolism is taking several different perspectives to try to uncover the factors and mechanisms that are responsible for these metabolic alterations. A few of these research directions are here briefly described.

Metabolic control Multiple bidirectional feedback and control mechanisms between metabolism and cellular regulation guarantee cellular and physiological homeostasis. While in the past the classical perspective was to consider metabolic alterations the result of genetic mutations, now the community is much more aware that the influences between the genome, the transcriptome proteome and metabolome are bidirectional, with many regulatory loops connecting the different layers. It has been shown, for instance, that specific concentrations of metabolites like serine, arginine and leucine are able to influence some key intracellular regulators, namely mTOR, AMPK and p53 [118].
These mechanisms of auto-regulation, however, are so complex that cannot

Identification of enzymes whose activity is fundamental for the emerging metabolic alterations

be clarified by single experiments. For this reason, studies on the distribution of control over the network have so far mostly focused on the identification of enzymes whose altered activity produces changes in the whole system. In [263] Vander Heiden and DeBernardinis categorize metabolic activities based on whether they are transforming, enabling or neutral with respect to cell transformation and tumor progression.

- Transforming activities are those activities that are able to initiate carcinogenesis. Blocking them might prevent the occurrence of cancer in susceptible patients and slow disease progression. Only very few metabolic activities can be considered as transforming based on genetic evidences
- Enabling activities carry out conventional metabolic tasks such as supporting energetics, generating macromolecules, and maintaining redox state and are required for tumor progression
- Neutral activities are dispensable for tumor growth and are not relevant to be considered as therapeutic targets

In cancer the activity of an enzyme can be altered for several reasons: a mutation conferring new properties to the enzyme, a mutation altering its kinetics, or an alteration in the abundance of the enzyme functional form. In [66] the authors reported an accurate review of the known altered enzyme activities in of glucose metabolism in cancer cells. Here we report a few of these findings:

- Glucose transporters (GLUTs). GLUT1, a glucose transporter with an elevated affinity for glucose, is often found overexpressed in human cancers. hypoxia-inducible transcription factor HIF-1 and signalling molecules c-myc and Akt. By contrast, the GLUT4 transporter, which is sensitive to the presence of insulin, is normally down-regulated. Additional gates for glucose uptake are guaranteed by the Na+-coupled glucose transporter SGLT1.
- Hexokinases (HKs). Additionally to HK-1, tumor cells express HK-2, which is able to promote glycogen synthesis and to divert glucose 6-phosphate towards the oxidative Pentose Phosphate Pathway (PPP)
- Phosphofructokinases (PFKs). PFK is a tetrameric enzyme known to be conditioned by a high number of allosteric regulators. It is thus believed that anomalous concentrations of these regulators, like ATP and citrate, may be crucial to promote the Warburg phenotype.
- Phosphofructokinase-2/Fructose-Biphosphatases (PFK-2/FBPase).Four isoforms (PFKFB1 to PFKFB4) have been identified for this enzyme. In specific, the synthesis of PFKFB3 is induced by several factors known to be implicated with carcinogenesis and cancer progression, like HIF-1, cMyc, Ras. Indeed, it was found that rapidly proliferating cancer cells constitutively express PFKFB3.

- Phosphoglicerate Mutase (PGM). The three PGM isozymes found in mammalian cells result from the combination of two subunits: muscle M and brain B. PGM-M, in specific, is known to be upregulated in many cancers.

- Pyruvate Kinase (PK) has two isoforms, PK-M and PK-L. It has been seen that cancer cells, and fast-growing cells in general, selectively express the M2 isoform (PK-M2). PK-M2, thanks to its kinetic properties, is able to redirect the flux of glucose carbons to the Pentose Phosphate Pathways and other biosynthetic pathways.

- A few mutations in the Trycarboxylic Acids (TCA) Cycle have recognized tumorigenic effects. Mutations in both succinate dehydrogenase (SDH) and fumarate hydratase (FH) have in fact been shown to result in paragangliomas and pheochromocytomas. FH was found moreover mutated in renal cell cancers. The mechanism by which this occurs seems to be linked to a stabilization of the hypoxia-inducible factor HIF-1.

- Consistent evidences support the belief that several alterations in enzymes of the glycolytic and PPP pathways promote a redirection of the carbon flux towards ribose 5-phosphate synthesis in tumor cells.

Regarding the available information on tumor-specific metabolic enzymes, it is important to underline that the majority of studies focused on gene expression analyses, while detailed enzyme-kinetic studies and metabolic flux quantifications are still rare [66].

Spatial factors In [266] the authors focused their work on cytoplasm solvent capacity and proposed an explanation for the apparently counterintuitive preference of cancer cells for fermentative pathways with low ATP yield. Considering that enzyme molecules have a finite volume and the total sum of their volumes cannot exceed the cell volume, they speculated that glycolysis, which produces low yields, is preferred as it is more efficient in terms of the required solvent capacity.

Intratumoral metabolic heterogeneity The discussion about the use of preferential substrates and pathways (glycolytic versus oxidative) requires to introduce the topic of metabolic heterogeneity in cancer. According to results produced in many studies, cancer appears to be a variegate disease also for its metabolic aspects. For example, many cancer cells do not show a Warburg effect under all conditions, and slowly proliferating tumor cells rely more on oxidative phosphorylation than rapidly growing cells.
Many experimental evidences recently collected come out in favor of a vision of cancer metabolism that is more and more linked to the microenvironment [5]. Within the tumor microenvironment lots of different cell types coexist and interact, and these interactions are conditioned by continuously changing gradients of pressure, concentrations and pH. It has been

finally recognized that cancer microenvironment needs to be studied as a complex ecosystem, in which all phenomena, including the metabolic phenomena, represent emergent properties of the system that cannot be generated by any isolated component. In a recent exhaustive review on cancer metabolism [123], it was proposed that cross-feeding relationships exist among cancer cells and tumor stroma, as well as among different metabolic subtypes of cancer cells. Cancer cells which strongly rely on glycolysis as the major energy producing pathways display the well known Warburg phenotype, characterized by the high production of lactate. Lactate, which is the waste product of these highly glycolytic cells, seems to represent instead a valuable source of energy for other metabolic subtypes of cancer cells, that mostly rely on the oxidative phosphorylation pathway. According to the authors, these mutualistic interactions, observed also between cancer and stromal cells, can ultimately promote aggressive and treatment resistance phenotypes.

Thanks to these results and to the ongoing improvement of experimental procedures, the interest of the scientific community for phenomena related to cancer metabolism is witnessing a renaissance. The observations here synthetically reported make us aware of the fact that metabolic systems play an eminent role in any cellular process. On the other hand, metabolic systems, even just at the intracellular scale, include highly complexes networks of molecular interactions that so far have been characterized just in small proportion.

In situations where phenomena are highly relevant, in which we aim to make predictions, but for which we have limited experimental tools at our disposal, mathematical modeling becomes essential. Computational models of cancer metabolism have gained more and more attention in recent years, in parallel with the successes achieved in the larger field of bioinformatics and systems biology. Differently from traditional "wet lab" experiments, computational models are now increasingly used to integrate experimental findings and to give back verifications, predictions and suggestions to the experimentalists.

In this thesis we will present the work done during a 3-years PhD project. The next chapters are organized as follows: In chapter 2, we will give an overview of the general process that has to be followed in order to build computational models of metabolism, describing its main steps with higher detail. For each of these steps the major obstacles and objectives that modeling poses will be discussed and the most important computational method proposed in the literature will be briefly mentioned. Some of the most remarkable results that computational models achieved in the most recent years will be cited as well. In chapter 3 we will present the computational approaches that we developed. Chapter 4 will be used to illustrate the experimental results achieved applying these methods and to discuss their ac-

curacy and limitations. In chapter 5 we will summarize our work and we will provide some insights on the research directions that are focus of our current and future efforts. Some of the concepts and approaches that will be exposed in this thesis have already been presented in two publications produced by our group, namely "Dealing with indetermination in biochemical networks" [250] and "Overcoming the lack of kinetic information in biochemical reactions networks" [249].

2 BACKGROUND

1 MODELING

Even if a scientific model, like a car, has only a few years to run before it is discarded, it serves its purpose for getting from one place to another. David L. Wingate,"Complex Clocks", Digestive Diseases and Sciences, 1983, 28:1139.

In recent years, the computational methods that modelers use to interpret biological phenomena have been continuously updated with new techniques. Although the array of available tools is in a phase of rapid expansion, modeling biological systems, and more specifically metabolic systems, remains a hard challenge.

A successful modeling process should follow some fundamental steps. Starting from a review of the scientific evidences and the available data, the model goals and some modeling hypotheses should be defined. This first step helps researchers to clarify the uncertainty that affects the system under study and to accordingly choose the most appropriate modeling technique. Then, through the specification of its structure and parameters, a draft of the model is created. Iterative processes of validation and refinement, in which model predictions are tested and new experimental data are integrated, help scientists to produce the final version of the model. Some a posteriori analyses should finally be used to quantify the uncertainty associated with model components.

1.1 Preliminary steps

As already anticipated, before the model is actually built, a series of preliminary modeling steps need to be carried out. Modeling means representing an abstraction of reality in a simpler but still meaningful way. Modeling requires a balance between complexity and accuracy. If we are studying an actual biological system, we do not want our representation to be too abstract and superficial: we would miss the goals it was meant to achieve. On the other hand, for a higher precision, we may be tempted to describe the system with as many details as possible. However, "the more the details, the more the precision" does not work as a general rule. The precision, in fact, might remain illusionary. We need to be aware that every time we add details, we also introduce errors, and the complexity of our representation increases.

Following these considerations a careful balance should be applied when we choose how wide is the network of processes we are representing, and how grained is the representation of each of these.
These choices should be oriented by several factors: the goals of the user, the type and amount of experimental data available, the presence of previous similar efforts in the literature, the computational costs that can be afforded. Each of these preliminary analyses and decisions conditions the others and is tightly influenced by the others, so a specific order does not have to be necessarily followed.
Ultimately, these steps support the choice of the best modeling approach. If, on the contrary, we disregard the context of these factors, no modeling technique can be defined as superior or more adequate with respect to the others.

Model goals

A computational model uses a mathematical language to reproduce some experimental finding. Reproducing observations represents then the first, and necessary, step of any modeling effort. In order to prove its utility, a model should shed light on the system under study and offer a deeper understanding of its behavior. It should then help users to formulate hypotheses around the mechanisms that generated the observed phenomena, or to predict new phenomena that were not used in the process of model building.
If we refer to metabolic systems, a model can be used to understand, predict or manipulate their behavior.
So far, computational models of metabolism have been extensively used in the field of metabolic engineering, with the intent to predict which genetic modifications are able to create cellular strains with a phenotype of industrial interest. In parallel, they have also been interrogated to answer more fundamental questions. In the most recent years the scope of metabolic models has expanded towards biomedical questions and applications. New challenges have been posed, and new modeling approaches now seek to identify, for instance, which molecular target a drug should better address to impact metabolism [16] [54]. Related to cancer research, computational models can pinpoint which enzymes would deserve to be silenced in *in-vitro* cell lines as a way to affect cellular energy production.
A clear and rigorous definition of modeling goals is paramount. This decision should in fact influence the kind of output we want the model to produce, and thus the modeling approach we need to choose and the modeling process we need to follow.

Size of the system and level of abstraction

A schematic overview of the main modeling approaches, is given in Figures 1 and 2, including a list of their principal dichotomic features. We report here the distinction, proposed elsewhere [223, 23], between interaction-

based, constraint-based and mechanism-based models. Cybernetic models, here not included, will be discussed separately in section 1.1. In order to choose wisely the best approach for our specific scientific question we need to define both the size of the system and the level of abstraction we intend to use in the model.

Metabolic phenomena can be observed and studied at different levels, at the level of the organism, of an organ, of a tissue, at the level of a cell or of a group of reactions. More generally, metabolic networks can, on one side, include just intracellular reactions, while on the other side, they can be expanded to comprehend environmental metabolic processes, i.e. all those occurring outside the cells. So far, the greatest part of metabolic models has concentrated on intracellular metabolic networks. These are composed of thousands of reactions compartmentalized in different organelles. Limiting to intracellular models, the size of the network they describe can help us to divide them into genome-scale (GS) models, if they include all intracellular metabolic reactions, core models, if they consider some of the main metabolic pathways, and toy models, in case they highlight some major feature of the system transcending the effects of single reactions.

Size of the network

The level of abstraction able to formally describe the functioning of the system should be identified taking into account the known biochemical, physical or regulatory properties of system components. This level of abstraction can be considered either fine-grained, in the case of mechanism-based models, or coarse-grained, as for interaction-based or constraint-based models.
The size of the system and the level of abstraction are closely linked: mechanism-based or cybernetic approaches are normally applied for toy or core models, while the interaction-based and constraint-based approaches are more suited for the analysis of genome-scale or core models.
A more extensive description of these different categories of approaches is reported in paragraph 1.1.

Level of abstraction

Moving from the coarse-grained (interaction-based, constraint-based) to the fine-grained (cybernetic, mechanism-based) approaches, models vary with respect to:

- the maximum tractable size of the system
- the computational costs that the analysis of the model requires
- the type of description of the system provided, either qualitative to quantitative
- the type of data they encompass and produce: static or dynamic

While mechanism-based models are in principle the most informative and what any modeler would aim for, in practice, due to their limitations, constraint-based models represent the standard technique of choice.

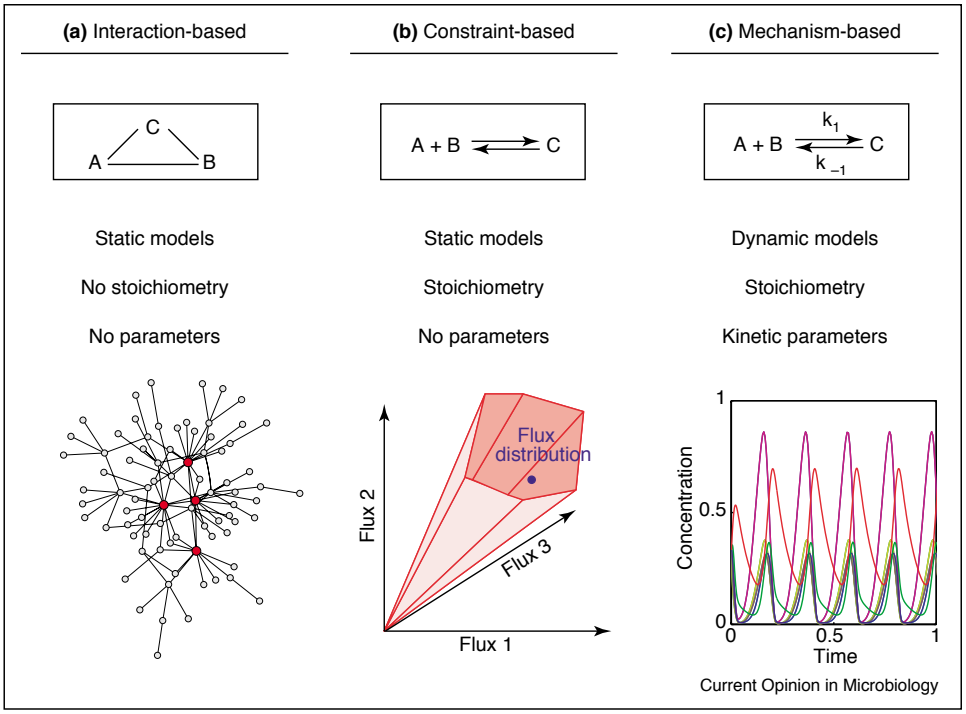

Figure 1: Schematic view of the computational approaches that can be used to model metabolism. Figure taken from [223].

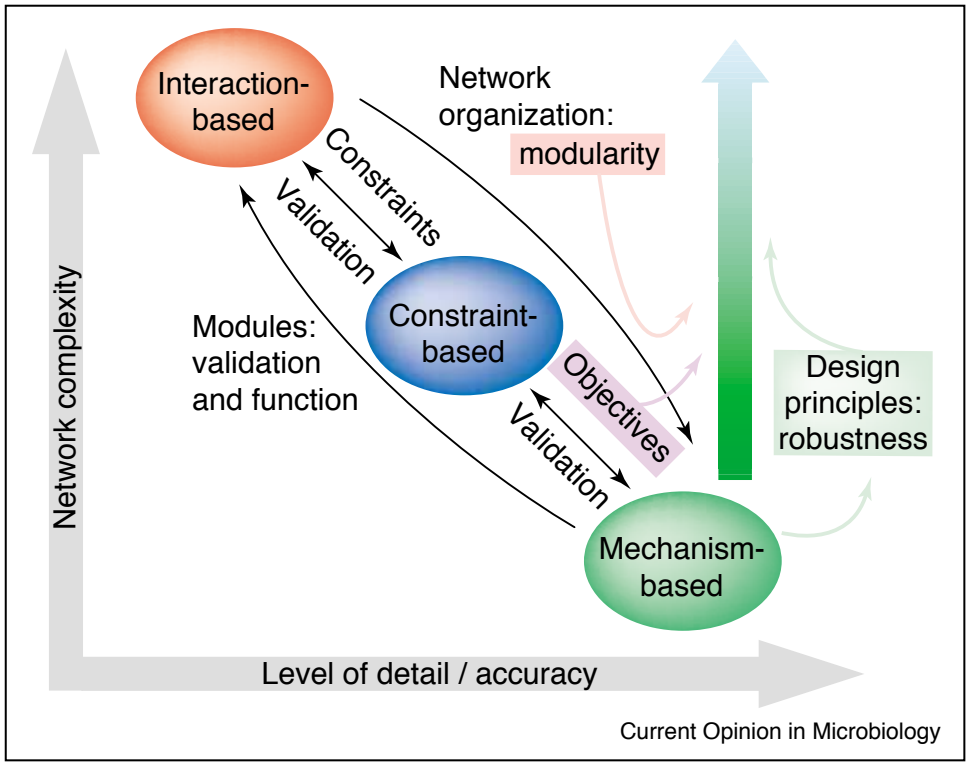

Figure 2: Mathematical modeling: scope and interactions. Figure taken from [223].

Modeling hypotheses

No model can be representative of any condition. A model can be either very generic, as it adapts to many different experimental conditions, either very specific.
Choosing between generality and specificity means to choose between models that are versatile in their use and models that are precise in their prediction. Caution should assist the modeler to find the right balance between extremes. The more assumptions are made, the more the model becomes specific. By marking the area of validity of the model, assumptions highlight model performances within those limits. However too many assumptions would restrict the scope of the model, to an extent that the model would then fail to represent experimental conditions of interest.
A tight discussion between modelers and experimentalists is arguable and helpful to clarify model assumptions and promote the phase of model construction

Data availability

EXPERIMENTAL TECHNIQUES Experimental data are necessarily the element that make models realistic and meaningful. A model can represent any type of processes, but again, its main goal is to reproduce some experimental findings. Data thus act as a constraint that guide the modeler to select the most appropriate modeling technique, as well as its specific structure and parameters. While for the great part of the last century the majority of data on metabolic systems was produced by biochemical assays, in the last decade all omics technologies have dominated the processes of data generation. As it is reasonable to think that these technologies will be pervading in the near future, we will limit our descriptions to these

- **Genomics**. Explosive advances in next-generation sequencing technologies (NGS) and computational analyses have enabled exploration of somatic protein-altered mutations in most cancer types. However, there is limited information on somatic mutations in non-coding regions, including introns, regulatory elements and non-coding RNA. Whole genome sequencing (WGS) approaches can be used to comprehensively explore all types of genomic alterations in cancer and help us to better understand the whole landscape of driver mutations and mutational signatures in cancer genomes and elucidate the functional or clinical implications of these genomic alterations. Whole exome sequencing (WES) is nowadays the main platform for cancer genome sequencing and vast amounts of mutational data in protein-coding regions have been accumulated for all types of common and rare human tumors.

- **Transcriptomics** is the large-scale study of RNA molecules by use of high-throughput techniques. It examines the abundance and makeup

of a cell's transcriptome. In contrast to DNA, which is largely identical across all cells of an organism, the actively transcribed RNA is highly dynamic, reflecting the diversity of cell types, cellular states and regulatory mechanisms. Because a transcriptome profile can be regarded as a signature or snapshot of the underlying cell state, the experimental profiling of samples and specimens can provide insights into their unique biology [27]. Classical methods such as Northern blotting and RT-PCR allowed the steady state measurement of selected transcripts. Microarrays and Affymetrix chips followed, but are now progressively being replaced by deep sequencing technologies. Next-generation sequencing methods provide with the mapping and quantification of several thousands of transcripts in single experiments[121].

In particular, transcriptome-wide gene expression profiling has proved useful to better understand the molecular mechanisms underlying prognosis and drug sensitivity, also for the study of cancer. Cancer cells are characterized by altered protein function and aberrant transcriptional patterns, which are the consequence of somatic mutations and epigenetic alterations. More recently, RNA editing, post-transcriptional modifications and various non-coding RNAs have represented essential aspects of transcriptomics. Of particular relevance to cancer, the base-pair resolution and coverage of modern techniques enabled the detection of expressed somatic mutations, including single nucleotide variants (SNVs) and gene fusions. As we will outline more specifically in section 1.5, several methods to integrated gene expression data in metabolic models have been developed.

- **Proteomics**. A qualitative proteomic analysis is focused on the study of proteins present in various types of biological materials, in particular to identify their functions, structures, and interaction sites or post-translational modifications. On the other side, comprehensive quantitative descriptions of biological systems at protein levels are more recent. In fact, the fast-evolving Mass Spectrometry (MS) techniques, the identification, and quantification of all of the proteins in a biological system are still an experimental challenge. Proteomics allows ones to unravel disease-related molecular mechanisms and to identify new disease biomarkers to be used in clinical applications for diagnosis, for evaluation of therapy outcomes and for follow-up analyses. Clinical proteomics, in fact, enables the quantitative and qualitative profiling of proteins and peptides that are present in clinical specimens like body fluids, cells, and tissues. Concerning proteins of metabolic interest models, the availability of public dataset reporting concentrations of metabolic enzymes is still limited [71].

- **Metabolomics**, the youngest of the omics technologies, is able to concurrently identify thousands of metabolites. Considering that the different amounts of metabolites obtained under perturbed experimental conditions reflect the changes in enzyme activity, metabolomics pro-

vides a biochemical snapshot of the physiological and pathological state of a cell or an organism. Metabolic profiling provides a complete functional picture a particular phenotype, derived from the integration of information stored in the genome and decoded via transcription and translation, together with information coming from the environment. Thanks to this comprehensive representation of cellular phenotypes metabolomics has recently found a valuable use in the clinical field to identify new biomarkers in neurological, cardiovascular and oncological diseases [206, 253, 269, 23]. Caution should be however used when we interpret metabolomics data. Studying and measuring metabolic processes occurring inside the human body is in fact an inherently hard challenge. Metabolites, due to their small molecular weight, are highly volatile compounds that can be easily and quickly altered. Metabolic reactions, on their side, occur at a very fast time scale, so their activity can change quickly with changes in the surrounding environment. Despite its limitations, metabolomics holds the promise to be the technique most widely employed for the overcoming years to study metabolic phenomena *in vitro* and *in vivo*.

- Fluxomics. Fluxomics is the determination of the actual reaction rates within metabolic networks. Extracellular fluxes between cells and their environment can be easily derived from time-dependent changes of extracellular metabolite concentration. By contrast, intracellular fluxes are not directly measurable but can be inferred from 13C-isotope tracer experiments. When cells are grown on a 13C-enriched substrate, 13C atoms propagate through the metabolic network according to the metabolic pathways and their activity. The 13C labeling patterns of metabolic intermediates and cellular components thus depend on the intracellular fluxes. Significant progress in the development of high-throughput fluxomics made in the last few years has contributed to deepening our understanding of cellular metabolism.

PUBLIC DATABASES In the process of model building, both the structure and the parameters need to be defined. Besides the experimental data, different types on biochemical information, both qualitative and quantitative, needs to be assembled and integrated in the model. To store and access this scientific information, a large number of open on-line databases have been created during the last years. We report here a synthesis of the fully detailed overview proposed in [29] and illustrated in figure 3:

- **Pathways databases**. Currently, there exist a vast number of databases containing information on biochemical reactions. MetaCyc [21] contains descriptions on large metabolic pathways and regulatory information. The KEGG database [76] includes reaction records that are linked to metabolic enzymes, genes and also to functional categories like pathways. The BIGG database [199] was instead constructed to store curated genome-scale metabolic reconstructions with a standard nomen-

clature, exportable in Systems Biology Markup Language (SBML), thus facilitating the comparison between different organisms.

- **Experimental data repositories**. The availability of experimental data is a fundamental requirement to build kinetic models. For this purpose, curated databases with metabolomics and/or fluxomics are essential. The experimental metabolite concentrations, flux data and enzyme levels are comprehensively collected from the literature in KiMoSys [30]. ArrayExpress [154] incorporates genomics data from high-throughput functional genomics experiments and PRIDE archive [181] includes proteomics data, including protein and peptide identifications [29].

- **Kinetic information databases**. The SABIO-RK [284] database includes kinetic parameters as well as associated mechanistic rate laws for a large collection of reactions and an automated service which offers the possibility to export this information with annotations in SBML format. The BRENDA [202] database is the main collection of enzyme functional data available to the scientific community. All these databases are linked to external sources such as UniProt [258], NCBI taxonomy and PubMed ID [198] to provide further semantic annotations. When we make use of these parameter values, we should always be aware that these depend on the conditions in which they were measured, and that these conditions are not always reported in the databases.

- **Model repositories**. Curated and reusable models describing biological systems can be found in publicly available on-line repositories such as BioModels database [106], JWS online model database [146] and the Physiome Model Repository (PMR2) [297].
  The BioModels database is probably the most well-known web source to store published curated models of biological systems. It provides a large variety of standard formats, including the SBML, and model entities are annotated with cross references to external databases such as ChEBI [37], KEGG [76] and UniProt [258].

Choice of modeling technique

Once the goal is set, the system is identified, experimental hypotheses are formulated, data availability is assessed, and the level of abstraction is chosen, we have collected enough information to select the computational technique that best suits our needs. As we anticipated, we expanded the classification proposed in [23], and we choose to organize the multitude of computational models that can be built to describe metabolic systems into four main categories that differ for the level of abstraction used: Interaction-based models, constraint-based models, cybernetic models and mechanism-based models. Each of them will be presented in the next paragraphs.

QUALITATIVE MODELS

Interaction-based models Interaction-based models are built with processes of network reconstruction and contain only structural information. They are qualitative maps of the system, that disregard quantitative details of the stoichiometry of reactions. Interaction-based models are normally exploited to study topological properties of the network with methods of network analysis.
In a given network, several topological features can be investigated: the degree distribution (statistical indexes on the number of arcs connecting nodes), centrality measures (indexes that indicate the relative importance of nodes and arcs), and the presence of hubs (highly connected components), motifs (repeated architectures), and clusters (portions of the network with a high node density) [4, 23]. These analyses unraveled that in metabolic networks most of the nodes have few edges and only a few nodes are hubs. This feature can be described mathematically as a degree distribution of nodes that follows a power-law. Networks of this kind, defined scale-free networks, prove to be robust if nodes or edges are randomly removed, but very fragile if hubs are disconnected. Also, metabolic networks show a small-world character, that is, any two metabolites in the network can be connected by paths that follow a relatively short number of reactions [45, 23].

QUANTITATIVE MODELS All quantitative modeling approaches have in common a process of network reconstruction. Reconstructing a metabolic network means to list all the reactions present in the system and to specify the reactants and the products for each of them. This structural information can be stored in a matrix, namely the **stoichiometric matrix**. The number of rows in the stoichiometric matrix corresponds to the number of metabolites present in the system while the number of columns equals the number of reactions in the system. The non-zero elements of each row represent the stoichiometric coefficients of the related metabolite in the reaction indicated by the column. If the metabolite is substrate of the reaction, the stoichiometric coefficient would have a negative sign; otherwise, if it is a product, the stoichiometric coefficient would have a positive sign. The stoichiometric matrix is generally a sparse matrix as the number of metabolites involved in each reaction tends to be much smaller than the number of reactions in the system.

The stoichiometric matrix

Network reconstructions are typically created in a bottom-up fashion based on genomic and bibliomic data. As already mentioned, network reconstructions can vary in size, from genome-scale networks to smaller core and toy models. The generation of networks derived from top-down approaches (inference of component interactions based on high-throughput data) will be instead discussed in section 1.2. Because of their extension and complexity, genome-scale (GS) metabolic network reconstructions result many years of collaborative work among different groups. As extensively described in [240], in order to speed up this procedure, automated strategies are typ-

Network reconstruction

ically exploited to aid the creation of an initial draft network starting from the genome sequence or annotations of an organism. In most of the cases additional process of manual curation and refinement are needed. This phase can also be supported by semi-automated procedures, that for instance can be used to identify missing reactions in the network [23].

To date, the most successful efforts to reconstruct a GS model of human metabolism is represented by Recon 1 and its successive extensions (Recon 2 [241], 2.2 [233] and 3D [18]). The initial reconstruction exploited an accurate human genome sequence annotation and from these genes, by taking into account the protein-gene relationships, it was possible to identify the metabolic enzymes and hence the reactions that they catalyze. These reactions were carefully formulated considering the stoichiometry of reactants and products, the substrate specificity, their directionality and reversibility and account for the conservation of mass and charge-based metabolite ionization. Moreover, metabolites in Recon 1 were correctly compartmentalized to properly consider transport and exchange reactions in the model. The entire reconstruction process consisted of various rounds of refinement and validation, supported by the review of online databases, primary articles, and textbooks.

After the network of the system is identified, a quantitative models of metabolism can be built to represent some characteristics of the system, like biochemical concentrations and reaction velocities. Due to their complexity and variability, biological processes are inherently difficult to model quantitatively. This stochastic behavior, indeed, can be only be represented by means of stochastic models. A few fundamental concepts of stochastic simulation are outlined in section 3.
Specific characteristics of metabolic systems allow modelers to exploit the so called fluid approximation, further described in 3 and to build a deterministic description of the system, based on Ordinary Differential Equations (ODEs), that well approximates the behavior of a stochastic model.
A quantitative description of a metabolic system is thus usually given as

$$\frac{dm}{dt} = S \cdot v \quad m(0) = m_0$$

that evidences how metabolite concentrations, m, and reaction velocities, v, are the object our focus. S, the stoichiometric matrix, defines how the changes in metabolite concentrations, $\frac{dm}{dt}$, depend on the values of v.

Constraint-based models Both constraint-based models and cybernetic models, considerably different in their approach, share a common perspective of intracellular metabolism that requires to introduce the term teleonomy, defined for the first time by the evolutionary biologist Ernst Mayr to describe the *apparently end-directed* behavior of biological systems. In contrast to the philosophical concept of teleology, teleonomy explains these goal-directed behaviors as the expression of genetic programs. A teleonomic view of nature hence considers that genomes of all living organisms store

Teleonomy

information acquired during a long history of natural selection, and thus that only genomes encoding phenotypes with optimal performances have been selected [292]. Alternatively rephrased, teleonomy considers that the direction of evolutionary pressures define the functioning of intracellular metabolic systems and, even though cells do not concretely maximize a specific metabolic task, starting from this assumption offers a description of intracellular metabolism that well approximates its real behavior.

Flux Balance Analysis (FBA)

Flux Balance Analysis (FBA) was the first method formalized in these group. The goal of flux balance analysis is to compute the distribution of reaction fluxes in a metabolic system at the equilibrium. The equilibrium, or steady state, is a situation that follows each perturbation, and in which metabolite pools remain constant, with no more net production or consumption over time. The transient period occurring after each perturbation of the system is, instead, neglected. Mathematically, the intracellular steady state is expressed as:

$$\frac{dm}{dt} = S \cdot v = 0 \tag{1}$$

where m is the vector of intracellular metabolites, and v is the vector of reaction fluxes we seek to identify.

Creating a FBA model consists in defining an optimization problem and solving it with linear programming algorithms. Thus, FBA requires a process of network reconstruction, with the creation of the stoichiometric matrix, the definition of an objective function, and of some equality and inequality constraints that define the space of feasible solutions explored in the optimization process.

FBA output, a vector of reaction fluxes at the equilibrium, corresponds to the left null space of the matrix S. This null space, in geometrical terms, is a convex polyhedral cone which in the context of metabolic network analysis is called the flux cone [85]. Equation 1.1 defines a system at the steady state, as it consists of a list of equality constraints that express the fact that the net production or consumption of each metabolite has to be zero.
Since metabolic networks typically include more reactions than metabolites, these equality constraints constraints alone leave the system under-determined. The space of all feasible solutions is further reduced adding inequality constraints that specify the minimum and maximum value of each reaction flux. These bounding values are defined according to some experimental evidences reported in the literature.
Flux balance analysis then retrieves the distribution of fluxes that maximizes, or minimizes, the value of an objective function J. This is defined on the

The objective function

user's choice and classically takes the form of a linear weighted sum of reaction fluxes like:

$$J = \sum_{k=1}^{N} c_k v_k = c^T v,$$

where c is a vector of coefficients that weight the contribution of each reaction to the objective function.
Supported by its large applicability in the field of Metabolic engineering, the maximization of biomass is the objective function most frequently used. Maximization of ATP or minimization of intracellular fluxes have been also successfully used to reproduce metabolic phenotypes [176, 23]. In all these cases the objective function expresses what is believed to be a physiological behavior of the cell.
Alternatively, the objective function can be defined as the maximization of the production of a specific metabolite. In this case flux balance analysis does not aim to describe the real behavior of the system but it is used to explore the metabolic capabilities of the cell. In a classical FBA formulation the objective function and the constraints are linear with this respect to the variables. When this is the case, the optimization problem can be solved with linear programming algorithms. FBA variants with non-linear objective functions or constraints have been proposed, and are reported in [68, 23].
In the solution of a linear programming problem, a unique maximum value of the objective function often does not coincide with unique solutions of the optimization problem. When this is the case, multiple flux distributions show equally optimal values of the objective function. Several techniques, that further explore the solution space, have addressed the issue of indistinguishability between optimal solutions, like the decomposition of the flux distribution into Elementary Flux Modes (EFM), or Flux Variability Analysis (FVA).

Elementary Flux Modes

Elementary Flux Modes EFMs are minimal pathways that connect the inputs with the outputs of the model. In mathematical terms, minimal pathways are support-minimal vectors of the flux cone. EFM are then formally defined as the nonzero, support-minimal vectors of the flux cone. Importantly, every steady-state flux distribution of the system can be represented as linear combination of EFMs [300, 85]. Thanks to their properties, EFMs have become useful in the analysis of medium-scale metabolic networks for the following applications, among others [85]:

- to identify minimal conversion routes in metabolic networks
- to predict enzyme and reaction essentialities
- to characterize equally optimal steady state flux distributions
- to investigate metabolic trade-offs

Flux Variability Analysys

Flux Variability Analysys (FVA) is another method used to identify equally optimal solutions. In this case the algorithm disregards the entire flux space,

and instead focuses on one reaction at time to identify the range of flux values that does not alter objective function optimality.

EFMs analysis, as well as FVA, are however, computationally intensive tasks that are normally difficult to handle for genome-scale models. Algorithms that speed-up these computations have been however proposed.

The reliability of a FBA result eventually depends on the goodness of the defined constraints and objective function. In fact, both flux constraints and the coefficients of the objective function are parameters that are eligible to be tuned. It is still debated how the most appropriate objective function should be defined. The choice of biomass maximization is normally motivated by theoretical speculations around the evolutionary pressures the cells historically had to face. However, simply from a theoretical standpoint, many counter arguments could be proposed. In the case of microorganisms in natural environments, for instance, processes of environmental stress, exploration, adaptation and selection always involve communities and not single microorganisms. If we accept the concept of co-evolution, then it appears clear how evolutionary pressures acting on the entire community would select for the fastest growing community rather than for the microorganism that grows fastest in isolation.

Choice of the most appropriate objective function

In the case of multicellular organisms, like humans, the situation even over complicates, as evolutionary pressures here act on the entire organism and not at the level of single cells. While deep and fascinating, theoretical considerations should in practice, and as more appropriate, leave space to more empirical ones. If the modeler acknowledges this uncertainty about the definition of the objective function, some educated guesses may help him to identify the specific objective function that results in the metabolic phenotype that is closest to experimental data, like in a process of parameter estimation.

FBA is the progenitor of many other techniques. Among these, dynamic Flux Balance Analysis (dFBA) and its extensions provide a dynamic description of extracellular metabolites concentrations, while the intracellular network is considered at the steady state.

Dynamic Flux Balance Analysis

In the literature, two slightly different implementations of dFBA can be found. In a first one, proposed by Varma and Palsson [265], the boundaries of uptake reactions are dynamically adjusted as metabolites become depleted in the extracellular environment, but no kinetic expression for uptake fluxes are defined.

In a second formulation [120], a Dynamic Optimization Approach (DOA) and Static Optimization Approach (SOA) were proposed.

DOA uses an optimization process over the entire time period of interest to obtain time profiles of fluxes and metabolite levels. The dynamic optimization problem is transformed to a non-linear programming (NLP) problem and the NLP problem is solved once.

In contrast SOA requires that the total simulation time is divided into sev-

eral time intervals and instantaneous optimization problems are solved at the beginning of each time interval, followed by ODEs integration over the interval. In SOA, the number of variables that have to be solved is far fewer in comparison, and the optimization problem is an LP problem as opposed to the NLP for DOA. The SOA version has thus received a much larger appreciation from the scientific community. Many works like [67] uptake fluxes are computed with mechanistic rate laws, starting from extracellular concentrations. These values are used to dynamically adjust FBA constraints at each simulation step.
Many other derivations and extension of FBA have been proposed recently. For an overview of all FBA-related approaches, here not addressed, the reader might refer to [64].

Cybernetic models The mathematician Norbert Wiener first coined the term cybernetics to describe a scientific field of "control and communication theory, whether in the machine or in the animal". He envisioned that the fields of engineering systems and biological systems show interesting commonalities with respect to regulatory processes. Also, he felt that knowledges and approaches developed in one field could be transferred and adapted to the other.
Building onto the definitions of cybernetics and teleonomy, in mid 80s Ramkrishna and coworkers first defined cybernetic modeling as a way to describe in mathematical terms the teleonomic principles of biologial systems. Similarly to FBA in fact, cybernetic modeling assumes that metabolic systems have been engineered by nature, through evolutionary forces acting for millennia, and thus show optimal performances. With a long series of articulated refinements, cybernetic models indeed proved how the approach of optimal control theory, traditionally applied in engineering applications, can help decipher intracellular metabolic phenomena [292].
In cybernetic models the state of the system is defined as a vector of concentrations of biochemimcal species, in specific metabolites **y**, subdivided into extracellular s and intracellular m, enzymes **e** and the biomass component **c**.

$$x = \begin{bmatrix} y \\ e \\ c \end{bmatrix} \quad , \quad y = \begin{bmatrix} s \\ m \end{bmatrix}$$

The dynamic behavior of the system is reproduced according to the following set of equations:

$$\dot{s} = S_s \operatorname{diag}(v) r c$$
$$\dot{m} = S_m \operatorname{diag}(v) r - \mu m$$
$$\dot{e} = \alpha^* + \operatorname{diag}(u) r_E - \operatorname{diag}(\beta) e - \mu e$$
$$\dot{c} = \mu c$$

where S_s and S_m represent the rows of the stoichiometric matrix S corresponding to s and m, respectively, and the "diag" operator forms a diagonal

matrix from its vector argument [293]. The regulatory strategies that the cell implements, like a rational controller maximizing its goals, are here described by u and v, the vectors of the so-called cybernetic control variables. The u vector mimics transcriptional and translational control of the rates of enzyme synthesis. The v vector, instead, mimics the control exerted by metabolite concentrations on enzyme activities, like the feedback inhibition of the product of a pathway on an upstream enzyme that contributes to its synthesis. The enzyme balance equation includes a constitutive synthesis term α^* that is necessary to maintain a basal level of each enzyme. A first-order degradation term is also included in the enzyme balance, with rate constants specified by the vector b. The specific growth rate μ is computed by summing the net specific production rates of all intracellular components,

$$\mu = h^T S_m \, \mathrm{diag}(v) r$$

where the conversion factors contained in h are required to express each metabolite concentration on a weight fraction basis. Unless otherwise noted, the kinetic expressions used to evaluate the elements of r and r_E are given by

$$r_j = k_j e_j \prod_{i \in I^-(j)} \frac{y_i}{K_{ij} + y_i}$$

$$r_{E_j} = \alpha_j b \prod_{i \in I^-(j)} \frac{y_i}{K_{ij} + y_i}$$

where $I^-(j)$ is the set of metabolite indices associated with the substrates of the jth reaction, $I^-(j) = i : S_{ij} < 0$. The multiplier b that factors into the expression for r_{E_j} represents the fraction of biomass ascribed to the enzyme synthesis machinery. Each model to be discussed includes balances on a B pseudocomponent, which encapsulates all of the DNA, RNA, protein, lipid and other core biomass constituents not explicitly considered in the remainder of the biochemical network. The factor b is then equivalent to the specific concentration of B [293].

In order to solve the system equations just defined, the vectors of cybernetic variables u and v need to be calculated. This is done with the Matching and Proportional Laws, whose definition [294] rely upon optimal control heuristics.

$$\sum_i u_i = 1$$

$$u_i = \frac{p_i}{\sum_{i=1}^{N_S} p_i}$$

$$v_i = \frac{p_i}{\max_{i \in \{1,2,\ldots,N_s\}} p_i}$$

These laws state that the implementation of cellular metabolic strategies should follow the economic principle of "return on investment". This can be explained intuitively as a policy by which the more a reaction contributes to cellular growth, the more it is activated by the cell.

Mechanism-based models Mechanism based the models have the potential to reproduce the dynamics of the system. After all metabolites enzymes and reactions in the system are listed, a dynamic description of the system

$$\frac{dx}{dt} = S \cdot v(x, k) \quad , \quad x(0) = x_0$$

can be produced starting from the following model elements: a scheme of the mechanism of molecular interactions among biochemical components of the system, a mathematical representation of these interactions, some specific values for the kinetic parameters appearing in this representation, and a set of values for the biochemical concentrations at the beginning of the simulation. All of these model elements have to be defined with caution, starting from a review of the information available in literature.

Interaction network The scheme of interactions among biochemical components of the system is part of model structure, which thus, for mechanism-based models, is not limited to the stoichiometry of the system. As a complete knowledge regarding the interactions between an enzyme and the full list of substrates, products, activators and inhibitors is rarely available, the structure of mechanism based models is often affected by indetermination. The assessment of model indetermination will be discussed in section 1.2

Reaction rate laws The values of reaction fluxes are computed with specific mathematical expressions. These are defined in accordance to some reaction rate laws.
Reactions rate laws are classified in different categories, depending on the level of accuracy of their description. An extensive and detailed summary of the meaning and use of the most important rate laws is reported in [192]. Screening the literature, however, it can be noticed that authors still do not fully agree on a unique classification of rate laws. The definition of mechanism-based, fully parameterized, canonical, approximate rate laws are all terms that in different papers take slightly different meanings.
In the absence of a terminology consistently and uniformly used by the community, in this thesis we will chose a less theoretical but rather practical scheme of classification. The term "canonical", for instance, which is used to indicate that the definition of the rate law is purely based on the metabolic interaction network and not on further assumptions, is here abandoned. We are aware that this may appear a gross oversimplification of formal classifications, but we decided to give our own re-interpretation of this classification to make it more affine to the concepts we will present in chapter 3

First of all, we consider that reaction rate laws are essentially mathematical models that describe how the reaction velocity depends on the concentration of the main biochemical entities involved. Thus, as for any other model, all reaction rate laws are representations of a real behavior and, as such, include

some level of approximation. From our perspective, it is more useful to distinguish two levels or grades of this approximation, depending on the process that generated it.

- In a first group, to which we will alternatively refer as **fully detailed**, **fully parameterized** or **mechanistic-based** rate laws, we include all the kinetic representations which depend on a enzyme-specific reaction mechanism and are consistent with thermodynamic laws. Typically these rate laws have been proposed and refined during decades of research in biochemistry. In most of the cases, they require the values of many kinetic parameters to be defined.
- In a second group, to which we will refer as **simplified** or **approximate** rate laws, we include those expressions that, exploiting some assumptions and disregarding intentionally some level of detail, simplify a fully detailed rate law.

Law of mass action

Fully parameterized rate laws

The simplest and most widely used reaction rate law is the law of mass action (LMA). LMA has been traditionally used to model the velocity of uni- or bi-molecular reactions. These reactions are called elementary reactions, due to the fact that they each represent one single molecular interaction and so they cannot be further decomposed into intermediate steps. The simplest process by which an enzyme transforms a substrate into a product can be seen as a sequence three elementary reactions: the binding between an enzyme (E) and a substrate S into the enzyme-substrate complex (ES), the backward dissociation of ES into E and S, and the dissociation of ES into free E and the product P [192].

$$E + S \underset{k_{1}-}{\overset{k_{1}+}{\rightleftharpoons}} ES \overset{k_{2}}{\rightarrow} E + P \tag{2}$$

The LMA representation of the elementary steps of synthesis and dissociation of the enzyme-substrate complex is the following:

$$v_1 = k_1^+ E \cdot S - k_1^- ES \tag{3}$$

The net flux through reaction v_1 equals the difference between the two unidirectional fluxes. Intuitively the law of mass action provides a relation between reaction rates and molecular concentrations in a constant volume assuming that molecules are equally distributed in the space.
It is clear that if all enzyme-catalyzed reactions in the system were decomposed into their elementary steps, the dimension and the complexity of the model would explode. For this reason, it is normal practice in biochemistry that all the elementary reactions participating in a single enzymatic reaction are modeled with a unique mathematical expression. The first general rate equation that describes enzyme-catalyzed reactions was derived in 1903 by Henri, and ten years later slightly modified by Michaelis and Menten,

who confirmed Henri's experimental work. Henri-Michaelis-Menten (HMM) equation is defined as:

Henri-Michaelis-Menten rate law

$$v = \frac{k_2 E \cdot S}{K_m + S} \quad , \quad v_{max} = k_2 E$$

Where E is the free enzyme and v_{max} is the maximum velocity that would be observed when all the enzyme is present as ES. In order to derive this equation, some specific assumptions had to be made. They limit the validity of the use of HMM equation to the cases in which:

1. the enzyme acts as a catalyst
2. the enzyme and substrate react rapidly and form an enzyme substrate complex
3. only a single S and a single ES complex are involved and the ES complex breaks down directly into the free enzyme and P
4. E,S and ES complex are at the equilibrium; that is, the rate at which ES dissociates into E + S is much faster than the rate at which ES breaks down to form E+P
5. The concentration of S is much larger than the concentration of E so that the formation of ES does not alter significantly the concentration of S
6. the overall rate of the reaction is limited by the breakdown of ES to form E and P, described as the catalytic step
7. the velocity is measured during the very early stages of the reaction so that the reverse reaction is insignificant

The assumption that only the early components of the reaction (E,S and ES) are at equilibrium is called a **quasi equilibrium or rapid equilibrium assumption**.
HMM equation can be derived from the individual elementary steps modeled with LMA. Thus, all the parameters that appear in HMM equation result from the combination of multiple elementary reaction rates. Much of the success earned by HMM equation is due to the fact that these aggregated, or macroscopic, parameters can be easily retrieved by fitting the results of enzymatic assays in standard conditions.

Since then, HMM equation has become the rate law most widely used to compute reaction velocity starting from the concentrations of substrates and products. We have to be careful that every time we use HMM equation we are implicitly accepting all Henri's assumptions. In particular, while in biochemical practice the quasi-equilibrium is almost always assumed, this choice should be accurately motivated by specific experimental conditions, reviewed in [200]. Some works indeed have observed that, specifically in the context of a human body, this does not necessarily hold true [242]. Considering that in any case this assumption has eventually fostered many successful

modeling efforts, during our works we decided to not address these theoretical considerations in detail and to exploit the assumption in the majority of our models.

After its formalization, the HMM equation worked as a baseline model upon which, during decades of biochemical studies, many more complex formulations where defined. Progressively more complicated reaction mechanisms were modeled, involving multiple enzymatic subunits and different types of regulators. Every time the reaction rate is modeled taking into account all the known molecular interactions between the enzyme and all substrates, products and allosteric regulators, we consider this reaction as modeled with a mechanism-based, fully detailed and parameterized equation.

Approximate rate laws

On the other side, a specific attention should be dedicated to approximate rate laws. These are mainly aimed to provide generic and compact expressions that can be be applied to compute any reaction flux in the system. Taking a "one-size-fit-all" approach, simplified kinetics introduce higher approximations in the model, and thus rise the risk of providing erroneous model behaviors. As a main advantage, however, they can be used to test more easily the scalability of new computational methods to wider systems. Automated procedures and modular network structures can both be exploited to make the use of simplified kinetics worth of exploration. The most successful modeling frameworks based on approximate kinetics will be mentioned later on in this paragraph.

Differently from what just seen, the LMA has been also used to represent a complete enzyme catalyzed reaction:

$$2A \underset{k_r}{\overset{k_f}{\rightleftharpoons}} B \qquad v = k_f A^2 - k_r B \tag{4}$$

Law of Mass Action

where A and B are the substrate and product respectively and the exponents, also called reaction orders, equal the stoichiometric coefficients related to the specific biochemical compound. In this case both rate constants are proportional to the concentration of the enzyme, which is not represented explicitly in the expression. In this sense, the LMA is here considered an approximate kinetics. Information about the mechanisms, in fact, are lost, but its formulation can be derived from the sole interaction network. The equilibrium K_{eq} constant defining the ratio between the forward and reverse rates of a bidirectional interaction, can be used to rewrite equation into

$$v = k_f A^2 - \frac{1}{K_{eq}} B \quad , \quad K_{eq} = k_f / k_r \tag{5}$$

where K_{eq}. recapitulates the thermodynamic properties of the reaction, and can be calculated with

$$\Delta G_r^\circ = -RT \ln(K_{eq})$$

where ΔG_r° is the Gibbs free energy, T is the temperature in Kelvin and R is the universal gas constant. From equation 2 it can be noticed that for each

LMA modeled reaction, if thermodynamic information available allows us to define the value of K_{eq}, so that just one kinetic parameter is needed. This aspect made LMA particular appealing to be exploited by the MASS approach [73], later described.

Other simplified rate laws, formulated as power-law or log-linear functions based on linear Taylor approximation, were proposed to expand the descriptive capabilities of LMA while still retaining a compact formulation.

Generalized Mass Action and Biochemical System Theory

Both Generalized Mass Action (GMA) and S-system were defined by Savageau inside the Biochemical System Theory (BST) [274] . A general formulation is the following:

$$v_i = \gamma_i \prod_{j=1}^{n} X_j^{f_{ij}}$$

where the rate constant γ_i is non-negative and the kinetic orders f_{ij} are positive for variables $X_j$that have an augmenting or activating effect on v_i, negative for variables that have an inhibiting effect, and 0 for variables that do not have a direct effect on v_i at all.

A S-System model represents a simplified version of a GMA model, where one single power-law term aggregates all the influxes/effluxes to/from a metabolite pool. For instance, if the concentration of a metabolite is modified by two reactions producing it and two reactions consuming it, the ODE of that metabolite would be composed of four terms (two negative and two positive) in a GMA system model, while just of two aggregated terms (one negative and one positive) in a S-system model. while some situations can cause inconsistencies in S-Systems, they generally provide a highly compact description of the system but still able to retain its most important features. GMA and S-system models have been used extensively for applications in both in systems and synthetic biology [116].

Loglin and Linlog

Another approximate rate laws, the so called **Loglin** and **Linlog** rate laws, are inspired from the developments of Metabolic control analysis (MCA). MCA, one of the methods of sensitivity analysis discussed in paragraph 1.3, aims to assess how, around a specific reference condition, the model is influenced by changes in its variables and parameters. A reference condition is normally chosen as a specific steady-state of the system, characterized by reference steady-state fluxes, and reference concentration of metabolites and enzymes. Seen in the context of metabolic engineering, MCA identifies the variables and parameters that should be modified in order to control system's behavior. Developed in parallel by Kacser and Burns and Heinrich and Rapoport [281], it introduced the key concepts of *elasticities*, which are reaction specific properties, and *control coefficients*, global properties of

the system. An elasticity, which quantifies the effect of a metabolite, like a substrate S, on a reaction flux v, is defined as

$$\varepsilon_S^v = \frac{\partial v}{\partial S}\frac{S}{v} = \frac{\partial \ln v}{\partial \ln S}$$

Loglin and Linlog

The Loglin, and the later developed Linlog rate laws, use elasticities as kinetic parameters and offer approximations of the system around a specific steady state. Regarding the performances of GMA, S-system, Loglin and Linlog models, it is hard to evaluate them independently from the reference condition that is considered, and they loose applicability in cases in which, for practical reasons, the reference conditions cannot be characterized [192].

Henri-Michaelis-Menten

Also a **reversible** form of the **HMM** rate equation, although linked to well defined biochemical principles and assumptions, inevitably loses accuracy and thus should be considered an approximate kinetics every time it is used with assumptions that oversimplify the reaction mechanism disregarding some molecular interactions.

Hill

Another simplified rate law, the **Hill kinetics** was defined for multimeric enzymes, i.e. enzymes with multiple catalytic sites. If the binding of one substrate induces structural or electronic changes that result in altered affinities for the vacant sites, the curve of reaction velocity will no longer follow Henri-Michaelis-Menten kinetics but will display, instead, a sigmoidal shape. An enzyme with with these properties will be classified as allosteric, a term originally defined by Monod Changeux and Jacob in 1963. If an allosteric enzyme has n binding sites and if their cooperativity in substrate binding is very marked then the reaction is most commonly modeled with a Hill equation.

$$v = \frac{v_{max}[S]^n}{(K_M)^n + [S]^n}$$

Studying the impact of rate laws aproximations

Besides the fascinating potential of simplified rate laws, they should be adopted with caution. Authors in [42] recently explored the validity of using approximate rate laws with varying levels of assumptions in the context of a red blood cell (RBC) kinetic model. They found that HMM rate law with measured kinetic parameters could consistently reproduce the behavior of the system. When, instead, they additionally assumed enzyme saturation, neglected entirely enzyme behaviors or assumed Michaelis constants equal to substrates concentrations, substantial dynamic and structural issues would have arisen. They concluded that if, on one side, fully approximate models can effectively contain useful information, on the other fully-detailed mechanistic models become necessary to predict the system dynamics with a reasonable accuracy.
In [20] Bulik et al. raised the issue that models built with simplified kinetics show good approximations near a specific reference condition, but their performance tends to become poorer moving farther from it. In their analysis,

they compared the reliability of approximate rate laws to reproduce complex metabolic behaviors assessing the range of physiological conditions under which a kinetic model of erythrocyte metabolism based exclusively on simplified rate equations still adequately describes the system's behavior.
In specific they produced synthetic time course data with a model of the red blood cell [20]composed of full mechanistic rate equations for 25 enzymes and five transporters. Then, they built new models of the same system, where various types of simplified rate equations replaced all of the original ones.
Mass action as well as LinLog, Michaelis–Menten and power law rate laws were tested. The goodness of these approximate models was assessed comparing their steady state behaviour after perturbations in the consumption of ATP and glutathione (GSH) with the respective behavior of the original model, considered as a reference standard. The authors concluded that in most tested cases, the simplified models failed to reproduce these post-perturbation responses even close to the reference in vivo state.

The approaches and results presented in this thesis concern specifically mechanism-based models. In the following sections, in order to be more coherent and focused on this subject, we will only refer to computational approaches developed for mechanism-based models.

1.2 Creation of kinetic models

Once the parametric structure of the model is defined, the following task consists in the proper model construction.

In this section, we will review some of the most relevant approaches used to build kinetic models. Looking at the overall process, the work-flow of building a kinetic model can follow a bottom-up or a top-down approach. With a **bottom-up** strategy the behavior of the system is retrieved from the integration fundamental components, whereas **top-down** approaches seek to describe the global view of it. Both approaches present strengths and limitations, and often is the case that the most accurate results can be obtained combining them.

- Bottom-up approaches seek to capitalize all the amounts of biochemical knowledge accumulated in the literature and stored in databases. Important public databases such as BRENDA [202], SABIO-RK, [284] and KiMoSys [30], contain specific enzyme kinetics and the associated parameters, while BioModels [106]and JWS [146] serve as model repositories.
  With their strategy, bottom-up models try to reconstruct the behavior

of the system integrating information of both the structure and the kinetics of its specific components. This information ultimately consists of other models, or portions of models, already built and validated in previous works. It is often the case that these models had been originally produced with different goals, in different experimental conditions and with different approaches. Moreover, the available information is characterized by a different level of detail depending on which portion of the system we are observing. It can then be easily understood why the main difficulties of bottom-up strategies are related to the heterogeneity of data, which can consistently affect model outcomes. Despite these limitations, several models have been built [24, 28, 192, 215]

- Top-down approaches, on the contrary, do not rely on the mechanistic information already available. Instead, they try to exploit phenomenological data of the system as a whole to infer details of its components. Thus, while in bottom-up approaches parameters are retrieved fitting the individual components to the data, in the case of top-down approaches, data fitting is always performed on the entire model.

In the process of model creation, many approaches so far proposed have made use of approximate rate laws.

For instance Jamshidi and Palsson, in [73], formulated the Mass action stoichiometric simulation (MASS) approach, in which they applied the LMA rate law to create a model of the red blood cell, comprehensive of glycolysis and the PPP pathway.

Mass action stoichiometric simulation (MASS) approach

In this system, LMA models enzymatic reactions in two different ways. For the majority of system reactions, LMA describes the full catalysis from substrates and products. Thus, LMA is here used as an approximate kinetics. On the contrary,a few specific reactions that have a well documented key impact on the overall system, namely phosphofructokinase (PFK), Hexokinase(HK), Diphosphoglyceromutase (DPGM), Glucose-6 phosphate dehydrogenase (G6PDH), were decomposed into their elementary steps.
The MASS approach, which gives the name to a toolbox implemented in MATLAB, provides a complete parametrization of the model starting from information about the stoichiometry, thermodynamics and about metabolite concentrations in the system. Importantly, the model aims to describe the the behavior of the system in proximity of a reference steady state, for which experimental data on reaction fluxes and concentrations are available. The procedure by which a MASS model can be built is elegant for its simplicity, and composed of the following steps:

- Specify a particular steady-state flux distribution
- Identify the metabolite concentrations at the specified steady state
- Retrieve equilibrium constants from the literature or approximate them

- For each reaction, solve the linear equation 5 to calculate the forward rate constants

Once all LMA rate constants have been calculated, the MASS model can be used to reproduce the behavior of the system in proximity of the specific steady state.

Other very well known computational frameworks devoted to the creation of kinetic models of metabolism on wider reaction networks have been proposed in recent years. Here, for the sake of completeness, we will mention the most discussed by the community. The structural Kinetic Modeling Framework (SKM) [224] dissects the dynamics of the system relying mostly on its structure and building a local linear approximation of its behavior [192]
The Ensemble Modeling approach (EM) [251] generates a large set of candidate models, composed of elementary reactions represented with LMA, that achieve a certain steady state flux distribution. The ensemble of models is created through sampling techniques that explore wide ranges of values of the kinetic parameters, still constrained by thermodynamic principles. Additional data, for instance produced with knock-out experiments, are later used to identify which models of the ensemble are able to reproduce them. These models are retained in the ensemble, while all the others are discarded. If the initial ensemble is big enough, and if the data are adequate, the ensemble is progressively shrunk and one single parametrization is identified.
The Optimization and Risk Analysis of Complex Living Entities (ORACLE) [129]. ORACLE integrates information on stoichiometry, thermodynamics, concentrations and fluxes and exploits a sampling approach to compute reaction elasticities, that are used to parametrize the models.

The Approximate Bayesian Computation and General Reaction and Assembly Platform (ABC-GRASP) [191] was defined instead as a sampling-based framework that exploits Bayesian inference and can be used to build detailed kinetic models of metabolism. An accurate comparison among these methods is reported in [192].

Hybrid models of metabolism: detailed + simplified kinetic rate laws

The work previously described [20], where the authors compared the impact of different approximate rate laws on the behavior of the system, also presents a different perspective on the use of these simplified rate laws. The authors indeed proposed the idea of building hybrid kinetic models, in which part of the reactions are modeled with fully mechanistic rate equations while others are described with simplified kinetics. As the model is built with two levels of detail, it is important that central regulatory enzymes, whose behavior has a strong effect on the overall system, are identified and modeled with fully mechanistic rate equations. Exploiting the Structural Kinetic Modeling (SKM) method [224] to identify regulatory enzymes, they compared hybrid approximate-mechanistic kinetic models with fully approximate models and verified that the first yielded better perfor-

mances for almost all variants of the simplified rate equations tested.

Hybrid models of metabolism: mechanistic + constraint-based approaches

The idea of a metabolic system described two with different levels of accuracy had already been proposed in [298]. Here the network is divided into sub-portions or modules, that are either interpreted as static or dynamic. The modeling approach is thus hybrid: dynamic modules are modeled with kinetic rate laws while the behavior of static modules is characterized just at the steady state by Metabolic Flux Analysis (MFA), a constraint-based approach similar to FBA that only requires reaction stoichiometry.
As the total number of kinetic parameters in the resulting model is significantly reduced, the main aim of this hybrid technique consists in providing a dynamic description of the system through easier modeling efforts.
According to authors' conclusions, the algorithm shows good results, although inconsistencies might occur when (i) there exist many bottleneck reactions, that is, boundary reactions that cannot be easily assigned to, or simulated by, either the static or the dynamic modules; (ii) bottleneck reactions are not clearly identifiable; (iii) there are large fluctuations in the rate of reactions included in the static module.
These hybrid approaches [20, 298] offered sources of discussion and revision that pushed us to work on an approach (see chapter 3) that is in fact hybrid, as different portions of the system are described with different levels of detail. Even if many authors embrace a "something better than nothing perspective", the actual utility of these simplified models still has to be clarified.

A PRIORI ASSESSMENT OF MODEL INDETERMINATION Either the case we are using a top-down or a bottom-up approach, we need to acknowledge that the structure and parameters of computational models of cancer metabolism are often undetermined. Data scarcity, moreover, often causes large uncertainties in the inferred interactions and parameters.
Figure 3 offers a condensed picture of all the components that need to be defined in order to produce a kinetic model of metabolism. The picture summarizes also all the main sources from which the necessary model components can be obtained.

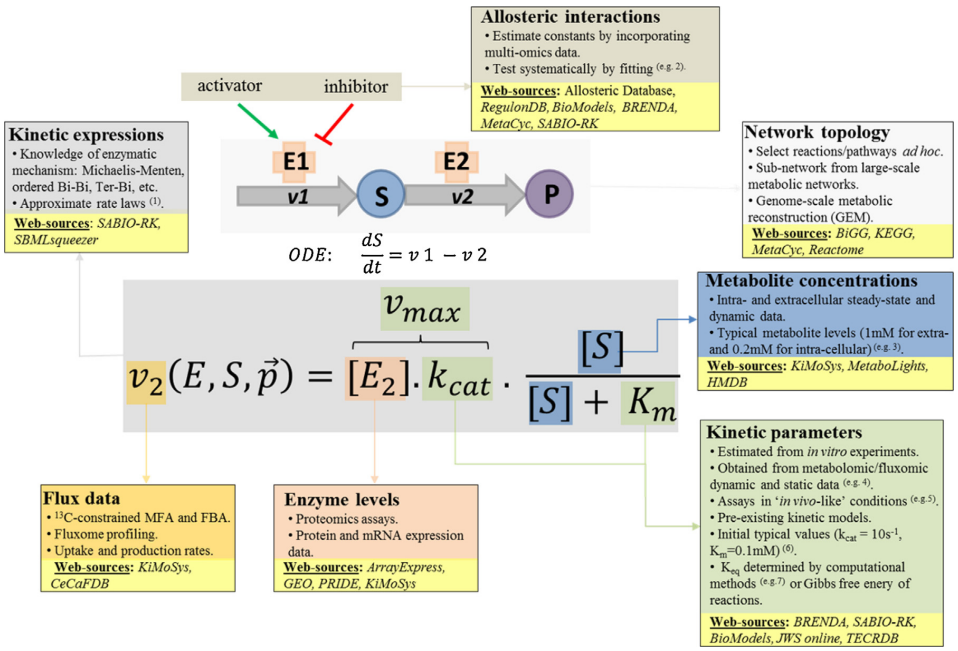

Figure 3: Overview of different types of data and related information that can be exploited to assemble ODE-based kinetic models of metabolism. Figure taken from [29]

All of the repositories of kinetic parameters we cited are far from being complete, and the experimental settings across different queries are often not fully detailed. Biochemical essays are commonly calculated executed in *in-vitro* conditions, characterised by stardard pH and temperature, and substrates and modifiers at saturating concentration. While for many biochemical models these parameters can be fruitfully used to describe different, *in-vivo*, conditions with good approximations, it has been seen that this is not generally true [242].

Many studies showed that, at the intracellular level, different isoforms of the same enzyme are present, and that every isoform has a specific affinity for its own group of molecular regulators [135]. In a given cell, the kinetic behavior of a metabolic reaction is then closely linked to the proportion of enzyme isoforms and to their kinetic differences [135] [12]. In many situations this information cannot be clarified, and, even if known, the list of modifiers for secondary isoforms often has not been fully clarified. For pathological conditions like cancer, the situation is even more complex. Mutations and copy number alterations of genes and isoform switches of transcripts can cause quantitative alterations of enzyme isoforms as well as changes in their affinity for regulators.

Structural identifiability

Among these preliminary analyses it should be additionally checked if the model parameters are structurally identifiable. Structural identifiability differs from practical identifiability as it guarantees that a unique parameter reconstruction would be possible if the response of the system to an arbitrary rich set of inputs was perfectly observed. Practical identifiability analysis

is instead performed *a posteriori*, and requires that parameter uncertainties are calculated and compared with a desired level of accuracy [13]. If the non-identifiability does not change for any data, these parameters are called structurally non-identifiable. On the contrary, if the non-identifiability can be remedied by data improvement, they are practically non-identifiable [107]. Structural identifiability is normally assessed calculating correlations among model parameters, as it was in fact done in the model presented in [186], that we will use in chapter 4

When the structure and the parameters of the models are unknown, approaches of Reverse Engineering (RE) and Parameter Estimation (PE) represent the two main classes of procedures aimed to infer them.

PARAMETER ESTIMATION Unless we are using a bottom-up approach and all kinetic parameters and initial concentrations of system variables can be found in literature, some procedure of parameter estimation has to be necessarily adopted. PE tries to carry out the automatic inference of kinetic parameters, starting from:

- a defined network of molecular interactions
- some experimental data. Time-dependent measurements, or time series, of biochemical concentrations in the system represent the most informative and useful type of data. The quality of the data has a great relevance: the more precise, densely sampled and repeated the measurements are, the more reliable will be the parameter inference. As it is almost never the case that we can access dynamic profiles for all biochemical concentrations, it becomes crucial that the data refer to components that pay a central role in the system [107]

Parameter inference can be performed with two different classes of methods: frequentist approaches or Bayesian approaches.

Frequentist approaches

Frequentist approaches employ optimization algorithms to identify the set of parameters that produce the maximal similarity between the model behavior and the data. This is done computing the Likelihood of each parameter set, which can be otherwise described as the plausibility of the set given specific observed data. Aim of the approach, which is thus based on Maximum likelihood estimation (MLE), is to use optimization algorithms to identify the parameter set with the maximum value of the likelihood function.

Bayesian approaches

Differently from frequentist methods, Bayesian methods use specific probability density functions to define parameters as random variables. Aim of these methods is to combine subjective information about the parameters (called priors and expressed as probability density functions) with the information contained in the data, and to compute an updated probability distribution, the posterior distribution.
Bayesian methods can proceed in two different ways. Similarly to frequentist inference methods they can use MLE and optimization algorithms to identify the candidate set of parameters that maximizes the likelihood function. On

the other hand Bayesian methods often exploit sampling-based techniques, like Monte Carlo approaches, to produce an ensemble of parameters that describe model behavior.
Despite their elevated computational costs, Bayesian methods are still the most affordable choice for PE of large and complex models. Intriguingly, thanks to the probabilistic definition of estimated parameters, Bayesian PE can be used iteratively considering the posterior in one experiment as the prior in a new experiment.
If PE is based on MLE, either in the frequentist or in the Bayesian case, it makes use of optimization algorithms.

Optimization processes: local vs global search methods

Optimization algorithms can be classified as local or global search methods. More specifically local methods are either gradient-free or gradient-based, depending whether they exploit or not the gradient (the matrix of the partial derivatives, or Jacobian) of the objective function with respect to the parameters.
Gradient-based algorithms are widely used for their high speed and convergence, but, if the landscape of the objective function is ragged, the solvers tend to remain trapped in local minima.
When the optimization problem is highly complex, as it is often the case in top-down approaches, a global search optimization should be better adopted. Global optimization methods are either deterministic or stochastic. Stochastic methods have lower computational costs but they need much more expertise to be properly parameterized. Examples of stochastic algorithms for global optimization are *simulated annealing* and many population-based meta heuristics, either based on evolutionary computation, like *genetic algorithms* or *differential evolution*, which exploit a process of Darwinian selection of candidate parameter sets, or based on swarm intelligence, like *Particle Swarm Optimization*, where parameters are described as autonomous agents.
When the models have complex structures and the data is scarce, MLE-based approaches often result in different parameter sets with equally optimal likelihood values and the parameters become non-identifiable [180]. In these situations, sampling-based, most often Monte Carlo based, approaches show a higher flexibility and better performances.

REVERSE ENGINEERING When mechanical interactions cannot be clearly defined, methods of reverse engineering try to infer them starting from the available data. It is often the case that both network structure and some kinetic parameters are unknown, and thus need to be inferred during the same optimization process. In these cases a PE procedure is embedded in a RE.
As an example, evolutionary techniques called Genetic Programming (GP), can for instance be exploited to generate a set of candidate networks of interactions, evaluate their behavior and identify the network that at best can represent the data. General limitations of RE approaches include:

- Indistinguishability of different network architectures that produce the same behavior
- failure of local search methods to identify the global optimum
- a very high computation time

PARAMETER UNCERTAINTY ANALYSIS After a specific parameter set has been identified, the uncertainty and quality of the fitted parameters should be assessed. In the frequentist statistical setting the goodness of the fit can be quantified with different approaches, here just cited: χ^2-test, likelihood ratio tests, Akaike Information Criteria (AIC), Bayesian information criterion (BIC) or Likelihood profile (LP) method [192].

Frequentist approaches

Monte Carlo-based methods can be also used to measure parameter uncertainty. In the case of frequentist Monte Carlo approaches, parameter uncertainty is normally calculated with bootstrapping methods, which, with or without the need of additional parameters, introduce some disturbance in the original experimental data and create new data sets, from which the parameter inference procedure is restarted.

In the case of Bayesian approaches, as parameters are defined in terms of probabilities, the analysis of parameter uncertainty is straightforward. Similarly to the process of parameter inference, Monte Carlo methods are often indispensable [192].

Bayesian approaches

In both of the frequentist and Bayesian setting, parameter uncertainty is quantified by confidence intervals. Once a set of parameters and their confidence intervals have been calculated, it is common to use these parameter values to calculate other quantities. During this process, it is important to consider how uncertainty propagates. Various methods, not mentioned here, can be adopted for the study of uncertainty propagation.

Uncertainty propagation

1.3 Sensitivity analysis (SA) and parameter sweep analysis (PSA

The role of a sensitivity analysis is twofold.

Sensitivity analysis (SA)

- SA can help elucidate the impact of parameter uncertainty on the behavior of the model. In this way, it can suggest which parameters require to be studied with more attention in order to reduce the uncertainty in model output, and, at the same time, which should not be given great relevance and could be withdrawn from the model.
- SA becomes fundamental to understand which are the elements of the models that have the larger influence on its behavior, unraveling the fundamental control mechanisms of the system. This kind of perspective can be exploited to identify, for instance, some liabilities of the system that could be targeted with some interventions. In this second use the SA can suggest which experimental tests deserve to be planned.

Fundamental requirements to perform a sensitivity analysis are a network topology, a defined set of parameters, considered as input of the analysis, a range of variability for these parameters, and the specification of a system variable that has to be treated as output of the analysis.

Sensitivity indexes

Sensitivity analysis algorithms aim to compute sensitivity indexes, which reflect how the output variable responds to changes in the input parameter. If the output is highly correlated with the input parameter, the sensitivity coefficient for that parameter is high, and small variations in the input result in wide changes in the output. When sensitivity coefficients are computed for many model parameters, they are normally ranked and can inform us on which parameters have the most-to-least influence on model outputs.

Computing sensitivity indexes: local vs global SA

If the variation in input parameters spans just the neighborhood of a reference value, SA is defined local, while if it spans a wide range of values is classified as global. **Local SA** should be used if we are in a situation where we have a good confidence on parameter values. When, in bottom-up approaches, instead, parameters are taken from many different sources and the uncertainty is high, methods of **global SA** should be better used, as they allow to scan a much wider range of parameter values. The use of global SA, although potentially permits better performances, however incurs in higher computational burdens. SA often takes a **"one at time" (OAT) approach**, in which one parameter at time is modified with all the other kept constant, and the output of interest is registered. This strategy is simple to implement and has reasonable computational costs, but interactions among parameters remain here hidden. In order to assess the impact of combinations of multiple parameters changes, instead, sampling methods like Latin Hypercube Sampling [205] become necessary to try to limit the computational efforts.

The calculation of sensitivity coefficients can be done with different methods. Two of the most used are

- Partial derivative methods. With these approaches, sensitivity indexes are calculated as the partial derivative of the output with respect to the input, in a given point in the input space

$$S_{X_i,Y} = \left.\frac{\partial Y}{\partial X_i}\right|_{X^0}$$

  The most famous example of a sensitivity analysis approach that uses partial derivatives is Metabolic Control Analysis, which, as already introduced, computes elasticities and control coefficients to identify where over the networks is the control distributed.

- Variance-based methods. In this case, sensitivity coefficients are calculated as the the proportion of the output variance that is caused by the variance of a specific input. In the first order coefficient, the contributions of a single parameter to the output variable can be calculated as

$$S_i = \frac{Var[E(Y|x_i)]}{Var(Y)}$$

Second order coefficients, instead, measure the contribution of a pair of parameters to the variance in the output.

Differently form sensitivity analysis, parameter sweep analysis (PSA) does not aim to compute sensitivity coefficients. Instead, PSA seeks to identify some combinations of model parameters that produce a desired outcome. In PSA, a large number of model behaviors are computed at different values of the input parameters and are later compared with some experimental data.

1.4 Model validation and refinement

In order to show their ultimate utility, computational models should be always validated and possibly refined. The process can follow multiple phases of validation and refinement, where the outcomes of model implementation are iteratively compared to the experimental data, some new hypothesis are generated, and some model component is modified accordingly.

Quantitative validation

Computational models of metabolism can be validated both qualitatively or quantitatively. A quantitative validation, in which model outputs are compared in value to some experimental findings, is always preferable. In an ideal scenario, time course experimental profiles of some biochemical compound are plotted and compared to the time-dependent behavior of the corresponding variables in the model. In analogy to what already observed for the parameter estimation phase, these data are often not available. In this case, model outputs can be compared to static measurements, like the values of fluxes and concentrations at the steady state. The ability of a model to reproduce new phenomena that were not considered during its construction is a very sound proof of its validity. In biology, a widely used procedure to achieve this, and at the same time inspire new experiments, is the evaluation of the outcomes of gene deletion experiments. Mechanism-based models generally show good flexibility in the simulation of these intervention. For instance, a gene deletion meant to suppress the synthesis of a specific protein, like a metabolic enzyme, can be easily reproduced *in silico* by simply changing the concentration of the enzyme to a low value or to zero.

Qualitative validation

If the system under study has an intrinsic complexity a quantitative validation might, however, be impossible to perform. In these cases, a qualitative validation becomes the only alternative. In a qualitative assessment, we are guided by the biological evidences available in the literature. We might check, for instance, that the model correctly consumes and produces some metabolite, we might verify that some required precursor is correctly synthesized, that some reaction fluxes are active or suppressed, or that some oscillations or bistabilities are present.

An observation should be done here. It is evident that in a qualitative validation many models could potentially produce equally plausible findings. This, however, should not be used as an argument to disprove the validity of the model. We believe, in fact, that as long as a model mirrors our current knowledge of the system, being it scarce or detailed, it deserves to be con-

sidered a possible, not necessarily true, explanation of the phenomena.

1.5 Data integration

A model that does not embed some information of the actual phenomena of study cannot aspire to describe reality.

A first way to customize metabolic models with experimental data is to relate the activity of the reactions as they are described in the model to the measured abundance of mRNAs and proteins.

Many efforts to incorporate both absolute and differential expression data into metabolic models have been attempted in recent years [14]. Thanks to its growing availability, gene-expression data represent an attractive source of information to be integrated in computational models. The integration of this data is, however, not devoid of complications. The main issues are related to the possibility to infer protein abundance from mRNA quantifications. Even though, according to the central dogma, the sequence of nucleotides in a gene determines the sequence of its mRNA product, and an mRNA's sequence determines the amino acid sequence of the resulting polypeptide, there is no trivial relationship between the concentration of a transcript and the concentrations of the proteins derived from a particular locus [114].

This subject has been largely reviewed and discussed in many works [121] [145]. The conclusions drawn by most of the authors come out in favor of a positive but weak correlation between mRNAs and protein abundance. However, it seems at least possible to treat cells in steady-state as a separate case. Cells can be considered at the steady state if the average protein and/or mRNA levels remain relatively stable over time (normally above several hours). If the experimental data are collected under these circumstances, gene-to-gene variation of protein levels can be primarily attributed to their respective mRNA levels [114]. In [9] the authors accurately analyze the procedure by which gene expression values, using gene-protein-reaction (GPR) rules, are used to infer protein abundance. Differently from what is done inside the COBRA Toolbox [64] and in other works, they explain how ANDs and ORs should be better replaced with minimums, and sums, respectively. Considering the biological explanation of these logics, we believe that, if two isoforms of the same enzyme are present, they are both functional units that can work in parallel. The activity of the resulting isoform mixture should then be defined as the sum of the single activities rather than the maximum. This motivates our choice to replacement of the ORs with sum() instead of max() in all our experiments. If the ORtosum() transformation is applied with an automated procedure, however, the conversion of the boolean logics into numerical operators is not trivial. GPR rules often contain several nested operators that can cause erroneous computations. As presented by the authors, we report here 1.5 an example of a problem that can occur with

this mapping where some genes' expression levels may be counted more than once.

$$r_1 := [A \text{ and } B] \text{ or } [A[A \text{ and } C] \longrightarrow e_1 = min(a, b) + min(a, c)$$

r_1 is a reaction rule and e_1 is the corresponding estimated complex abundance level. Lower case letters are shorthand for the expression level of the corresponding gene ID in uppercase; for example, $\alpha = E(A)$, where $E(A)$ is the expression of gene A. Supposing A is the minimum, then if we just evaluate r_1 directly, A will be counted twice. In their paper the authors propose an algorithm that deals with these situations.
In case proteomics data for the conditon of interest are available, these should in principle be preferred to gene expression data, and used to refine the models. The data produced by proteomics techniques are normally linked to the IDs of the related genes, thus they should be integrated into the models through GPRs mapping, analogously to what is done with gene expression data.

Despite the fact that these regulatory processes can be still understood and modeled with great uncertainties, in the last decade several approaches to integrate gene and protein expression data into metabolic models have been proposed. Considering that high-throughput technologies offer us information at the cell level, most of the efforts have focused on developing methods to integrate these data into genome-scale constraint-based models. These methods are aimed at constructing cell or tissue-specific networks of metabolic reactions starting from a GS map of the human metabolism like Recon 1, Recon 2 or the more recent versions. The algorithms use (mostly, but not exclusively) RNA sequencing (RNA-seq) data to predict which subset of reactions are active in each cell line, depending on a user defined threshold.

A critical and quantitative comparison between several of these algorithms can be found in [119]. In [147] the authors gave a quantitative evaluation of how different pipelines of data integration resulted in different genome-scale models predictions. Hundreds of models of four cancer cell lines were built using three sets of constraints based on exometabolomics data, six algorithms, and four gene expression thresholds. Referring to the distinction proposed in [147] we can group these algorithms into three main families:

- The GIMME-like family minimizes flux through reactions associated with low gene expression.
- The iMAT-like family finds an optimal trade-off between removing reactions associated with low gene expression, and keeping reactions whose genes/enzymes are highly expressed.
- In the MBA-like family, the algorithms identifies a sets of core reactions that should be retained and active, while remove other reactions if possible.

Following the descriptions reported in [147] we will briefly explain here the basic procedures implemented in some of them.

- **GIMME** Gene Inactivity Moderated by Metabolism and Expression (GIMME) [10], finds a flux distribution that is consistent with a given biological objective and that minimizes the utilization of reactions classified as inactive, weighted by the difference between their expression level and a given threshold. The authors used this method to model adaptive evolution in E. Coli strains, and to create tissue-specific human cell models.

- **iMAT** The integrative metabolic analysis tool (iMAT) [278] considers gene expression to divide reactions into two groups: highly and lowly expressed. It then finds a flux distribution that maximizes the consistency with this classification. iMAT has the advantage of not requiring the definition of a biological objective, facilitating the analysis of biological systems, such as multi-cellular organisms, where this definition is not so clear.

- **mCADRE** Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE) [307] uses the gene expression levels and the network topology to calculate connectivity-based evidence scores for all reactions in a model. These scores are used to determine which reactions should be removed from the generic model to create a context-specific model.

- **INIT** The Integrative Network Inference for Tissues algorithm (INIT) uses proteomic data from the Human Protein Atlas, but can also use transcriptomic data to build tissue-specific models [2]. It maximizes the activation of certain reactions based on a qualitative confidence score while minimizing the utilization of reactions associated with absent proteins. One of the novel aspects of this method is the relaxation of the steady-state condition to allow a small net accumulation rate for internal metabolites. If there is evidence for the presence of a metabolite, this accumulation is imposed in order to prevent the removal of the reactions necessary for its synthesis.

A recent work presented single-cell Flux Balance Analysis (scFBA) [32] a computational framework to translate single cell transcriptomes into single-cell fluxomes. In this framework scRNA-seq data are integrated into a multi-scale stoichiometric model of cancer population. The authors showed that, with this integration process, some clusters of cells with different growth rates within the population can be identified.

Metabolomics and fluxomic instead generate phenomenological data that, due to their nature, represent the best source of information to be exploited

by parameter inference methods. Either time course metabolite concentrations, steady state metabolite concentrations and steady state fluxes can be effectively used for these goals.
The specific procedures of data extraction and manipulation for all of these experimental techniques will not be described here.

2 QUANTITATIVE MODELS OF CANCER METABOLISM

In this section we will review the most important efforts of modeling cancer metabolism. Very different types of models can be found in the literature. Schematically, we will divide them into 3 groups: intracellular constraint-based models, intracellular mechanism-based models and population-based models.

CONSTRAINT-BASED MODELS The algorithms of data integration presented in section 1.5 allowed the generation of several models of cancer metabolism:

- In [305] the authors studied the hallmarks of metabolic alterations in cancer calculating metabolic flux states for the NCI-60 cell line collection and correlated the variance between these states with the phenotypic characteristics of each cell line. To obtain a flux distribution for all cancer cell lines they started from the definition of a core metabolic model of a cancer cell that included the pathways related to the known most relevant metabolic functions.
  The optimization problem was then constrained by uptake fluxes based on cell line-specific measurements. These were retrieved from a recent study [72] where exometabolomic data were used to infer uptake and secretion profiles for the NCI-60 panel. Non-cell-line-specific approximations, specifically ATP maintenance, oxygen uptake, and certain flux splits, were used as additional constraints. They also defined a partially cell line-specific biomass function for each cell line based on cell sizes and on the typical composition of mammalian cells. FBA was performed with an NADPH production objective, which interestingly determined the highest agreement with 13C tracing data. The resulting flux distributions showed that glutamine was taken up on average 32 times more than its biosynthetic requirement, a finding that confirms the addiction of cancer cells for glutamine. These results led the authors to speculate that the compliance to oxidative stress might represent a fundamental requirement for cancer cells, and that this might be achieved via the catabolism of glutamine through the mitochondrial NADPH-producing malic enzyme pathway.

- In [79] a hybridoma cell line was used as a model for cancer to produce experimental data that were used to build a constraint-based model of 152 reactions including TCA and PPP pathways, the electron transport chain (ETC), and the one-carbon metabolism pathway. The model was interrogated to investigate metabolic requirements of mammalian cell proliferation. The authors focused on the study of the process of production of 1C units from serine and glycine catabolism, as well as the contribution of glutamine to the total cellular nitrogen and carbon to clarify the most important requirements for biomass production. Moving forward from these analyses they used constraint-based FBA simulations to model the metabolic effects of metformin, a well known antidiabetic drug, considered as a potential cancer therapeutic.

- In another work, [44] Liquid-Cromatography-MS-based experimental data, together with an FBA model were used to characterize cancer metabolism, more specifically the contribution of aerobic glycolysis and oxidative phosphorylation to the total ATP production, and the relative contribution of glucose, glutamine and other nutrients to the maintenance of the reducing power. FBA was used to calculate a steady-state flux distribution that could match metabolite uptake and secretion rates measured [72] and 72-h dynamic profiles of some intracellular metabolite concentrations. FBA predictions supported literature evidences that glutamine is able to drive TCA cycle flux, but also interestingly showed that oxidative phosphorylation remains the largest quantitative contributor to ATP production and that Ras oncogene has no net effect on ATP production.

- In [54] A genome-scale constraint-based model specific for clear cell renal carcinoma (ccRCC) cells was presented. Model network was reconstructed with the INIT algorithm [3], integrating gene and protein expression data. Exploiting experimentally measured fluxes for a number of exchange metabolites FBA was used to predict essential genes. Essential genes are the genes whose function is critical for the survival of the cell and thus may be considered as potential therapeutic targets. This scenario was simulated *in-silico* with knock-out experiments in which the flux of the reaction under observation was constrained to zero. Discussing FBA results, the authors argued that ccRCC depends on the expression of AGPAT6, GALT, GCLC, GSS, and RRM2B, which, although essential for cancer cells, are potentially nonessential in normal cells.

- After some studies demonstrated how cancer cells could use alternative glycolytic pathways with net zero ATP production, in [267]. the authors built a genome-scale constraint-based model of a Myc-driven tumor accounting for cytoplasm solvent capacity.
  The study uncovered a novel pathway for ATP generation that starts from 3-phosphoglycerate and involves reactions of the serine biosynthesis, of one-carbon metabolism and of the glycine cleavage system.

The results showed that cancer cells may exploit different pathways for ATP generation, that either maximize ATP yield per mole of substrate or ATP yield per occupied volume fraction.

- In [48] genome-scale constraint-based models of NCI-60 cell lines were used to identify genes essential for cellular proliferation. Starting from Rz, the subset of active reactions was defined with the Model Building Algorithm (MBA) [75] which automatically integrates gene expression microarray data. The model was able to predict 52 cytostatic drug targets, 40% of which already targeted by known anticancer drugs, as well as combinations of synthetic lethal drug targets.

- In a different study [290] more than 280 models of normal and cancer cell-lines were built with PRIME algorithm using gene expression data, while measured proliferation rates supported both the testing and validation phases. The model predicted that Malonyl-CoA decayboxylase (MLYCD) gene can be targeted to affect cancer cell growth. The result was tested in both leukemia and renal cancer cell lines, vs normal lymphoblast and renal cell lines, confirming the prediction. The authors however acknowledged that assuming maximal cell growth, used as FBA objective function, and neglecting enzyme variants, potentially cause limitations in model predictions.

- Differently from the approaches mentioned so far, the work in [182] presented a model of cancer metabolism based on dynamic FBA, which, starting from an initial number of cells C_0 and an initial glucose concentration in the media GLC_0, could simulate both intracellular flux distributions and cell growth. The model includes glycolysis, the TCA cycle, the Pentose Phosphate Pathway (PPP), glutaminolysis and the oxidative phosphorylation. Hela cell lines were used to study the growth kinetics and qualitatively compare it with *in silico* predictions. Noticeably, the authors acknowledged the difficulties related to the definition of an appropriate objective function for cancer cells. The problem was addressed starting from a review of the literature and general considerations on the functioning of metabolic systems, and then defining a generic objective function as a linear combination of extracellular lactate, ATP, and mitochondrial oxaloacetate, citrate, ribose 5-phosphate and NADPH. 1000 instances of the objective function were then chosen by randomly sampling the stoichiometric coefficients. Metabolic reactions with a central role in cancer cell growth were identified by two constraints: low flux variability and high enzymatic essentiality for cancer cell growth. Together, these constraints constitute computational criteria for selecting those reactions that ensure a low redundancy on metabolite synthesis with a maximal effect for decreasing cell growth. The model successfully identified some enzymes that are currently considered as potential drug targets.

MECHANISM-BASED MODELS Aiming to move beyond the characterization of flux distributions at the steady state, in [82] the first kinetic model of cancer metabolism was developed. The model, containing 58 reactions modeled in a Log-linear form, was built with the ensemble modeling (EM) approach, cited in section 1.1. Sampling for reaction reversibilities and enzyme fractions under thermodynamic and steady state constraints an ensemble of models was generated.Then, all models in the ensemble were computationally perturbed and the steady state fluxes they predicted were compared to experimental perturbation results. Models that captures the experimental results were retained. The resulting models predicted transaldolase (TALA) and succinyl-CoA ligase (SUCOAS1m) to cause a significant reduction in growth rate when repressed. Furthermore, the results suggested that the simultaneous repression of the two enzyme targets would result in a 3-fold increase in the repression of growth rate.

POPULATION-BASED MODELS In [31] is presented one of the first attempts to represent intra-population metabolic heterogeneity in tumor. In this work, the authors introduced popFBA, an extension of FBA that takes into account intra-tumor heterogeneity and interactions among different cell populations within the same tumor. popFBA was applied to a model of 10 clones of the metabolic network of human central carbon metabolism, simulating a plasma supply of glucose, glutamine and oxygen, assuming equal bounds for the reactions of the 10 clones and an internal exchange of lactate, glutamine, glutamate and ammonia. The models showed that clones may follow several different metabolic paths and cooperate to maximize the growth of the total population. Also, the model showed how alternative nutrients in plasma supply and/or a inhomogeneous distribution of oxygen provision may affect the landscape of heterogeneous phenotypes.

Metabolic heterogeneity in cancer

Metabolic heterogeneity non in cancer

Modeling efforts that represent intra-population metabolic heterogeneity in a non-cancer scenario deserve to be briefly presented. From a modeling perspective a population of different bacterial species and a population of cancer sub-clones with heterogeneous metabolic traits can be represented similarly. Even if the context and the environment are hugely different among the two situations, both can be viewed as multitudes of independent cells that consume substrates and secrete wasteful compounds, thus exchanging metabolites with a shared extracellular environment. In the most recent years growing efforts to model natural and synthetic communities of bacteria have been made. Many of the approaches that are valid for population of microorganisms could be in fact potentially applied to describe metabolic phenomena in cancer.
The modeling strategies that can most adequately represent metabolic interactions at the population level have been accurately reviewed recently in [218].
In principle, all the classes of modeling approaches that have been discussed so far could be applied also to represent metabolic systems spanning populations of cells. It should be highlighted, however, that the increased size

of the biological system strongly conditions the applicability of these methods. If high number of cells participate to metabolic interactions, in fact, the intracellular metabolic system of each of them should be represented more schematically. Considering the costs we currently need to afford to build and solve our models, the size of intracellular network should be limited to toy or core models, and the intracellular steady state should be assumed. Limiting to quantitative models, great part of the works available in literature describe intracellular reactions with constraint-based modeling models. We list here a few of these examples:

- In their pioneering work, Stoylar et al. [226], built a stoichiometric model for a microbial consortium, in which microbial species were treated as internal compartments and an intracellular flux distribution for each organism was obtained maximizing the weighted sum of species biomasses.

- Under the name of community Flux Balance Analysis (cFBA), was instead proposed an extension of FBA that can be used to model populations of cells. cFBA predicts for communities at balanced growth the maximal community growth rate, the required rates of metabolic reactions within and between microbes and the relative species abundances [81].
  Zomorrodi and Maranas [306] developed instead OptCom, a generalized computational platform for cFBA, in which a bi-level optimization problem where both community-level and individual cell-level objective functions are used.

- Thanks to its ability to describe changes in extracellular concentrations, dynamic Flux balance analysis is well suited to model metabolic interactions in a population of cells.
  In a variant of OptCom, an implementation of dynamic flux balance analysis that includes uptake kinetics was also presented. Differently form OptCom, DyMMM, a framework based on dFBA used to model competitive and syntrophic (cross-feeding) communities, considers instead just a community-level objective function. [218]
  The Computation of Microbial Ecosystems in Time and Space (COMETS) represents instead a framework in which dFBA is integrated with diffusion on a lattice to reconstruct the metabolic behavior of a colony of cells. Each different metabolic subpopulation is represented with a different FBA model and at each time point metabolites concentrations in the extracellular space are updated according to the computed uptake and release fluxes. COMETS was able to reproduce the composition of two and three-species communities at the equilibrium [61].

3 DATA MINING

Data mining is a process by which large amounts of data are analyzed with the goal of discovering patterns and rules [90]. This discovery implies a process of learning, which can be classified as supervised and unsupervised . In supervised learning some instances of data, considered as targets of the analysis, are described by a model generated through the rest of the data [90].

Supervised and unsupervised learning

Supervised learning algorithms can be used for classification, prediction or estimation purposes [90]. With unsupervised learning algorithms, on the other hand, relationships between the data are highlighted, without any training phase. Association, visualization and clustering are some examples.

The approaches of data integration presented in this thesis motivated us to focus on tasks of data classification and clustering. More specifically, the classification and clustering problems we faced in our work can be defined as multiclass multivariate time-series classification and clustering problems. Time series classification (TSC) problems are differentiated from traditional classification problems because the attributes, e.g. metabolite concentration profiles, have a temporal ordering [211]. It is often the case that datasets are not huge, with the result of the number of objects in the train/test splits being relatively small. This prevents the use of the majority of machine learning algorithms, which require a high number of objects to produce accurate results.

Time series classification problems

Distance measures

Both for classification and clustering purposes, the distance between objects needs to be quantified. Many distance measures are available. The simplest and best known is the Euclidean Distance. In the case of time-series objects X and Y composed of the same number of temporal observations N, the Euclidean distance among them can be computed as:

$$d = \sqrt{\sum_{i=1}^{N} (Xi - Yi)^2}$$

While the Euclidean distance is easy to calculate, the standard benchmark elastic distance measure is, however, dynamic time warping (DTW). Core DTW strength is the ability to deal with times-series of different lengths and to transform ("warp") them non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. In DTW algorithms, first a nxm local cost matrix (LCM) is created, where the number of rows n and columns m respectively equal the number of observations in the two time series [56]. Each element of LCM is defined as the squared Euclidean distance computed between a couple of elements of the two time series. A warping path P is a contiguous set of matrix elements that spans all rows and columns of the LCM and defines a mapping between the two objects. Starting from the LCM, DTW algorithm searches for an optimal warping path between X and Y having the minimal

total distance among all possible warping paths. The DTW distance between X and Y is then defined as the sum of all LCM elements that compose the optimal warping path.
In the case of multivariate time series, calculating the DTW distance becomes more complex and computationally expensive. For this specific problems, different approaches have been proposed in the literature [53]. In order to cluster or classify a group of k objects, DTW distances need to be calculated pairwise among all k objects. Output of this procedure is a kxk distance matrix that can be used as input for a classification or clustering algorithm. The list of clustering and classification algorithms developed and available is extensive, and we will not provide a more detailed overview of them here. Suffice it to say that the several factors such as the number of variates and classes, differences in the number of observations, or the presence of motifs that repeat in time should affect the choice of the algorithm which is used.

3 METHODS

REPRESENTING STOCHASTICITY The development of all the approaches we will present later in this chapter is grounded on a stochastic representation of the behavior of metabolic systems based on Stochastic Petri Nets and its extensions.

Before we move to explain our approaches we thus give here an introduction on stochastic models, and explain how, in the context of metabolic system, they can be well approximated by deterministic models composed of ODEs.

To present stochastic modeling, we need to start defining a random variable. A **random variable**, by the name itself, is a variable which can take random values with a specific probability. If the number of possible values the variable can take, D, is countable, then the variable is said to be *discrete*. For a discrete random variable X, its *probability mass function* $p(a)$ can be defined by *Random variable*

$$p(a) = P\{X = a\} \quad , a \in D$$

On the other hand, we say that X is a *continuous* random variable if there exists a nonnegative function $f(x)$, defined for all real $x \in \mathbb{R}$, having the property that for any set $B \in \mathbb{R}$

$$P\{X \in B\} = \int_B f(x)dx$$

The function f (x) is called the *probability density function* of the random variable X.

A **stochastic process** $X(\nu), \nu \in T$ is a collection of random variables. This means that, for each $\nu \in T, X(\nu)$ is a random variable. The set T is called the index set of the process. The index ν is often interpreted as time and, as a result, we refer to $X(\nu)$ as an instance of the process at time ν. *Stochastic process*

A stochastic process is said to be a *discrete-time process* if T consists of a countable set. If $\{\nu \in \mathbb{R}\}$, instead, the stochastic process is said to be a *continuous-time process*. For instance, $X(\nu)$, $\nu \in \mathbb{N}$ is a discrete-time stochastic process indexed by the nonnegative integers, while $X(\nu), \nu > 0$ is a continuous-time stochastic process indexed by the nonnegative real numbers.

The *state space* of a stochastic process is defined as the set of all possible values that the random variables $X(\nu)$ can assume.

A stochastic process has the **markovian property** if the conditional distribution of the future $X(\nu + \tau)$ given the present $X(\nu)$ and the past $X(u)$, $0 \leqslant u < \nu$, depends only on the present and is independent of the past. *The markovian property*

A stochastic process $X_\nu, \nu \in \mathbb{N}$ displaying the markovian property is termed **Discrete Time Markov Chain**. If $X_\nu = i, \nu \in \mathbb{N}$, then the process is said to be in state i at time ν. We suppose that whenever the process is in state i, there is a fixed probability P_{ij} that it will next be in state j. That is, we suppose that

$$P\{X_{\nu+\tau} = j \mid X_\nu = i_\nu,\ X_{\nu-1} = i_{\nu-1}, ...,\ X_0 = i_0\} = P\{X_{\nu+\tau} = j \mid X_\nu = i_\nu\} \tag{6}$$

with $\tau \in \mathbb{N}$ and $i_\nu, j \in \mathbb{R}^+ \forall \nu$.

Moving to the case of continuous time, a process

$$\{X(\nu)\}_{\nu \in \mathbb{R}^+}$$

Continuous Time Markov Chains

is a **Continuous-Time Markov Chain** (**CTMC**) if for all $\tau, u \in \mathbb{R}^+, u \leqslant \nu$ and $i, j, x(u) \in \mathbb{N}$:

$$\begin{aligned} P\{X(\nu+\tau) = j \mid X(\nu) = i, X(u) = x(u), \quad & 0 \leqslant u < \nu\} \\ & = P\{X(\nu+\tau) = j \mid X(\nu) = i\} \end{aligned} \tag{7}$$

The importance of CTMCs for the description of physical and chemical processes was evidenced in the works by Daniel T Gillespie. In 1977, he developed a theory based on the hypothesis that collisions among molecules, in constant volumes and at constant temperatures, are random. With his theory, he was able to show that the kinetics of the chemical reactions deriving from these collisions corresponds to an underlying stochastic process that is a Continuous Time Markov Chain (CTMC).

In order to simulate the evolution of stochastic processes like CTMCs, we need to introduce the concept of exponentially distributed random variables. An **exponential random variable** with parameter λ is a continuous random variable with a probability density function given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{8}$$

Random variables that are exponentially distributed with rate λ can be used to represent the times that separate two consecutive events in the evolution of stochastic processes like CTMCs.

Memoryless random variable

The exponential distribution, in fact, has the unique property to be memoryless. A random variable X is said to be without memory, or memoryless, if

$$P\{X > x + y \mid X > x\} = P\{X > y\} \quad \forall\, x, y \geqslant 0. \tag{9}$$

In 9 the first term expresses the *conditional probability* that $X > x + y$, *knowing that* $X > x$.

A CTMC can represent any direct graph with labeled transitions, where the value of the label describes the rate associated with that change of state. The graph in figure 4, for instance, is described by a CTMC where two states are possible, and λ and μ define the rates at which the two states can be left. The rates of transition between states are recapitulated in the transition rate matrix, Q, also defined infinitesimal generator of the CTMC.

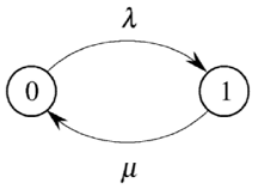

Figure 4: State transition rate diagram

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

Infinitesimal generator of the CTMC in figure 4

Diagonal elements of Q q_{ii} are defined such that

$$q_{ii} = -\sum_{j \neq i} q_{ij} \quad \text{and} \quad \sum_{j} q_{ij} = 0 \quad \forall i$$

and thus they represent the probability that the system remains in state i.

The exact solution of the CTMC at time t amounts to the computation of the solution of a set of differential equations called **forward Chapman-Kolmogorov equations**, defined as:

Chapman-Kolmogorov equations

$$\frac{d\pi_j(t)}{dt} = \sum_{i \neq j} \pi_i(t) q_{ij} - \pi_j(t) q_j \quad ,$$

$$\frac{d\pi(t)}{dt} = \pi(t) Q \quad ,$$

where $\pi_i(t)$ describes the solution of the system and consists in the probability that the system is in state i at time t. Q is the infinitesimal generator. For instance, the Chapman-Kolmogorov equations that describe the behavior of the system in figure 4, if the system is in state 0 at time 0 with probability 1, is

$$\begin{cases} \frac{d(\pi_0)(t)}{dt} = -\lambda \pi_0(t) + \mu \pi_1(t) \\ \frac{d(\pi_1)(t)}{dt} = -\mu \pi_1(t) + \lambda \pi_0(t) \\ \pi_0(0) = 1 \\ \pi_1(0) = 0 \end{cases} . \tag{10}$$

The number of equations that need to be written equal the number of states in the CTMC.

Fluidification

FLUIDIFICATION As the number of Chapman-Kolmogorov equations that need to be written equals the number of states in the CTMC, the simulation of the system often becomes computationally prohibitive. In fact, even small extensions of the network, or increments in the number of the entities described, can cause state space of the process to rapidly expand. In the case of metabolic models, where our intention is often to describe several reactions, and where the number of biochemical compounds is high, solving the system of the Chapman-Kolmogorov equations becomes practically infeasible.

With the goal of circumventing these issues, the pioneering works by Kurtz showed that the stochastic process can be approximated as a deterministic one, in which each system quantity is described by one ODE. Since then, in fact, Ordinary Differential Equations have represented the formalism most used for mechanism-based models of metabolism. This transformation from a stochastic model to an approximated, deterministic, one, is also described as "fluidification".

STOCHASTIC PETRI NETS Since working at the level of CTMCs can be computationally expensive and require advanced mathematical and modeling skills, in this section we introduce the Petri Net (PN) formalism which allows one to model the system as a parametric graphical diagram that makes easier and faster the model creation and its comprehension. Moreover, PN provide the possibility of automatically derive qualitative and quantitative properties with both numerical and analytical methods

Petri Nets PN and their extensions [161, 162] are a family of graphical modeling formalisms well suited for modeling *Discrete Event Dynamic Systems* (DEDS): they have been satisfactorily applied to fields such as communication networks, computer systems, manufacturing systems, but also applied to biological systems. Simple and intuitive in its graphical appearance, PNs are able to facilitate the process of model creation.

PNs are bipartite directed graphs with two types of nodes: *places*, graphically represented as circles, that correspond to the system variables (e.g. enzymes and metabolites), and transitions, graphically represented as rectangles, that correspond to the events (e.g. interactions among biochemical entities) that lead the system to evolve. An explanatory PN model representing glycolysis in human red blood cells is shown in Fig. 9. Places can be marked with tokens, graphically represented as black dots, that, in the context of systems biology, normally describe the number of molecules of the corresponding entities. The state of a PN, called *marking*, is defined as the number of tokens in each place of the net. An example of marking for the PN in Fig. 9 is showed in the third column of the Table in Fig. 10 .

Arcs connect places to transitions and vice versa, and express the relationships between states and event occurrences. Arcs are divided into input and output arcs: input arcs are directed towards a place, while output arcs move away from a place. Each arc is labeled with its multiplicity, a number that represents the amount of tokens that are moved when the transition occurs. This occurrence is called *firing*. A transition *fires* only if it is enabled, meaning the markings on its input places equal or exceed the multiplicities of the corresponding input arcs.

Stochastic Petri Nets In our work we focused on Stochastic PNs (SPNs) [133], in which transition firings are dictated by exponentially distributed random

delays, which are interpreted as durations of certain activities. SPNs are thus suited to give a high level graphical representation of a stochastic process. In specific, the stochastic process which underlies the behavior of an SPN is a CTMC, whose state space is isomorphic to the reachability set of the SPN. Thanks to this assumption the temporal behavior of the system can be modeled with a random process governed by the so-called Chapman-Kolmogorov differential equations [46]. These equations correspond to the Master Chemical Equations [55] that are used to describe the behaviour of biological systems, thus making this formalism quite attractive for these types of applications.

The formal definition of an SPN is the following:

Definition 1 (SPN). *A SPN is a tuple:*

$$\mathcal{N}_{SPN} = \langle P, T, \mathcal{I}, \mathcal{O}, \lambda, \mathbf{m}_0 \rangle$$

where:

- P *is a finite and not empty set of places;*
- T *is a finite and not empty set of transitions, such that* $P \cap T = \emptyset$*;*
- $\forall p \in P, t \in T,\ \mathcal{I}(p,t), \mathcal{O}(p,t) \to \mathbb{N}$ *are the pre- and post- incidence matrices, whose elements represent respectively the multiplicity of the input and output arcs connecting place* p *to transition* t *or vice versa;*
- λ*: is a mapping from T into* $\mathbb{R}$ *that gives the* firing intensities *of the transitions.*
- $\mathbf{m}_0 : P \to \mathbb{N}$*, called the* initial marking *of the net, is the initial state of the net;*

The marking $\mathbf{m}$ of the net, including the initial marking $\mathbf{m}_0$, associates with every place a natural number that corresponds to the number of tokens contained in such a place

In order to define the set of places in input or output to a transition the following shorthand notation was introduced:

1. ${}^\bullet t = \{p \in P : \mathcal{I}(p,t) > 0\}$ is the subset of P containing all the places in input to transition t.
2. $t^\bullet = \{p \in P : \mathcal{O}(p,t) > 0\}$ is the subset of P containing all the places in output to transition t.

All of these features schematically refer to two different SPN aspects: the network **structure** on one side, and its **behavior** on the other.

Static analyses Structural properties of Petri nets are obtained from the incidence matrix, independently of the initial marking. This information can be exploited by approaches based on graph theory, also used for interaction-based models. In addition, two structural properties that can be checked

Static analyses

specifically with Petri Nets are the presence of P semiflows, that refer to the places of the net, and T semiflows, that refer to their transitions.

Given a Petri Net, pre- and post- incidence matrices $\mathcal{I}(p,t)$ and $\mathcal{O}(p,t)$, let $C(p,t)$ be the Incidence Matrix $C(p,t) = \mathcal{O}(p,t) - \mathcal{I}(p,t)$. Each element of C, $c_{p,t}$, thus describes the effect of the firing of transition t on the number of tokens in place p.
In an SPN, a set of places is said to be covered by P semiflow if the weighted sum of their markings does not change with time. Formally, if $x \in Z^{|P|}$ is a place vector; then a P-semiflow is a place vector x such that it represents an integer and non-negative solution of the matrix equation $xC = 0$.
A sequence of transitions, instead, is said to be a T semiflow if, through a sequence of events, it can reproduce at some point in time t the marking of the net that was present at time $t = 0$. Formally, if $y \in Z^{|T|}$ is a transition vector; then a T-semiflow is a transition vector y such that it represents an integer and non-negative solution of the matrix equation $Cy = 0$. As already said, both P and T semiflows are structural properties that can be assessed by the analysis of the incidence matrix C.

Dynamic analyses The net behavior can be obtained starting from the net structure and the initial marking and applying the evolution rules for the marking. An evolution rule defines the preconditions for the occurrence of a transition and the state change produced by such occurrence. Both the preconditions and the state change are encoded in the arcs connected to the transition.

Dynamic analyses

The separation of the net structure from its behavior is reflected directly upon methods of analysis based on the structure of the net and those based on the state space (net behavior). The dynamic behavior of the net is described by means of its *Reachability Graph* (RG) an oriented graph whose nodes are the possible states (or markings) that the system can reach from its initial marking m_0 by applying a transition firing rule. The arcs of the RG represent the transition firing that produce the change state.

Reachability Graph

In order to introduce the RG the following definitions are needed: *transition concession*, *enabled transition*, *transition firing*, *firing sequence* and *reachability set*.

Definition 2 (Concession and enabling in SPN)**.** *A transition* $t \in T$ *has concession to fire in a marking* $\mathbf{m}$ *iff:*

$$\forall p \in {}^\bullet t, \mathbf{m}(p) \geqslant \mathcal{I}(p,t)$$

A transition t *with concession in* $\mathbf{m}$ *becomes enabled.*

Definition 3 (Transition firing)**.** *A transition* t *enabled in marking* $\mathbf{m}$ *can fire and its firing causes a state change from* $\mathbf{m}$ *into* $\mathbf{m}'$ *denoted* $\mathbf{m}[t\rangle\mathbf{m}'$*. The state evolution happens according to the following rule:*

$$\mathbf{m}[t\rangle\mathbf{m}' \Leftrightarrow \mathbf{m}' = \mathbf{m} - \mathcal{I}[t] + \mathcal{O}[t] \wedge t \in \varepsilon(\mathbf{m})$$

This definition can be extended to a transition sequence as follows.

Definition 4 (Transition firing sequence)**.** *A transition firing sequence* σ *is a list of transition firings* $t_1, t_2, \ldots, t_k$*. A transition firing sequence* σ *is enabled in marking* $\mathbf{m}$ *if* $\exists \mathbf{m}_1, \ldots, \mathbf{m}_{m-1}$ *s.t.* $\mathbf{m}[t_1\rangle\mathbf{m}_1[t_2\rangle \ldots \mathbf{m}_{m-1}[t_k\rangle\mathbf{m}_m$

The marking $\mathbf{m}_m$ reached by firing a sequence σ from the marking $\mathbf{m}$ is given by:

$$\mathbf{m} = \mathbf{m}_m - \sum_{i=1}^{k} \mathcal{I}(t_i) + \sum_{i=1}^{k} \mathcal{O}(t_i)$$

We shall denote $|\sigma|_t$ the number of occurrences of a transition $t \in T$ in sequence σ

Now we can define the set of markings that can be reached from the initial marking $\mathbf{m}_0$ by applying the above firing rule.

Definition 5 (Reachability Set (RS))**.** *Let* $\langle P, T, \mathcal{I}, \mathcal{O}, \mathbf{m}_0\rangle$ *be an SPN, its* Reachability Set *(RS) is the smallest set satisfying the following properties:*

- $\mathbf{m}_0 \in$ *RS;*
- $\mathbf{m} \in RS \wedge \mathbf{m}[t\rangle\mathbf{m}' \Rightarrow \mathbf{m}' \in RS$*.*

The RS contains no information about the transition sequences fired to reach each marking. In order to have this information we must introduce the RG. Each node in the RG represents a reachable state, and there is an arc from $\mathbf{m}$ to $\mathbf{m}'$ iff the marking $\mathbf{m}'$ is directly reachable from $\mathbf{m}$. This arc will be labeled with t iff $\mathbf{m}[t\rangle\mathbf{m}'$. Note that one or more arcs can connect two nodes (it is possible for two transitions to be enabled in the same marking and to produce the same state change), so that the RG is actually a multi-graph.

Definition 6 (Reachability Graph (RG))**.** *Let* $\langle P, T, \mathcal{I}, \mathcal{O}, \mathbf{m}_0\rangle$ *be an SPN, its* Reachability Graph *(RG) is a graph RG =(RS, A) where:*

- *RS is the reachability set of the system;*
- $A \subseteq RS \times T \times RS$ *is a set of labeled arcs such that* $(\mathbf{m}, \mathbf{m}', t) \in A$ *iff* $\mathbf{m} \in RS \wedge \mathbf{m}[t\rangle\mathbf{m}'$*.*

It is important to highlight that for each SPN the underlying stochastic process corresponds to a Continuous Time Markov Chain (CTMC) that can be represented as a graph isomorphic to the RG of the net labeling the edges with transition rates

FROM SPN TO ODES It often happens that, in case of very complex models, the underlying CTMC can not be derived or/and solved due to the well-known state space explosion problem. To cope with this difficulty, whenever the stochasticity of the modeled system can be neglected (e.g. due to huge number of molecules), the so-called deterministic approach can be exploited, assuming that the behavior of entities contained in a place of the net is approximated by an Ordinary Differential Equation (ODE) and that the whole model is specified with a system of ODEs, one for each place of the net.

In the literature, different laws (e.g. Michaelis-Menten, Hill-equation, etc.) have been proposed to encode each reaction of the biological system into an ODE. Here we focus on the Mass Action (MA) law [273][1] in which the ODEs describing the model have the following form:

$$\frac{dx_{p_i}(\nu)}{d\nu} = \sum_{j:\mathcal{O}(p_i,t_j)\neq 0} \mathcal{O}(p_i,t_j)\lambda(t_j) \prod_{h:I(p_h,t_j)\neq 0} x_{p_h}(\nu)^{I(p_h,t_j)} - \sum_{j:\mathcal{I}(p_i,t_j)\neq 0} \mathcal{I}(p_i,t_j)\lambda(t_j) \prod_{h:\mathcal{I}(p_h,t_j)\neq 0} x_{p_h}(\nu)^{\mathcal{I}(p_h,t_j)} \tag{11}$$

where $x_{p_i}(\nu)$ represents the amount of the entity in place p_i at time ν assuming that $x_{p_i}(0)$ is defined through the initial marking of the net so that $x_{p_i}(0) = m_0(p_i)$.

For instance, considering the PN model in Fig.9 the behaviour of place *GLC* is described by the following ODE equation assuming the MA law:

$$\frac{dx_{GLC}(\nu)}{d\nu} = +\lambda(K_{F_1}) \cdot x_{HK} \cdot x_{GLC} \cdot x_{ATP} \tag{12}$$

$$-\lambda(K_{R_1}) \cdot x_{HK} \cdot x_{G6P} \cdot x_{ADP} \tag{13}$$

Petri Nets in Systems Biology The use of Petri Nets and computational approaches based on them is not new in systems biology. In the literature in fact can be found various examples of how they are applied to the description of biological systems [63, 40].

OPTIMIZATION PROBLEMS The approach to indetermination we will present in next section requires that we introduce here some concepts related to optimization processes. In Mathematics, Computer Science, and Operations Research, optimization or mathematical programming consists of minimizing (or maximizing) a function by systematically choosing the values of its variables from a set of feasible possibilities properly exploiting analytical or numerical methods. In Systems Biology optimization is not a new concept since it has been already proposed to reconstruct gene regulatory networks, transcriptional regulatory networks, protein interaction networks, conditional specific sub-networks, and active pathways [97], and to perform FBA. Formally an optimization problem with inequality constraints can be defined as follows:

$$\begin{aligned} \underset{\mathbf{y}}{\text{minimize}} \quad & \mathcal{F}_{opt}(\mathbf{y}) \\ \text{subject to} \quad & \mathcal{G}_i(\mathbf{y}) \geqslant b_i, \quad 1 \leqslant i \leqslant l \\ & \mathcal{L}_i(\mathbf{y}) \leqslant c_j, \quad 1 \leqslant j \leqslant m \end{aligned}$$

where the vector $\mathbf{y} = (y_1, \ldots, y_n)$ is the *variable vector*, the function $\mathcal{F}_{opt} : \mathbb{R}^n \to \mathbb{R}$ is the *objective function*, the functions $\mathcal{G}_i(\mathbf{y}) : \mathbb{R}^n \to \mathbb{R}$ and $\mathcal{L}_i(\mathbf{y}) :$

1 Observe that this choice does not affect the generality of our approach that can be applied independently of the assumed law.

$\mathbb{R}^n \to \mathbb{R}$ are *inequality constraint functions*, and the constants $b_1, \ldots, b_l, \quad c_1, \ldots, c_m$ are the *bounds* for the constraints. A vector $\mathbf{y}^\bullet$, called *optimal*, is the solution of the OP if, among all vectors that satisfy the constraints, it is that which yields the smallest (largest) value of the optimization function: $\forall \mathbf{z}$ s.t. $\mathcal{G}_1(\mathbf{z}) \geqslant b_1, \ldots, \mathcal{L}_1(\mathbf{z}) \leqslant c_m$ we have that $\mathcal{F}_{opt}(\mathbf{z}) \geqslant \mathcal{F}_{opt}(\mathbf{y}^\bullet)$.

We recall that an OP is called a *linear program* if the objective and the constraints are linear with respect to the variables and a *non-linear program* otherwise. As shown in the next sections of this paper, we will focus on non-linear programs in which constraints are non-linear as well. To solve this type of OPs, several algorithms have been proposed in the literature, and the reader can find a complete survey of these methods in [1].

Now that we have introduced concepts and definitions related to stochastic processes and Stochastic Petri Nets, we present the three main directions of our work, together with some approaches we developed

1 MODEL INDETERMINATION

In this Section, we present a method to deal with indetermination of mechanism-based kinetic models. As already highlighted in Chapter 2, either we are taking a bottom-up or a top-down approach in order to properly set up a kinetic model we need a model structure and specific values for its kinetic parameters. Defining them is a well-known nontrivial task: small differences in this choice can have huge impacts on model dynamics. Many methods of parameter estimation and reverse engineering have indeed been proposed in literature, and have been reviewed in Chapter 2. Experimental time-course data represent the general requirement for these methods to be effectively applied. At the current moment, modeling efforts are strongly limited by the scarcity of these data. While their availability is growing for in-vitro studies, the same is not true for in-vivo systems like cancer.

In Chapter 2 we analyzed how in common situations uncertainty largely affects multiple components of kinetic models of metabolism. Both uncharacterized isoform mixtures and far-from-standard environmental conditions not only overcomplicate the procedure of parameter estimation but also undermine the accuracy of the estimated parameters. For these and the aforementioned motivation it is often the case that he validity of the calculated parameter values is restricted to the proximity of a specific steady state. If we wanted our model to effectively capture the behavior of the system during large environmental changes, these in practice might be better described with different parameterizations, each valid around of a specific steady state.

Acknowledging that many factors have the potential to shape the behaviour of a metabolic reaction, however, should not induce us to drop any modelling effort. Coherently to authors' statements in the realization that "sloppy" parameters can still yield a predictive kinetic model has opened

the door for 'unusual' modelling strategies in the field. We intend, thus, to proceed, exploiting some specific biological assumptions.

Determined and undetermined reactions

As a *first assumption* for our method, we consider that our knowledge of molecular interactions and kinetic parameters is non uniform across the network, and depends on the biochemical event we are observing. We then decide to divide metabolic reactions, the events of metabolic systems, into two classes: *determined* reactions and *undetermined* reactions.

In the case of determined reactions, all molecular interactions are assumed to be known, and we also assume that a mechanistic, fully-detailed, mathematical description of their activity is available in the literature. Thus, reactions of this class can be clearly characterized and modeled.

For undetermined reactions, instead, we assume that either 1) they belong to portions of the network that have been studied with less detail, so that a fully detailed reaction kinetics has not been reported in the literature, either, 2), that the available kinetics is not adequate to represent the specific actual conditions (e.g isoform mixtures, pH, temperature, metabolite concentrations) our model aims to represent.

We should clarify here that undetermined reactions are not regarded as completely blank elements though: we consider that their stoichiometry is known. While also some activator/inhibitor might have been identified, we make the general assumption that some additional mechanism that we cannot quantify is potentially affecting their behavior. We are not interested to speculate further on the nature of this indetermination: either unidentifiable isoform concentrations, unknown molecular interactions, uncharacterized environmental conditions, or a combination of these factors, we decide to remain agnostic to its specific causes.

More importantly, this assumption makes us state that a fully parametrized mathematical expression that appropriately describes undetermined reactions is lacking. As we described in 1.2 many authors have already tried to circumvent the problem exploiting approximate kinetics. Aware the limitations of these approaches, we still believe their use is legitimated by the absence of adequate alternatives and becomes attractive for their compact mathematical representation, which eases the development of new computational tools. In our experiments, presented in Chapter 4, for reasons of clarity we chose to illustrate the potential of our approach with the simplest formalism, the LMA rate law, used to model both determined and undetermined reactions. To the reader, this probably seems to contradict the definition of determined reactions we just gave. In fact we stated that a fully detailed kinetic characterization is available for determined reactions. This, indeed, holds true. However, the scope of the example here presented is rather explanatory, hence for determined reactions, and just for them, a LMA with defined kinetic parameters is here meant to represent a fully detailed kinetic expression retrieved from the literature. This use of LMA is assumed to represent the behavior of determined reactions without approximations.

On the contrary, due to their undetermined nature, an approximate kinetic formulation has to be forcedly adopted for undetermined reactions, where

it represents their approximate behaviour.
In our method, we try to mitigate the effects of this approximation with the following modeling choice. Just in the case of undetermined reactions, we transfigured the significance of their kinetic parameters: fixed parameter values are replaced by time-varying coefficients. As we stated that many, and potentially important, factors might affect the behavior of undetermined reactions, an approximate kinetic description would hardly be able to recapitulate their behaviors. Time-varying coefficients are then defined as functions that change in time, depending on some component of the system, either known or not characterized.
In this view, these time-varying coefficients are meant to condense the effects of multiple biochemical processes into a single value. With this new perspective, approximate kinetics, like LMA, acquire extended capabilities to represent different kinetic behaviours. In order to avoid confusion, two points should be stressed here. The introduction of time varying coefficients with the resulting conceptual change of kinetic rate laws

1. involves only undetermined reactions;

2. reflects the fact that, with this approach, we intend to focus on the behavior of the whole model rather than on the biochemical meaning of kinetic parameters, which in fact here is lost;

3. even if if the effects of unknown isoform mixtures, pH, temperature or metabolite concentrations could realistically affect also determined reactions, we decide to treat determined reactions as if these factors could be neglected.

We are aware that the assumptions we are posing here can cause non-trivial approximations and can be seen as limiting, but from our perspective they can become empowering as they allow to progressively test different biological hypotheses.

Finally, as anticipated in Section 2, we assume that the behavior of our model is goal-directed. This does not mean that the system under study necessarily is, but, instead, that if an "accurate" objective function is defined, this can be used to formulate an OP and to reproduce the behavior of the original system.

The discussion then shifts to which objective functions can be considered accurate. From our point of view, the objective function can be interpreted and thus defined in different ways. For instance, the objective function can be related to its definition in FBA or cybernetic models and formulated similarly. In FBA, according to its teleonomic perspective, the objective function describes a function that we believe is physiologically relevant for the cell, like the production of biomass to sustain growth. An objective function with this meaning can then be used to predict a physiological behavior of the cell.

In alternative, taking inspiration from other successful applications of FBA, the objective function can express a non-physiological goal of the cell,

like the maximization of ATP production, and can be used to identify some property of the network, or to impose an engineering (e.g. maximization of the production of a particular amino acid) or therapeutic (e.g. minimization of some reaction fluxes vital for cancer cells) goal that we would like the system to reproduce.
More in general, the objective function is here intended to represent any relevant biological behavior that has been observed experimentally or that we would like to artificially recreate in the system. Our approach thus tries to investigate if, and how, this specific biological phenotype can be reproduced in a biochemical system where the topology and the parametrization are just partially known.

In details, we propose here a method that exploits iterative processes of static optimizations: at each simulation step, an objective function with the aforementioned biological meaning allows us to estimate the activities of undetermined reactions, and thus to obtain a complete description of system behavior.

As we anticipated, the kinetic parameters of determined transitions are retrieved from the literature and are kept at fixed values. On the contrary, for each (unidirectional) undetermined transition, one time-varying coefficient is defined.

To facilitate the construction of the model we propose a new graphical formalism based on PN, which allows to automatically translate the model into its mathematical representation, consisting of the ODEs system and the Optimization Problem (OP).

According to this we decide to present our approach firstly introducing this new graphical formalism, and then providing its automatic translation into a ODE system in which indeterminate transitions are tackled through an OP. For this purpose we use the model of Fig.9 as a “running example" that we comment in the rest of the paper to discuss the features of this new modeling formalism.

SPN with Indetermination. The formal definition of a new PN extension called Stochastic Petri Net with Indetermination (SPNI) is the following:

Definition 7. *A stochastic Petri net with indetermination is a tuple*

$$\mathcal{N} = (P, T, \mathcal{I}, \mathcal{O}, \mathbf{m}_0, \lambda_n, \Lambda_u, \mathcal{F}^{\mathcal{N}}_{opt})$$

where:

- $T = T_n \cup T_u$ *is a finite, non-empty set of* timed transitions *with* $T_n \cap T_u = \emptyset$. T_n *is the set of* determined transitions, *while* T_u *is the set of* undetermined transitions.
- $\lambda_n : T_n \to \mathbb{R}$ *gives the* firing intensity *of* T_n *transitions.*
- $\Lambda_u : T_u \to \mathbb{R}^2$ *gives the range of variation of the flux of* T_u *transitions.*

- $\mathcal{F}^{\mathcal{N}}_{opt} : T \times P \to \mathbb{R}$ *is an* objective function *whose terms are represented by place markings and transition firing intensities.*

We use the notation $\Lambda^{L}_{u}(t)$ (resp. $\Lambda^{U}_{u}(t)$) to denote the lower (resp. upper) bound values of the interval in which the flux of a $t \in T_u$ can vary; $\Lambda_u(t)$ then represents a possible flux value of the (undetermined) transition t compatible with its specified lower and upper bounds.

From SPNI to ODE and OP.
Due to the indetermination associated with the T_u transitions, it is not possible to directly use Eq. 11 to represent the deterministic behavior of an SPNI model. We can however re-write Eq. 11 as follows:

$$\begin{aligned}\frac{dx_{p_i}(\nu)}{d\nu} = &\sum_{j:\mathcal{O}(p_i,t_j)\neq 0} \mathcal{O}(p_i,t_j)\mathcal{M}_{t_j}(\nu) \prod_{h:\mathcal{I}(p_h,t_j)\neq 0} x_{p_h}(\nu)^{\mathcal{I}(p_h,t_j)} \\ &- \sum_{j:\mathcal{I}(p_i,t_j)\neq 0} \mathcal{I}(p_i,t_j)\mathcal{M}_{t_j}(\nu) \prod_{h:\mathcal{I}(p_h,t_j)\neq 0} x_{p_h}(\nu)^{\mathcal{I}(p_h,t_j)}\end{aligned} \tag{14}$$

where $\mathcal{M}$ is a function defined in the following way:

$$\mathcal{M}_t(\nu) = \begin{cases} \lambda_n(t) & \text{if } t \in T_n \\ y_t(\nu) & \text{otherwise} \end{cases} \tag{15}$$

The parameter $y_t(\nu)$ encodes the indetermination associated with the undetermined transition t at time ν and must be properly estimated to solve the ODE system.

Independently of the context of the modeling experiment, it is usually the case that we want to minimize (or maximize) certain measures defined on the portion of the state of the system that is not directly affected by undetermined transitions. These measures, that may assume complex definitions, become the optimization functions that we use to reproduce the model behavior.

To cope with this problem we thus propose to exploit an optimization process in which the objective function depends on the solution of the ODE systems in Eq.14. In practice, the optimization process solves the ODE system for a specific time interval while, simultaneously, it uses the obtained solution to compute the objective function of the optimization problem. The maximum/minimum value of the objective function allows to identify the values of unknown parameters of undetermined reactions.

Given an SPNI model, the corresponding OP, whose solution will be used to estimate the firing intensity values of the T_us, is derived using the following definition.

Definition 8. *The OP derived by the SPNI is a tuple*

$$Opt = (\mathbf{y}_\nu, \mathcal{F}_{opt}, \mathcal{G}, \mathcal{L})$$

where:

- $\mathbf{y}_\nu$ *represents the optimizing values of undetermined transitions at time* ν*, i.e.* $\forall t \in T_u \Rightarrow y_t(\nu) \in \mathbf{y}_\nu$*;*

- $\mathcal{F}_{opt} = \mathcal{F}^{\mathcal{N}}_{opt}$;
- $\mathcal{G}$ *is defined by*

$$\forall t \in T_u \Rightarrow y_t(\nu) \prod_{h:\mathcal{I}(p_h,t)\neq 0} x_{p_h}(\nu)^{\mathcal{I}(p_h,t)} \geqslant \Lambda_u^L(t)$$

- $\mathcal{L}$ *is defined by*

$$\forall t \in T_u \Rightarrow y_t(\nu) \prod_{h:\mathcal{I}(p_h,t_j)\neq 0} x_{p_h}(\nu)^{\mathcal{I}(p_h,t_j)} \leqslant \Lambda_u^U(t)$$

For instance, considering the SPNI in Fig. 9, where the gray boxes highlight the transitions affected by indetermination, the vector $\mathbf{y}(\nu)$ has size six and represents the optimal values of the firing rates of transitions T_{uf1}, T_{ur1}, T_{uf3}, T_{ur3}, T_{uf12}, T_{ur12}. An example of objective function could be the maximization of the Lactate (LAC) as described in our case study in the Chapter 4.

Moreover in our example
$\Lambda_u^L(T_{u_{f_1}}) = 1620 \cdot x_{HK} \cdot x_{GLC} \cdot x_{ATP}$
$\Lambda_u^U(T_{u_{f_1}}) = 2.592e+08 \cdot x_{HK} \cdot x_{GLC} \cdot x_{ATP}$,
with limit values chosen as explained in the Chapter 4.

How to compute the model behavior. Let $\mathbf{x}(\nu)$ represent the behaviour of the model at time ν The numerical integration of Eq. 14 provides the behaviour of the model at time $\nu + \tau$, in terms of the behaviour $\mathbf{x}(\nu)$ computed at time ν and of a set of parameters deriving from the structure of the SPNI ($\mathcal{I}$, and $\mathcal{O}$), the firing intensities of the definite transitions of the net (λ_n) and of the firing intensities of the undetermined transitions estimated at time ν and collectively represented as $y(\nu)$. The values of $\mathbf{x}(\nu + \tau)$ are thus the results of the evaluation of a function whose input parameters are represented by a tuple $B(\nu) = (B, B_u(\nu))$ where $B = (\mathcal{I}, \mathcal{O}, \lambda_n)$ and $B_u(\nu) = (\mathbf{x}(\nu), \mathbf{y}(\nu))$ [2]. The integration step s identifies the time points $\nu_i = i * \tau$ where the evaluation of the model behaviour is of interest.

Role of the estimation phase of our method is to find a set $\mathbf{y}(\nu)$ that, being compatible with the constraints of the SPNI model (Λ_u), minimizes the objective function at time $\nu + s$. The optimization phase identifies a number K of initial conditions, that we denote with $B_u^{[k]}(\nu), k = 1, ..., K$, consisting of the behaviour of the model computed at time ν and of K random points within the space of firing intensities of the undetermined transitions identified by the constraints Λ. From each of these configurations the method numerically integrates the system of ODEs up to time $\nu + s$ to derive $\mathbf{y}(\nu)$. Letting $B^{[k]}(\nu) = \left(B, B_u^{[k]}(\nu)\right)$, with $B_u^{[k]}(\nu) = (\mathbf{x}(\nu), \mathbf{y}^{[k]}(\nu))$, the solutions obtained from the integration of the ODEs with parameters $B^{[k]}(\nu), k = 0, 1, ..., K$ and up to time $\nu + s$ are compared to identify the

2 In the sequel of the paper we will indifferently use $\mathbf{y}_t(\nu)$ or $\Lambda_u(t, \nu)$ to represent the undetermined parameters of our models as provided by the optimization problem at time ν.

Algorithm 1 Algorithm to solve ODE system with Indetermination

```
1: function SolveODEI(ODEI,G,L,F_opt,y_t,τ,FinalTime)
2:     ν = 0.0;
3:     ODEI.Init(Value);
4:     while (ν ⩽ FinalTime) do
5:         print(ν,Value);
6:         Res=SolveOpt(y_t, Value, ODEI, ν + τ, G, L, F_opt);
7:         Value=Res.Value;
8:         y_t=Res.y_t;
9:         ν += τ;
```

Algorithm 2 Algorithm to solve ODE system with Indetermination

```
1: function SolveODEI(ODEI, G, L, F_opt, y_t, τ, FinalTime)
2:     ν = 0.0;
3:     ODEI.Init(Value);
4:     while (ν ⩽ FinalTime) do
5:         print(ν,Value);
6:         if Heurist(Value,time) then
7:             Res=SolveOpt(y_t, Value, ODEI, ν + τ, G, L, F_opt);
8:             Value=Res.Value;
9:             y_t=Res.y_t;
10:        else
11:            Value=ODE.SolverODE(Value, ν + τ, Rate_{T_u});
12:        ν += τ;
```

choice of $B^{[k]}(\nu)$ which provides the best evaluation of the objective function, thus identifying $\mathbf{y}^{[k]}(\nu + s) = \mathbf{y}(\nu)$. Crucial in this optimization step is that the numerical integration of the ODEs is performed with a method capable of identifying an integration step h small enough to allow a precise solution of the ODEs during these "tentative" evaluations that are used to select the firing intensities of T_us.

In general, this whole method is repeated for each time point ν_i starting from $\nu_0 = 0$. However, solving the OP for each value of ν_i can be excessively costly and we can thus reduce this computational effort by identifying a time interval ρ that is a multiple of τ and that determines the time points where the optimization is requested. By doing so, if we set $\rho = m \cdot \tau$, we assume that for $m - 1$ intermediate evaluation steps the values of $\Lambda_u(\nu)$ (i.e. $\mathbf{y}(\nu)$) remain constant and an approximation is introduced.

Having discussed how to derive from an SPNI model (i) an ODE system with indetermination (see Eq.14), and (ii) an OP (see Def.8), we can devise an algorithm which combines them to derive the model behaviour.

The pseudo-code of this algorithm is shown in Alg. 1. It takes as input the ODE system with indetermination (i.e. *ODEI*), the OP (i.e. described by functions $\mathcal{G}$, $\mathcal{L}$ and $\mathcal{F}_{opt}$), the initial guess for the rate of undetermined

transition (i.e. y_t), the step size used for the optimization schema (i.e. τ), and the final time (i.e. *FinalTime*). The output of the algorithm is represented by the values generated for each system entity at different time points (i.e. v_i). In details, the method *Init()* at line 3 initializes the vector *Value* encoding the initial values assumed for all the entities of the model. Then, the code from line 7 to line 12 is repeated until the time horizon is reached. In each iteration the function *print()* is called to print the current values of the system entities. Subsequently, the function *SolveOpt()* solves the optimization and returns the new values of the system entities and of the rates of T_us (i.e. $Res.Value$ and $Res.y_t$ respectively). It takes as input an initial guess for the rate of T_us (i.e. y_t), the current values of the entities (i.e. *Value*), the final time in which the objective function will be evaluated (i.e. $v + \tau$), the ODE system (i.e. *ODE*), and a set of functions encoding the OP (i.e. $\mathcal{G}, \mathcal{L}$ and $\mathcal{F}_{opt}$). The functions $\mathcal{G}$ and $\mathcal{L}$ are used by the optimization solver to test if a new vector $\mathbf{y}$, randomly generated according to the parameter constraints, is a feasible solution. Indeed the functions $\mathcal{G}$ and $\mathcal{L}$ verify if $\mathbf{y}$ satisfies the inequality constraints. The function $\mathcal{F}_{opt}$ is instead called by the optimization solver to compute the value of the objective function associated with a feasible vector $\mathbf{y}$.

This function, takes as input the vector $\mathbf{y}$, the current values of the entities (i.e. *Value*), the ODE system (i.e. *ODE*), and the final time in which the objective function must be evaluated (i.e. $v + \tau$). First it computes the quantities values at $v + \tau$ assuming the missing rates to be equal to $\mathbf{y}$. Then, the computed values are used to evaluate the objective function, whose derived value is returned. When the optimization step is terminated the vector *Value* is updated with the new computed values.

Moreover, in Alg. 2 we report an extension of the previous pseudo-code in which the optimization solver could be executed less frequently, so not at each time step. Indeed, we propose to exploit a heuristic function to decide when the optimization phase must be performed. Hence, when the optimization solver is not called the previous value y_t are considered during the solution of ODE system (i.e. method *SolveODE()*). In Chapter 4 an example of such a heuristic function is discussed and some experimental results are presented.

2 DATA INTEGRATION

Our method aims to use high-throughput gene expression data produced with RNA-Seq technologies to transform kinetic models of metabolism, already built and validated for a specific condition, into new models that describe new conditions of interest. The transformation is intended to adjust the value of some model parameters depending on the expression of the metabolic enzymes in the specific condition.

The level of expression of metabolic enzymes has in fact a high impact on the velocity of the reactions in the system. For all reactions catalyzed by enzymes, either modeled with fully detailed or approximate kinetics, the influence of enzyme concentration on reaction kinetics is captured by some of the kinetic parameters that appear in the rate law. In the case of LMA and GMA, for instance, the influence of enzyme concentrations on the reaction flux is modeled implicitly. Variations in enzyme concentrations, in fact, would impact the values of the forward and reverse reaction rate constants. In Michaelis-Menten, as well as in the vast majority of fully detailed reaction rate laws, instead, a v_{max}, forward or reverse, parameter that defines the maximal catalytic activity an enzyme can reach, is always present. As it is also highlighted in figure 3, the value of v_{max} can be linked to the values of its multiplication factors, E_{tot} and k_{cat}, by the relationship $v_{max} = E_{tot} \cdot k_{cat}$.
k_{cat}, or turnover number, consists of the same parameter described as k_2 in reaction 2. k_{cat} determines the catalytic efficiency of the enzyme as it represents the rate of the limiting elementary step in the reaction.

In order to test our approach of data integration, we assume that the differences in the estimated values of v_{max} (or, equivalently, in k_f and k_r for LMA and GMA rate laws) across different conditions, like different cells or different tissues, are uniquely caused by differences in enzyme concentrations. k_{cat} is then assumed to hold a constant value, specific of each enzyme, in all the conditions we are considering. With this assumption v_{max} can be directly calculated knowing the level of enzyme present in the system.

Estimates of enzyme concentrations are normally obtained with western blots and proteomics technologies. These analyses, however, require labor-intensive preprocessing procedures and high costs. For these reasons, these types of data are publicly available in limited amounts. Gene expression data, on the other hand, is currently a much more readily available source of information.

In our method, we explored how enzyme concentrations can be inferred with a simple procedure that exploits RNA-Seq gene expression data. In specific, we use RNA-Seq outputs, namely the Fragments Per Kilobase Mapped Reads (FPKMs), to recast some model parameters so that the model reflects the gene expression profile of the condition of interest.
With respect to this goal, it is important to highlight that the amount of mRNA of an enzyme-coding gene should be used with caution. First of all, the current understanding of transcription and epigenetics teaches us that the proportion of transcript that is converted into a functional enzymatic form is the result of many processes of post-transcriptional and post-translational modification [304]. The impact of these can be potentially very high, as reviewed in [115]. Also, these mechanisms are still far from being completely understood, and almost impossible to model mathematically.

Aware of these extremely complex events of regulation, we anyway believe that the utility of simplistic methods of data integration should be explored.

We give here an example of our recasting approach. In the case of v_{max}, the parameters for a new, desired, condition N can be calculated the known v_{max} values of a reference condition R

$$v_{max^N} = v_{max^R} \cdot \mathrm{FPKM}^N / \mathrm{FPKM}^R$$

.
In the case of LMA, instead

$$k_f^N = k_f^R \cdot \frac{\mathrm{FPKM}^N}{\mathrm{FPKM}^R}$$

$$k_r^N = k_r^R \cdot \frac{\mathrm{FPKM}^N}{\mathrm{FPKM}^R}$$

In this way, a model that is fully parameterized for a reference condition can be recast into a new model, representative for a new condition of interest, if RNA-Seq data of the two conditions are available.

It should be noticed that when we make use of these simple transformations we do not necessarily need to postulate that all post-transcriptional processes have a null effect on enzyme concentrations. More precisely instead, we are proposing that epigenetic mechanisms affecting the final concentration of metabolic enzymes are consistently and equally present across the conditions we are studying. In fact, if in two different conditions some post-transcriptional modifications have a high but equal impact on enzyme abundance (we can envision, for instance, that a very low proportion of mRNA is effectively converted into a functional protein) and we know from the existing literature the accurate v_{max} value for one of the two conditions, then the computed v_{max} value for the new condition would be accurate as well.
We remark that, in order to limit additional sources of uncertainty, the level of mRNA should be better measured with RNA-Seq technologies rather than with microarrays, as the former provides a more accurate absolute quantification of its abundance.

RNA-Seq data, however, are not the only type of data that can be used for this approach. For those situations in which proteomics data are available, in fact, these can be used to improve the accuracy of the recasting process. The type of data used does not alter the procedure of data integration, simply FPKM values should be replaced by protein concentrations. A complete discussion on the potential insights that can be gained applying this approach in the context of kinetic models of cancer metabolism, supported by some figures and data, is postponed to chapter 4.

3 REPRESENTING METABOLIC HETEROGENEITY

A metabolic system can present features of metabolic heterogeneity at different levels. In this thesis, we will focus more specifically on two forms of heterogeneity: heterogeneous kinetic properties of different enzyme isoforms and the heterogeneous metabolic traits that can be displayed by different cells in a population. We will see here how an extension of SPN, namely Stochastic Symmetric Nets, has peculiar properties that facilitate the representation of the metabolic heterogeneity present in the system.

3.1 Stochastic Symmetric Nets

We have seen how SPNs have the capability to provide a high level representation of a stochastic process, a CTMC specifically, in a graphical and compact formulation, from which a system of ODEs can be directly derived. As for many biological systems, however, it can be the case that an SPN representation is still highly complex and hard to represent graphically without confusion. In these cases, a modeler can exploit repeated and modular structures that might be present in the network, which are seen as *symmetries* in an SPN, to create a much more concise representation of the system. *Stochastic Symmetric Petri Nets* (SSN) are the formalism that permits such a compact model description.

One important new feature of SSNs is the possibility of having distinguished tokens, so that the tokens could be represented graphically as dots of different color [3]: in practice the "color" attached to them may be any kind of information. The type of the information associated with tokens can differ depending on the place where they are located, hence the definition of an SSN must include the definition of a *color domain* for each place (denoted $cd(p), p \in P$) that specifies the type of data attached to the tokens in that place.

The advantage of this new feature can be better understood considering the following biochemical example. Figure 33 shows the elementary step of a simple reaction in which a substrate is converted to a product in the presence of a competitive inhibitor. We can imagine that one or more isoforms of this enzyme are actually present in the system, and that these isoforms display a slightly different functionalities. It could be, for instance, that one of the isoforms is insensible to the interaction with the inhibitor, or that the catalytic step occurs at different rates. In such situations it becomes important to model how many molecules of substrate interact with one isozyme or with the other, as if the tokens contained in places E, ES, EI, EP could be separated depending on the specific isoform proportions. If we maintain the separation, our model is able to represent what actually happens in the system, i.e. the substrate molecules are channeled into separate, parallel routes, each characterized by its own kinetic properties. In this situation a color class, as defined in 10, can be associated to all the enzymatic forms,

3 this explains the adjective colored

bounded or unbounded. The resulting SSN representation of the system, later reported in figure 33, is able to retain all the information in the model in a much more compact format.

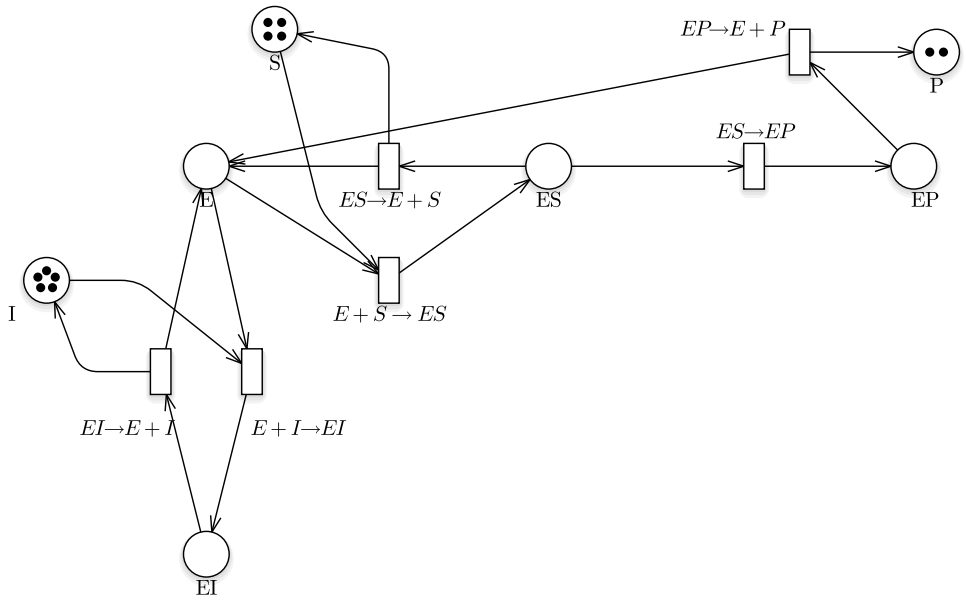

Figure 5: SPN representation of the elementary steps of the conversion of substrate S into product P in the presence of a competitive inhibitor I

The state, or marking, in SSNs is represented by the multiset of colored tokens associated with each place. As in SPN systems the state change is performed by transition firing. Since the tokens in SSNs are distinguished, some additional information is needed to define the colored tokens that are withdrawn from the input places and put into the output places of a given transition when it fires. Hence the transitions can be considered to be procedures with parameters. The possible values of these parameters define the so called *transition color domain*. The set of parameter is defined through the arc function connecting the transition to input/output places.

In order to enable and fire a transition, it is necessary to instantiate the actual values for its parameters. A transition whose parameters have been instantiated to actual values is called *transition instance*.

If the number of possible colors is finite then an SSN has the same theoretical modeling power as an SPN and it is always possible to derive an equivalent SPN applying an unfolding algorithm[4].

Modeling complex systems, like biochemical systems for instance, with SSN is more convenient, not only for their compactness and readability but also for their significantly higher degree of parameterization. Model symmetries can in fact be automatically exploited to generate a lumped CMCT

4 It is important to observe that while the unfolding of an SSN is unique, the inverse operation of folding an SPN to obtain a more compact, colored representation of the same model, may lead to several alternative SSN models, depending on the point of view of the folding and the desired degree of compacting.

from which it is possible to compute the same indexes computed for a non-lumped CTMC [25, 59, 70].

Before giving a formal definition of an SSN it is necessary to introduce the following concepts: *multiset*, *basic color class*, *color domain*, *standard predicate*, *elementary function*, *class function* and *arc function*.

Definition 9 (Multiset). *A multiset* a *over a nonempty set* A *is a mapping* $a \in \{A \to \mathbb{N}\}$, *we use the notation* $Bag(A)$ *to denote a multiset over* A. *Intuitively, a multiset is a set that can contain several occurrences of the same element. It can be represented by a formal sum:* $a = \sum_{x \in A}(a(x))x$. *The coefficient* $a(x)$ *is called multiplicity of* x *in* a.

Observe that the elements with multiplicity zero are omitted in the formal sum representation

Definition 10 (Basic Color Class). *A basic color class, denoted* C_i, *is a finite not empty set of colors identifying objects of the same nature.*

It is called ordered basic color class if a successor function (denoted $!$*) induces a circular ordering on its elements.*

Moreover a basic color class can be partitioned into n_i *disjoint subsets* $C_{i,j}$ *with* $j = 1, \ldots, n_i$ *called* static subclasses*; colors belonging to different static subclasses represent objects of the same type but with different behavior.*

We will denote $\mathcal{C} = \{C_1, \ldots, C_h, \ldots, C_n\}$ *the set of pairwise disjoint basic color classes. We use also the convention that classes with index up to* h *are not ordered, while classes with higher index are ordered.*

Definition 11 (Color Domain). *The information associated with tokens comprises one or more fields, each field in turn has a type selected from the set of basic color classes* $\mathcal{C}$.

The transition color domains are used to define the parameters of transitions and their type. Each parameter has a type selected from the basic color classes. In addition, the possible color instances of a transition can be restricted by means of a guard, *a* standard predicate *which will be defined below. Hence the definition of a transition color domain comprises a list of typed parameters.*

Each parameter is associated with a variable appearing in some arc function of the input, output or inhibitor arcs of the transition. We shall denote $var_i(t)$ *the subset of transition* t *parameters of type* C_i, *and* $var(t)$ *the whole set of transition* t *parameters. In practice this approach of using the variable-based notation makes the model more readable.*

Definition 12 (Standard Predicates and Guards). *A guard is a boolean expression defined on a transition color domain. In an SSN it is expressed by a* standard predicate, *that is a boolean expression whose basic terms are* basic predicates. *Basic predicates allow to compare color elements from the same color class* C_i *or test whether a color element belong to a given static subclass, and can take the following form:*

- $[X_i^j = X_i^k](c)$, *it evaluates to true iff the* j^{th} *component of type* C_i *in* c *is equal to the* k^{th} *component of the same type;*
- $[d(X_i^j) = C_{i,h})](c)$, *it evaluates to true iff the* j^{th} *component of type* C_i *in* c *belong*

to the static subclass $C_{i,h}$*;*
- $[d(X_i^j) = d(X_i^k)](c)$*, it evaluates to true iff the* j^{th} *and* k^{th} *components of type* C_i *in* c *belong to the same static subclass.*

Before introducing the arc functions, it is necessary to define the two following concepts: *elementary function* and *class function*.

Definition 13 (Elementary Function)**.** *An elementary function is a linear mapping from* $cd(t)$ *to* $Bag(C_i)$ *(for some* $i \in 1, \ldots, n$*) chosen among the following functions:*

- *the* projection function *denoted* X_i^l *defined as:*

$$X_i^l(\ldots, c_1^{j_1}, \ldots, c_2^{j_2}, \ldots, c_n^{j_n}, \ldots) \mapsto c_i^l$$

- *the* successor function *denoted* $!X_i^l$ *defined as:*

$$!X_i^l(\ldots, c_1^{j_1}, \ldots, c_2^{j_2}, \ldots, c_n^{j_n} \ldots) \mapsto !c_i^l$$

- *the* diffusion function *(also called synchronization function, depending whether it annotates an output or an input arc), which is constant, denoted* S_i *and defined as follows:*

$$S_i(\ldots, c_1^{j_1}, \ldots, c_2^{j_2}, \ldots, c_n^{j_n} \ldots) \mapsto C_i$$

  Notice that in practice the symbols X_i^l *used above to denote the projection function are substituted by names of transition variables (representing the transition parameters) in the models; each variable has a type* C_i*. The variable-based notation usually makes the model more readable.*
  The diffusion (synchronization) function can be restricted to a static subclass, denoted $S_{i,j}$ *and defined as follows:*

$$S_{i,l}(\ldots, c_1^{j_1}, \ldots, c_2^{j_2}, \ldots, c_n^{j_n} \ldots) \mapsto C_{i,l}$$

Definition 14 (Class Function)**.** *A color function* f *on class* C_i*, also called* C_i *class function, is a linear combination of elementary functions (with same domain and codomain):*

$$f_i = \sum_j \alpha_j \cdot X_i^j + \sum_{q=1}^{|\widehat{C_i}|} \beta_q \cdot S_{i,q} + \sum_j \gamma_j \cdot !X_i^j$$

The coefficients $\beta_q \in \mathbb{N}$*,* $\alpha_j, \gamma_j \in \mathbb{Z}$ *must satisfy the following constraint: if* f_i^- *and* f_i^+ *are respectively the multisets of elements with negative and positive coefficients in the formula above (so that* $f_i = f_i^+ - f_i^-$*), then it must hold* $f_i^- \leqslant f_i^+$*.*

Definition 15 (Arc Function)**.** *An arc function is a weighted (and possibly guarded) sums of tuples, the elements composing the tuples are in turn weighted sum of class functions.*

$$F = \sum_k \lambda_k \cdot [pred_k] \bigotimes_{i=1}^{n} \bigotimes_{j=1}^{e_i} f_i^{j,k}$$

where $f^k_{i,j}$ *is class function,* $\lambda_k \in \mathbb{N}$, $[pred_k]$ *is a standard predicate and* e_i *is the number of occurrences of class* C_i *in* $cd(p)$. *The symbol* $\bigotimes$ *denotes the Cartesian product, in the text we shall also use the representation* $\langle f^1_1, f^2_1, \dots f^{e_n}_n \rangle$, *briefly called function tuple (or simply tuple).*

The formal definition of an SSN is the following

Definition 16 (SSN). *A Stochastic Symmetric Petri net is a tuple:*

$$\mathcal{N}_{SSN} = \langle P, T, \mathcal{C}, \mathcal{I}, \mathcal{O}, cd, \phi, \lambda, \mathbf{m}_0 \rangle$$

where:

- P,T *are defined as for SPN;*
- $\mathcal{C}$ *is a finite set of finite color classes,*
- $\mathcal{I}(p,t), \mathcal{O}(p,t) : cd(t) \to Bag(cd(p))$ *are the pre- and post- incidence matrices associating a function with each arc;*
- $cd : P \cup T \to \mathcal{C}$ *is a function defining the color domain of each place and transition;*
- ϕ *is the vector of guard functions and maps each element of* T *into a function assigning to each color* $cd(t)$ *a value in* $\{false, true\}$;
- λ *is a set of functions* $\lambda_t : cd(t) \to \mathbb{R}$ *expressed in the following form:*

  $\lambda(t)=$
  {
  case $cond_1$ *:* r_1
  case $cond_2$ *:* r_2
  $\dots$
  case $cond_n$ *:* r_n
  default*:* $r_{default}$
  }

- $\mathbf{m}_0 : P \to Bag(cd(p))$ *is the initial marking, mapping each place* p *on a multiset on* $cd(p)$

The evolution of an SSN system is defined through a firing rule, but in this case the firing concerns a transition instance rather than a transition.

Definition 17 (Transition instance). *A transition instance is an assignment of actual values to the parameters of the transition. We use the notation* $\langle t, c \rangle$ *for an instance of transition* t, *where* $c \in cd(t)$ *represents the assignment of actual values to the transition parameters.*

Let us define the concession and the enabling of transition instance and their firing.

Definition 18 (Concession and enabling of transition instance)**.** *A transition instance* $\langle t, c\rangle$ *has concession in* $\mathbf{m}$ *iff:*

$$\forall p \in P, \mathcal{I}(p,t)(c) \leqslant \mathbf{m}(p) \ \wedge \phi(t)(c) = true$$

A transition instance $\langle t, c\rangle$ *is enabled in marking* $\mathbf{m}$ *iff* $\langle t, c\rangle$ *has concession in* $\mathbf{m}$*.*

Definition 19 (Transition Instance Firing)**.** *An enabled transition instance* $\langle t, c\rangle$ *may fire, and its firing leads to a new marking* $\mathbf{m}'$*:*

$$\forall p \in P, \mathbf{m}'(p) = \mathbf{m}(p) + \mathcal{O}(p,t)(c) - \mathcal{I}(p,t)(c)$$

The firing is denoted $\mathbf{m}[t, c\rangle\mathbf{m}'$ *or* $\mathbf{m} \overset{\langle t,c\rangle}{\rightarrow} \mathbf{m}'$*.*

As we described in chapter 2 Petri Nets can be effectively used to describe a system of metabolic reactions, both for its structural properties and its dynamic behavior. In this section we are presenting how heterogeneity at different levels can be represented with Symmetric Stochastic Petri Nets, and the notions of color domains.

HETEROGENEITY OF ENZYME ISOFORMS Enzyme isoforms, or isozymes, can be defined as highly similar gene products that perform essentially the same biological function. Due to these slight differences, isozymes convert substrates into products with slightly different kinetics. To better present our approach, we will temporarily restrict our focus to consider a system composed of a single enzyme catalyzed reaction, like the one we presented above, depicted in figure 33. The figure shows a Petri net that represents the elementary steps of a simple reaction mechanism, in which an enzyme, a substrate, a product and a competitive inhibitor participate. If we wanted to expand our model to take into account the fact that in our system additional isozymes are present, our reaction network would expand significantly. In the simple case in which we have a mixture of just two enzyme isoforms, the two different enzyme species, together with the reactions they independently catalyze, would increase both the rows and the columns of the stoichiometric matrix associated to the system. A graphical representation with an SPN would equally expand, almost doubling in size. A visual example of how a model with two isozymes would appear is reported in figure 6

It should become evident that the size of the SPN would increase as many times as the number of different isoforms we want to include in the model. If we further wanted to describe systems of more that one reactions the representation would become rapidly incomprehensible. In order to describe larger systems where multiple isozymes are present while limiting modeling difficulties associated to network reconstruction, visualization and model simulation, we propose that SSN should be more effectively employed. In figure 7 we report an SSN representation of the same system in figure 33, in which 3 enzyme isoforms are specified. This SSN has one color class $Et = \{A, B, C\}$ divided into two static subclasses $EtA = \{A\}, EtB = \{B, C\}$. The

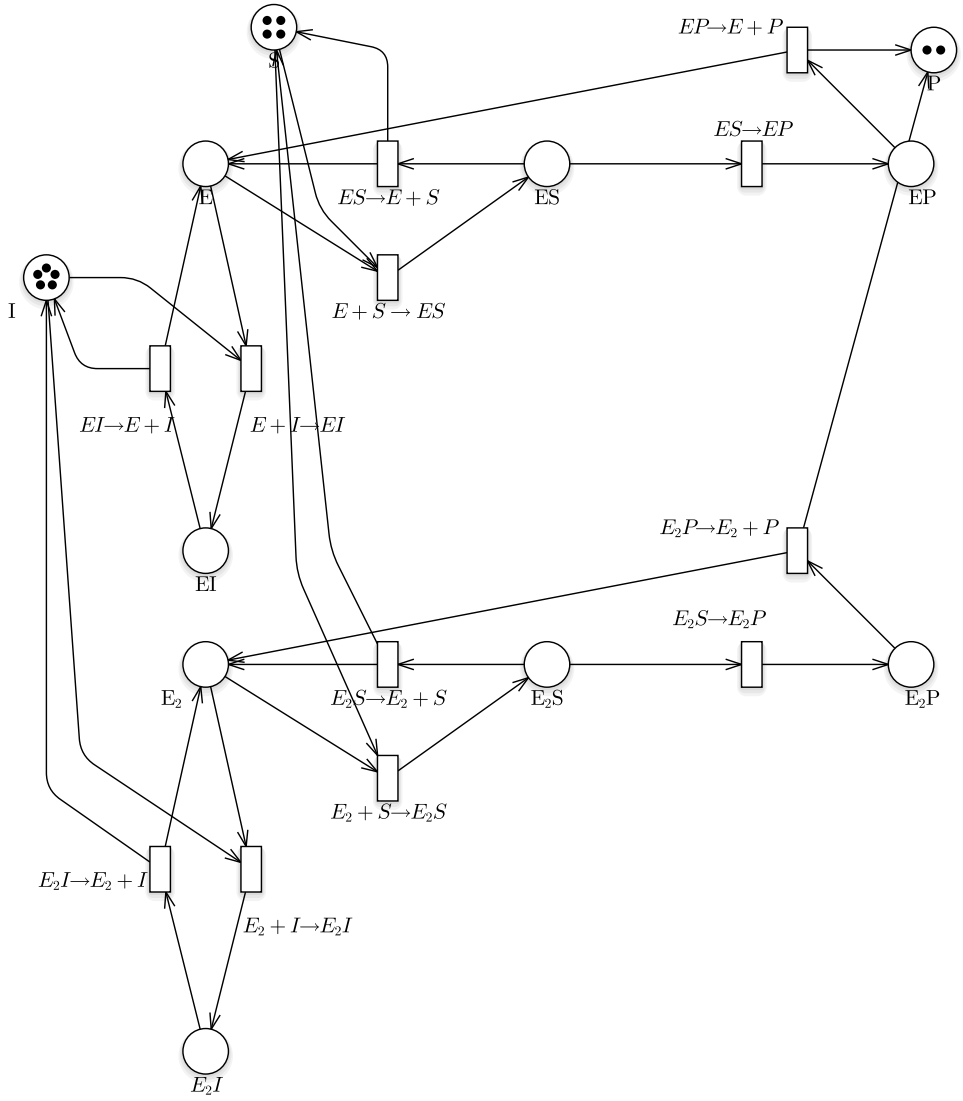

Figure 6: SPN representation of a system analogous to the one in figure 33, in which two enzyme isoforms are described

color domain of all the places is Et and represents the enzyme in its bounded and unbounded forms. The resulting SSN model, much more elegant in its appearance, is able to retain the same exact information on the structural and behavioral properties of the original model.

For this explanatory model, for instance, heterogeneity was introduced in the model as follows:

- The initial marking of place E_0, $m_0(E) = 2\langle A\rangle + \langle B\rangle + 3\langle C\rangle$ reflecting the relative isozymes abundances as 2 color A tokens, 1 color B token, and 3 color C tokens
- All the arcs are labeled with the projection function $\langle x\rangle$
- the label $[d(x) = EtA]$ associated with the transition $E + I \to EI$, represents a guard which defines that just isozymes of color A can interact

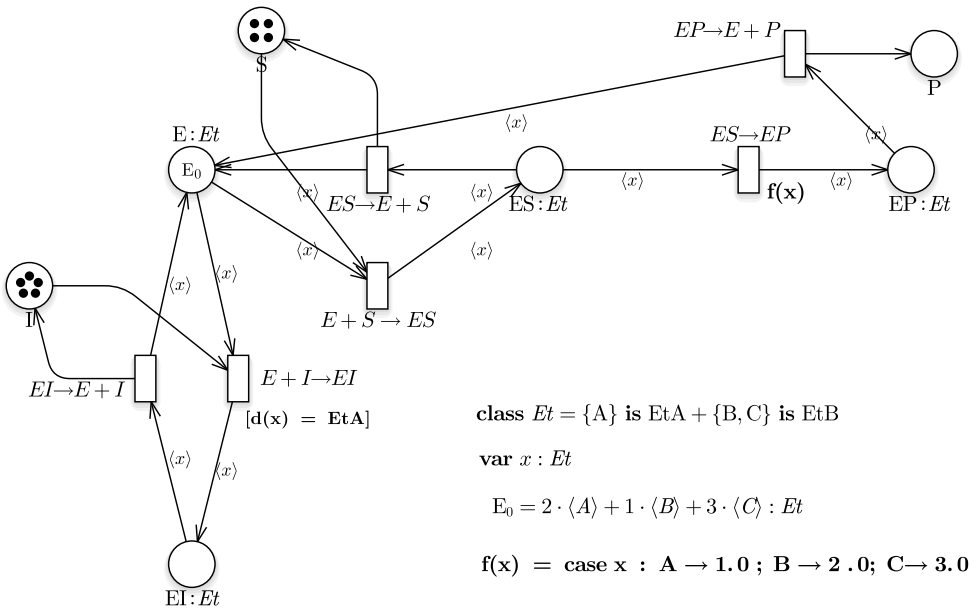

Figure 7: SSN representation equivalent to the SPN representation in in figure 6

with the inhibitor, while the other two remain unaffected by this inhibition.

- The kinetic properties of the isozymes differ with respect to the k_{cat}. This kinetic heterogeneity can be easily encoded in an SSN specifying that three different firing rates $f(x)$ are associated with the isoforms.

INTRAPOPULATION METABOLIC HETEROGENEITY The high versatility of SSNs can also be shown in the context of metabolic heterogeneity in a population of cells that share the same environment and the same energetic substrates. This multicellular system can be envisioned as a portion of a tissue, a cancer, or a bacterial community. A population in which different metabolic traits are present would be normally modeled with an equally numerous population of interacting models. It is important to notice that a model of a population of cells cannot be reduced to multiple, separated sub-models, each for one single cell. A similar transformation would miss the key aspect that these metabolic models compete for the same substrates available in the environment. Each single cell sub-model, thus, should always "sense" the presence of the other sub-models. If these sub-models are characterized by a high similarity (e.g they just differ for the values of their kinetic parameters), they can be described with a single SSN, where the colors reflect the differences (e.g. different parametrizations) among the sub-models in the population. In chapter 4, we will show how an ensemble of single cell sub-models describing the different metabolic phenotypes identified in a sample of cancerous tissue can be represented and modeled with the support of the SSN formalism.

4 EXPERIMENTAL RESULTS AND DISCUSSION

1 MODEL INDETERMINATION

Our implementation. To perform our experiments a prototype implementation of the proposed method integrated in the *GreatSPN* framework [7] was developed. In detail, we extended the *GreatSPN* tool PN2ODE providing the automatic translation from SPNI to the ODEs system and OP system. A scheme representing this process is reported in Fig. 8. The generated ODEs system and OP are encoded in R language and saved into a file that is processed through the *R environment* to obtain the system behavior. The R package *deSolve* [78] is used to solve ODE integration, while package *GenSA* [234] is used to solve the OP.
The translation tool takes as inputs:

- the SPNI model, drawn with the GreatSPN GUI and saved into two files with extension.net and.def;
- One text file listing the undetermined bounded transitions;
- One text file containing the objective function.

The translation is performed with the following pipeline:

- The program extrapolates from.net and.def files all the PN information, such as places, transitions, arcs, initial marking and firing rates. Unknown transition rates are marked as NA.
- The program verifies the correctness of the file containing undetermined bounded transitions. For each undetermined transition T_u two values $\Lambda_u^L(t)$ and $\Lambda_u^U(t)$ that bound its flux are required. Optionally, the starting point, from which the OP solver starts searching the optimal solution, can be specified. If no starting point is provided, then a default value is computed as half of the sum of $\Lambda_u^L(t)$ and $\Lambda_u^H(t)$.
- The objective function, stored in the .txt file, is processed through a *lex and yacc* parsing tool. It can be a generic expression whose terms are the places and the transitions of the net.
- The whole translation process is executed from the command line as follows:
  PN2ODE SPNI_file -M -P -T./transitions_file -F./obj_fun_file ,
  where -M enables Mass Action policy, -P enables export format in R with the optimizer, while -T and -F are respectively used to specify the

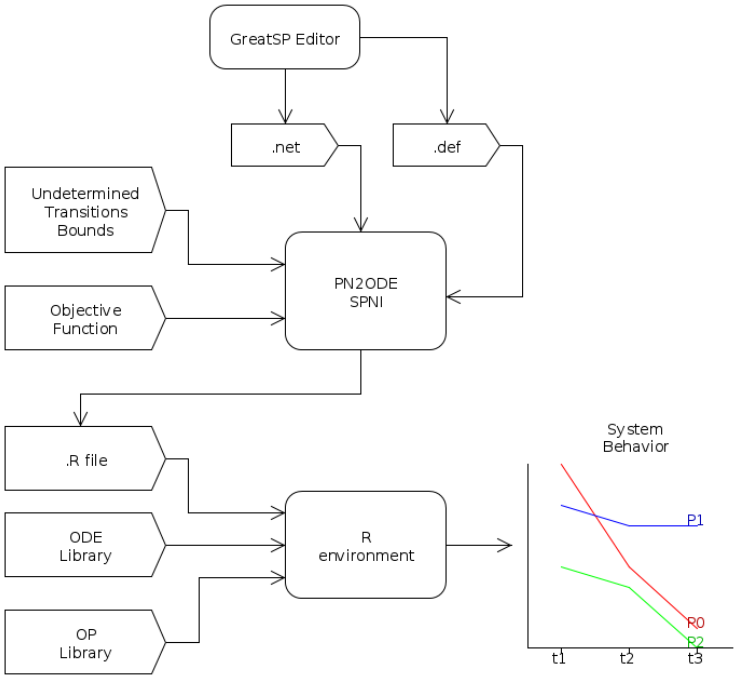

Figure 8: PN2ODE translation process.

two text files containing the undetermined bounded transitions and the objective function.

Case study. The proposed approach is used to investigate the metabolic behavior of cancer cells to illustrate its practical applicability. The model represents the glycolytic pathway in a generic human cell. It is inspired by the model presented in [74], which describes glycolysis in human red blood cells. Glycolysis is the most important and best studied intracellular metabolic pathway. In every cell of the human body, it leads to the consumption of Glucose (GLC) and a progressive production of Pyruvate (PYR) and energy, in the form of Adenosine Triphosphate (ATP). Then, in physiological conditions, in the presence of oxygen, PYR is metabolized by other pathways to generate the majority of the energy consumed by the cell. In absence of oxygen, PYR is converted to LAC without further energetic yields. The model is characterized by seventeen metabolic reactions, the related equations are reported in the first column of the Table in Fig. 10, and it can be graphically described by the SPN model in Fig. 9 where place names are chosen to recall the corresponding biological compounds. The first and the last transitions are included to reproduce the inflow of GLC and the outflow of LAC in and from the cell. All the other transitions describe forward and reverse reactions, catalized by specific metabolic enzymes.

Differently from normal cells, cancer cells exhibit an enhancement of glycolysis and production of LAC even in the presence of oxygen, a phenomenon known as Warburg Effect [62]. This phenomenon represents the central focus of our experiments. It has been recently shown that metabolic alterations

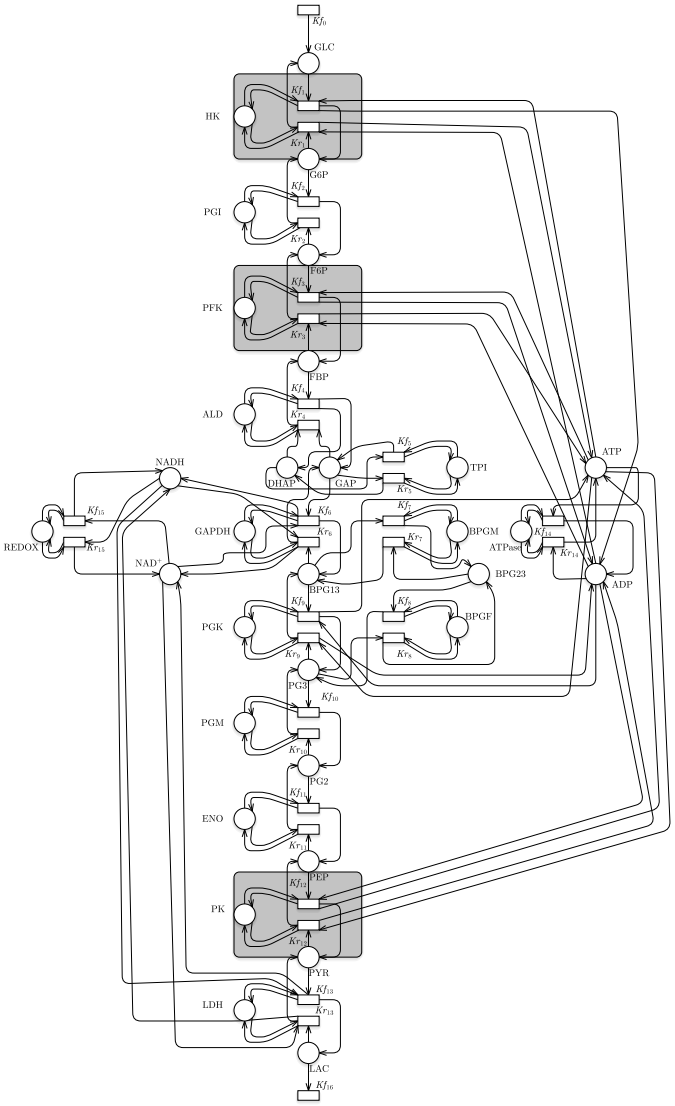

Figure 9: Case study: Glycolysis in *Homo Sapiens*.

Reactions	Rate Equations
$\emptyset \xrightarrow{Kf0} GLC$	$Kf_0 = 6.48E+06$
$HK + GLC + ATP \underset{Kr1}{\overset{Kf1}{\rightleftarrows}} HK + G6P + ADP$	$Kf_1 = 6.48E+05, Kr_1 = 4.9E+03$
$PGI + G6P + ATP \underset{Kr2}{\overset{Kf2}{\rightleftarrows}} PGI + F6P$	$Kf_2 = 1.15E+03, Kr_2 = 2,68E+03$
$PFK + F6P + ATP \underset{Kr3}{\overset{Kf3}{\rightleftarrows}} PFK + FBP + ADP$	$Kf_3 = 1E+09, Kr_3 = 8.47E+04$
$ALD + FBP \underset{Kr4}{\overset{Kf4}{\rightleftarrows}} ALD + DHAP + GAP$	$Kf_4 = 1.46E+02, Kr_4 = 1.18E+00$
$TPI + GAP \underset{Kr5}{\overset{Kf5}{\rightleftarrows}} TPI + DHAP$	$Kf_5 = 7.93E+00, Kr_5 = 4.53E+06$
$GAPDH + GAP + NAD^+ \underset{Kr6}{\overset{Kf6}{\rightleftarrows}} GAPDH + BPG13 + NADH$	$Kf_6 = 1.42E+05, Kr_6 = 5.28E+06$
$BPGM + BPG13 \underset{Kr7}{\overset{Kf7}{\rightleftarrows}} BPGM + BPG23$	$Kf_7 = 1E+08, Kr_7 = 1E+05$
$BPGF + BPG23 \underset{Kr8}{\overset{Kf8}{\rightleftarrows}} BPGF + PG3$	$Kf_8 = 6.84E+02, Kr_8 = 1E-09$
$PGK + BPG13 + ADP \underset{Kr9}{\overset{Kf9}{\rightleftarrows}} PGK + PG3 + ATP$	$Kf_9 = 2.61E+04, Kr_9 = 1.45E+01$
$PGM + PG3 \underset{Kr10}{\overset{Kf10}{\rightleftarrows}} PGM + PG2$	$Kf_{10} = 5.38E+01, Kr_{10} = 7.92E+00$
$ENO + PG2 \underset{Kr11}{\overset{Kf11}{\rightleftarrows}} ENO + PEP$	$Kf_{11} = 5.82E+02, Kr_{11} = 3.44E+02$
$PK + PEP + ADP \underset{Kr12}{\overset{Kf12}{\rightleftarrows}} PK + PYR + ATP$	$Kf_{12} = 5.17E+02, Kr_{12} = 5.17E-01$
$LDH + PYR + NADH \underset{Kr13}{\overset{Kf13}{\rightleftarrows}} LDH + LAC + NAD^+$	$Kf_{13} = 1.04E+03, Kr_{13} = 2.34E+00$
$ATPase + ATP \underset{}{\overset{Kf14}{\rightleftarrows}} ATPase + ADP$	$Kf_{14} = 9.74E-01, Kr_{14} = 9.74E+00$
$REDOX + NAD^+ \underset{Kr15}{\overset{Kf15}{\rightleftarrows}} REDOX + NADH$	$Kf_{15} = 9.74E-01, Kr_{15} = 9.74E-04$
$\emptyset \xrightarrow{Kf16} LAC$	$Kf_{16} = 1$

Figure 10: Table: Reactions, Equations and Initial marking of glycolysis in *Homo Sapiens*.

seen in cancer cells are promoted by specific mixtures of isoforms of their metabolic enzymes. In particular, it seems that isoforms of **Hexokinase** (HK), **Phosphofructokinase** (PFK) and **Pyruvate Kinase** (PK) may play an eminent role [124]. Despite these discoveries, it is still complicated to characterize the *in vivo* kinetics of these isoforms. Conditioned by these constraints, we chose to set the reactions involving HK, PFK and PK as undetermined transitions, i.e deficient of a complete list of regulators and of a specific mathematical expression containing its kinetic parameters. Our approach is used here as an attempt to acquire a deeper understanding of cancer metabolic dynamics. The idea is to use an objective function that encodes the Warburg Effect, considering every type of cancer at every possible tumour stage. We decided to formalize it as the maximization of LAC production at every integration step. Thus, the optimization process searches the values of the firing intensities of all T_us that allow to maximize LAC. The fluxes of undetermined transitions $T_{u_{f_1}}$, $T_{u_{r_1}}$, $T_{u_{f_3}}$, $T_{u_{r_3}}$, $T_{u_{f_{12}}}$ and $T_{u_{r_{12}}}$ were allowed to vary in a wide range that agrees with the available biological knowledge. Specifically the boundary conditions were set as follows:

$$
\begin{aligned}
\Lambda_u^L(T_{u_{f_1}}) &= 1620 \cdot x_{HK} \cdot x_{GLC} \cdot x_{ATP} \\
\Lambda_u^U(T_{u_{f_1}}) &= 2.592e+08 \cdot x_{HK} \cdot x_{GLC} \cdot x_{ATP} \\
\Lambda_u^L(T_{u_{r_1}}) &= 12.24 \cdot x_{HK} \cdot x_{G6P} \cdot x_{ADP} \\
\Lambda_u^U(T_{u_{r_1}}) &= 1.9584e+06 \cdot x_{HK} \cdot x_{G6P} \cdot x_{ADP} \\
\Lambda_u^L(T_{u_{f_3}}) &= 2.5e+06 \cdot x_{PFK} \cdot x_{F6P} \cdot x_{ATP} \\
\Lambda_u^U(T_{u_{f_3}}) &= 4e+11 \cdot x_{PFK} \cdot x_{F6P} \cdot x_{ATP} \\
\Lambda_u^L(T_{u_{r_3}}) &= 211.864 \cdot x_{PFK} \cdot x_{FBP} \cdot x_{ADP} \\
\Lambda_u^U(T_{u_{r_3}}) &= 3.38983e+07 \cdot x_{PFK} \cdot x_{FBP} \cdot x_{ADP} \\
\Lambda_u^L(T_{u_{f_{12}}}) &= 1.29234 \cdot x_{PEP} \cdot x_{PK} \cdot x_{ADP} \\
\Lambda_u^U(T_{u_{f_{12}}}) &= 206774 \cdot x_{PEP} \cdot x_{PK} \cdot x_{ADP} \\
\Lambda_u^L(T_{u_{r_{12}}}) &= 0.0058511 \cdot x_{PK} \cdot x_{PYR} \cdot x_{ATP}
\end{aligned}
$$

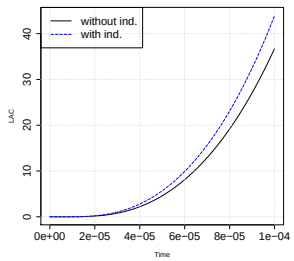

Figure 11: Behaviours of place *LAC*

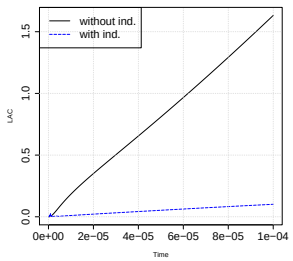

Figure 12: Behaviours of place *F6P*

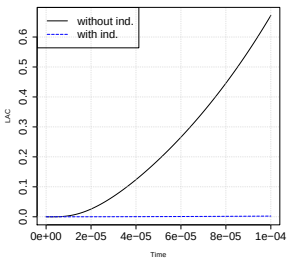

Figure 13: Behaviours of place *PEP*

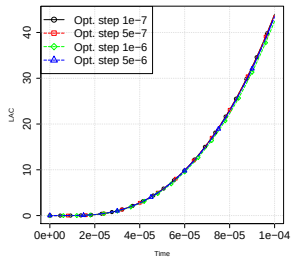

Figure 14: Behaviours of place *LAC*

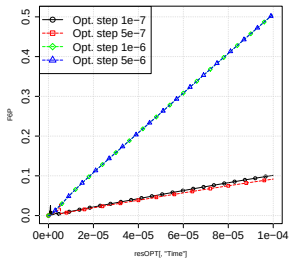

Figure 15: Behaviours of place *F6P*

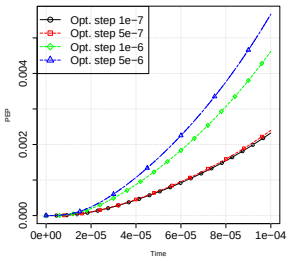

Figure 16: Behaviours of place *PEP*

$$\Lambda_u^u(T_{u_{r_{12}}}) = 0.620322 \cdot x_{PK} \cdot x_{PYR} \cdot x_{ATP}$$

Fig. 11 shows the evolution of LAC over time. The black line represents the results of the model of a normal cell, where all parameters are well characterized. The blue dashed line show the time evolution of LAC when uncertainty is applied. It can be noticed that this objective function is able to drive the system to accelerate LAC production. Even if the difference might not seem large enough to represent the Warburg Effect, we point out that these diagrams show the behavior of our model for a very short time interval. Fig. 12 shows the rapid accumulation of Fructose 6-Phosphate (F6P) in the normal cell model compared to a more balanced production-consumption dynamics in the cancer model. F6P, a high glycolitic intermediate, increases as a direct consequence of GLC degradation and is later processed by PFK.

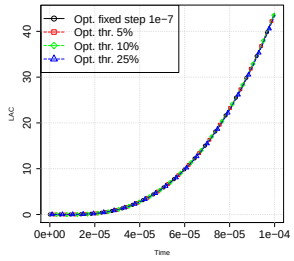

Figure 17: Behaviours of place *LAC*

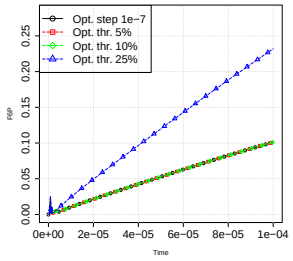

Figure 18: Behaviours of place *F6P*

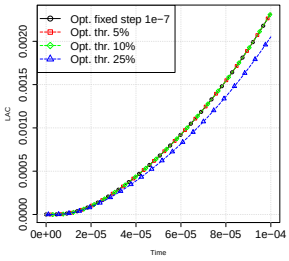

Figure 19: Behaviours of place *PEP*

It is then significant to see which parameters the optimization solver independently chooses to tune in order to maximize its objective function. While parameter values of HK and PK did not vary markedly if compared to the normal cell model (data not shown), the highest difference regarded PFK. Many articles as [135] have demonstrated that PFK kinetics is highly nonlinear and depends on many allosteric interactions. Our results seem to reinforce Mulukutla's thesis [135] that the regulation of PFK activity has a crucial impact on the glycolitic flux and may be relevant to explain metabolic alterations in cancer. In the absence of quantitative data to perform a quantitative validation of our approach, we consider that the emergent finding of the importance of PFK parameters for the behavior of the system can be regarded as a qualitative validation of our approach.

With an additional set of experiments we studied if it was possible to reduce the computational costs of our approach while maintaining a good accuracy of the solution. With this intent, we progressively decreased the frequency at which the optimizator was invoked and then explored the performance of our algorithm. When the optimization process is not repeated at every integration step the time-varying parameters associated with T_us are then transformed into piecewise constant parameters. This can be motivated with the assumption that in the time interval between two consecutive optimization processes some fixed parameter values can well approximate the real behavior of all T_us. The time interval that separates different optimization processes, in alternative called optimization step, was set to four different values: $1e^{-7}$, $5e^{-7}$, $1e^{-6}$, $5e^{-6}$ h. As reported in Table 1 the resulting computational times are compared. As expected, reducing the number of optimization processes allowed to significantly diminish the overall computational efforts of the algorithm. We then studied how these changes affected the dynamics of the system. Intriguingly, we found that some places, like LAC, as shown in Fig 14, displayed little or no changes, while for other places, like F6P and PEP, the changes were much more relevant, as shown in Fig. 15 and 16. We can then observe that for the more sensible places the reduction of computational time comes at the cost of the precision of the solution, which nevertheless maintains the capability to provide some qualitative information about the behavior of the model. In a further set of

ODE	**Opt int 1e-7**	**Opt int 5e-7**
0.732235s.	2441.878sec.	569.7389sec.
	Opt int 1e-6	**Opt int 5e-6**
	254.9282sec.	119.2946sec.

Table 1: Execution times for different optimization time intervals. All executions were performed on an INTEL i7 64bit 2.60GHz processor

experiments we tested if the time step of the optimization processes could be adaptively tuned along the solution of the system. The rationale behind this choice is to invoke the optimization process just if it produces significant effects. In order to do it, we added in our algorithm an heuristics, presented

in Chapter 3, that considers the relative change in the markings of input places of all T_us. When the relative change in one of these places crosses a user-defined threshold, the optimization process is executed. We studied how three different thresholds values (i.e. 5%, 10% and 25% of change) impacted both the computational time and the accuracy of the solution. Figures 18, 19 and 17 display the differences in the dynamic behaviors of F6P, PEP and LAC, when the different threshold values are considered. The execution times in the three cases are reported in Table 2, compared to the performance of the algorithm in the absence of the implemented heuristics. Data clearly show that, for instance when a 5% threshold value is used, no negative effects reflect on model solutions, while the computational cost is reduced of a factor greater that 10. Regarding the accuracy of the solution, also greater threshold values are satisfactory, with the advantage that they allow considerably higher time savings.

Opt int 1e-7	Thr=5%	Thr=10%	Thr=25%
2441.878sec.	195.0151sec.	135.4613sec.	82.67076sec.

Table 2: Execution times for different threshold values of relative change in place markings. All executions were performed on an INTEL i7 64bit 2.60GHz processor

It is important to observe that our approach, moving forward from the strict teleonomic perspective adopted by FBA and CM, allows to define objective functions that may mimic some experimentally observed behavior. It is worth underlying that teleonomy has shown great utility to help dissect metabolic characteristics in microorganisms. In the case of multicellular and more complex organisms like humans, however, the process of selection that justifies the teleonomic view is much more complex, and a clear goal-directedness of intracellular metabolism is not equally reasonable to consider. However, the LAC maximisation function, which we adopted in our case study of cancer glycolysis, besides reflecting the fact that a high production of LAC in cancer cells is a renown experimental finding, may also have a teleonomic value. In fact, it can be hypothesized that the process of evolution of cellular phenotypes observed in cancer, i.e. in the selection of aggressive cancer clones during its progression (current focus of extensive studies [286]), tends to select a goal-directed cellular phenotype characterized by LAC maximization.
Moreover our proposed approach, differently from FBA and CM, does not assume the steady state of the intracellular metabolites, as in FBA, and it does not require a complete knowledge of all kinetic parameters, as in CM. Finally, it is useful to highlight again that the lack of experimental data, which is very frequent in cancer cell scenarios, does not affect our approach as it does for techniques of RE and PE.

2 DATA INTEGRATION

In chapter 3 we presented a method that can be used to convert a validated, condition-specific kinetic model, into a new model, specific for a new condition of interest. Fundamental requirements to use this approach are gene expression data in the form of FPKMs, or in case they are available, protein quantifications by proteomics technologies.

In this section we intend to illustrate the process that accompanied our modeling efforts and discuss some theoretical as well as practical considerations arisen from these results. The structure of this section will be thus more conversational.

The work that led us to elaborate this approach started with a process of literature review, in which we scrutinized the most significant examples of computational models of cancer metabolism. A summary of some of these is reported in section 2. As already mentioned, modeling cancer metabolism is an intrinsically difficult challenge, both for the complexity of its pathophysiological aspects, both for the low availability of biological data. So far these substantial limitations have strongly conditioned both the number of modeling attempts as well as the methods adopted. The vast majority of modeling experiences focused in fact on constraint-based models, while just a few authors proposed kinetic models.

Acknowledging the methods already implemented in the literature, we asked if these modeling efforts could be complemented and improved through additional procedures of data integration. With this goal in mind, we decided to aim attention at kinetic models of metabolism. The description of the system they are potentially able to provide is attractive, especially if we consider the strongly dynamic characteristics of cancer. As an additional motivation of our choice, many other technique, reviewed elsewhere [119, 147], had already been proposed for constraint-based models.
Moving in a relatively unexplored field, our investigation was not devoid of complications. We will accompany the discussion of our results reviewing the main difficulties and obstacles that can be encountered in the process.

In order to build and validate kinetic models of metabolism, the most informative type of data are time-course profiles of metabolite concentrations. The scientific community is confident in the belief that these data will become cheaper and cheaper, and more and more available in the near future. At the moment, however, they are hardly accessible. In the absence of such a dynamic overview of the system we are studying, top-down modeling strategies are hampered. We hence devoted our efforts to a bottom-up strategy and we inspected the sources of kinetic information that are currently at our disposal. As we saw in section 1.1 these can be distinguished into databases where kinetic parameters are archived, and on-line repositories of kinetic models that have been already validated and published. Regarding the former, we already discussed that the data collected in these databases always has a highly heterogeneous provenience, thus their use should be prudent. Concerning the latter, we searched a model that we could use as a scaffold

to test our data integration approaches.
In this search, it can be rapidly discovered that stored models are often incomplete of some necessary element that prevents a straightforward reproducibility of the published results.

RED BLOOD CELL MODEL Among the models we checked, no kinetic model specific of cancer metabolism was available. Thus, looking for models of non-transformed, human cells, we selected the red blood cell model, in which Jamshidi and Palsson proposed their MASS approach based on the approximate mass action rate law [73]. As a motivation of our choice, its formulation with the LMA makes the model easy to implement and check. Interestingly the stoichiometric matrix, the full list of estimated kinetic coefficients, the values of the fluxes at the steady state and the initial conditions were listed. The model was represented with an SPN, using the *GreatSPN* framework [7], a tool that simplifies the creation and simulation of SPN models. Exploting the features of the software, an SPN model was automatically translated into a system of ODEs describing its behavior.
We intended to test our approach trying to recast this RBC model into kinetic models representative of different, non pathological tissues.

In doing this, our intention was not to create models that could reproduce quantitative metabolic behaviors with accuracy. Rather, we wanted to verify if some qualitative differences that resembled the known metabolic diversity of these tissues could be reproduced. We downloaded FPKM values for five different tissues: muscle, liver, brain, pancreas and reticulocytes.
A factor we had not previously considered, as red blood cells have no nucleus, their mRNA production is no more active. We hence decided to use FPKMs as an estimate, even if rough, of enzyme expression levels in RBCs. Then all the reactions in the network were manually assigned to the corresponding reaction ID in the Human metabolic reconstruction Recon 2 [241]. Mapping model reactions to Recon 2 was necessary to compute the conversion. It should be highlighted in fact that the integration of FPKM values requires that these, which have a one-to-one relationship with the genes, are mapped into a coefficient that defines the level of activation of the reaction. This mapping procedure requires gene-protein-reaction (GPR) rules, logic expressions in which gene identifiers are linked by AND and OR operators. In chapter 2 we mentioned that other more recent reconstructions, like Recon 2.2 [233] and Recon 3D [18], have already been published. We however decided to use Recon 2 for practical reasons of the mapping process. Referring to the analyses carried out in [147], we believe that the results obtained with our data integration approach can be still evaluated regardless the choice of the human reconstruction. Anyway it is part of our future intentions to try to assess the differences we would obtain in our results if we used newer metabolic reconstructions.

A specific feature of the original model, the reactions are described with two different levels of details. For some of them, all enzymes in their complexed and uncompleted forms are represented as variables in the system. For the others, instead, the enzymes are not modeled, and the reaction directly catalyzes the conversion of substrates into products. This diversification was taken into account using two different modes of integration of FPKM data. Where an enzyme was not specifically modeled, FPKM values were used to recast k_f and k_r values of the original model into the new one

$$k_f^M = k_f^R \cdot \frac{FPKM^M}{FPKM^R}$$

$$k_r^M = k_r^R \cdot \frac{FPKM^M}{FPKM^R}$$

Where, in this example, M indicates the target condition (muscle) while R the reference condition (reticulocyte). For all the reactions in which enzymes were explicitly modeled, instead, FPKM values were used to recast the initial concentrations of total enzyme moieties, starting from the initial concentrations specified in the supplementary material of the paper. With the recast completed for all the enzymes in the network, we were able to produce time course profiles for each metabolite/enzyme in the model.
The analysis of these results was unfortunately inconclusive, hindered by the inability to check that the original model (not recast) was correct, as the paper did not report dynamic profiles that could be used as a comparison. In our simulations the dynamic trends of the most important intracellular metabolites showed unreasonable behaviors that could not be more accurately interpreted. Moreover, following the procedure exposed in the paper, we tried unsuccessfully to retrieve the steady state fluxes starting from the reported kinetic rate constants and the steady state metabolite concentrations, using LMA formula:

$$v_1 = k_1^+ A^2 - k_1^- B \tag{16}$$

The recasting process was however able to recreate distinguishable profiles for the different tissues, but deeper speculations on whether these quantitative differences were consistent with the biological differences among the tissues were impeded. We hypothesized that these apparently anomalous behaviors were caused by some error in the model description. It is out of the scope of this thesis to be more detailed here, however we do believe it is interesting and important to witness that the reproducibility of published data is an issue that should be more and more discussed by the community in the future.

ALTERNATIVE GPR MAPPING METHODS Even though the recast models could not deliver interpretable predictions, they were however used to analyze the impact of two alternative implementations of GPR rules conversion.

As we have previously explained in section 1.5, some authors proposed that the logic operators AND and OR contained in the rules are converted into $min()$ and $max()$ functions. According to other perspective found in the literature, we believe that the $OR-to-sum()$ conversion is more appropriate. Here, we assessed the effect of choosing one policy of translation versus the other. As we mentioned in section 1.5, computing the transformation from gene expression levels to reaction expression levels through GPR rules is not trivial, and some specific algorithms have been proposed for this task [9]. Unable to run these for compatibility reasons, we implemented the $AND-to-max()$ and $OR-to-sum()$ conversions manually. Figures 20 show the differences in the dynamic behavior of the model if we use the MATLAB functions extractGPRs() and mapGeneToRxn() included in the COBRA Toolbox versus the mapping we implemented manually. Even though the differences are not uniformly marked, and the trajectories tend to converge to the same steady state, it can be seen how the choice of the type of conversion influences significantly the transient behavior.

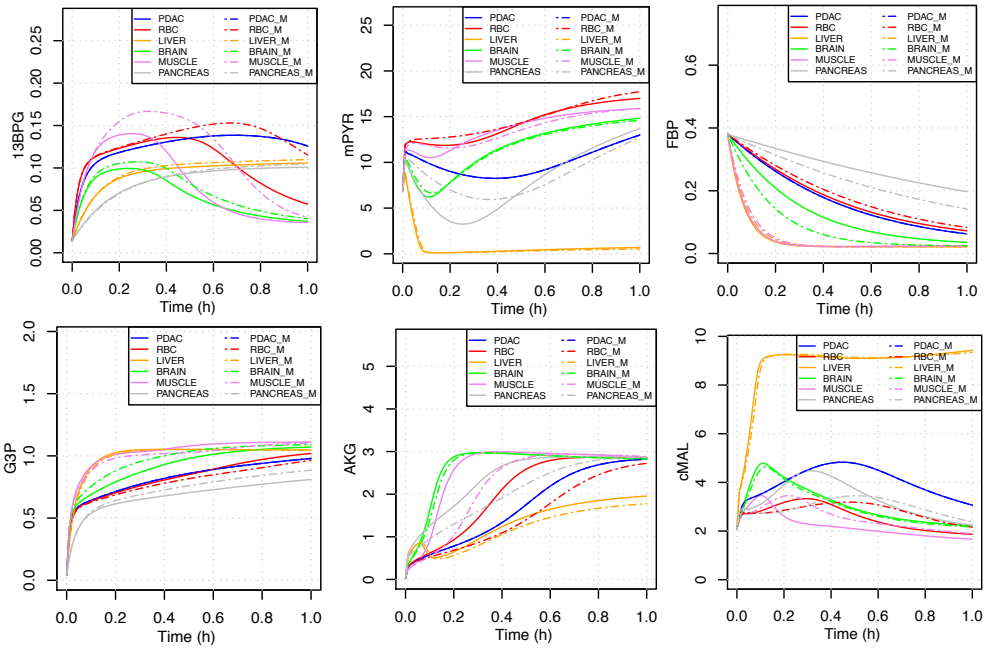

Figure 20: Continuous lines: behavior obtained with COBRA Toolbox functions; dashed lines: profiles obtained with $OR-to-sum()$ implemented manually.

PANCREATIC DUCTAL ADENOCARCINOMA MODEL In order to explore the prediction capabilities of our recasting method, the MASS RBC model was inadequate, and thus abandoned. A new recently published kinetic model [186], was chosen as a scaffold to test our approaches. In this work the authors built a kinetic model of the core metabolic pathways which for the first time represents metabolic alterations in KRAS-mediated pancreatic ductal adenocarcinoma (PDAC). The model is composed of a set of ordinary differential equations that describe the time evolution of metabolite concentrations

in glycolysis, glutaminolysis, tricarboxylic acid cycle and the pentose phosphate pathway. The model, comprising 46 metabolites and 53 reactions, was fit to published enzyme knockdown experimental data specific to PDAC cell lines. The building phase was carried out with processes of identifiability analysis and parameter estimation. The analyses reported in the paper explain the ability of the model to simulate *in silico* enzyme modifications and evaluate the effects on cell proliferation. Attaining to what the authors declare, the model identified some potential combinations of enzyme knockdowns, metabolite inhibitions, and extracellular conditions that impede cell proliferation. A scheme of the system of reactions contained in the original model by is reported in figure 21

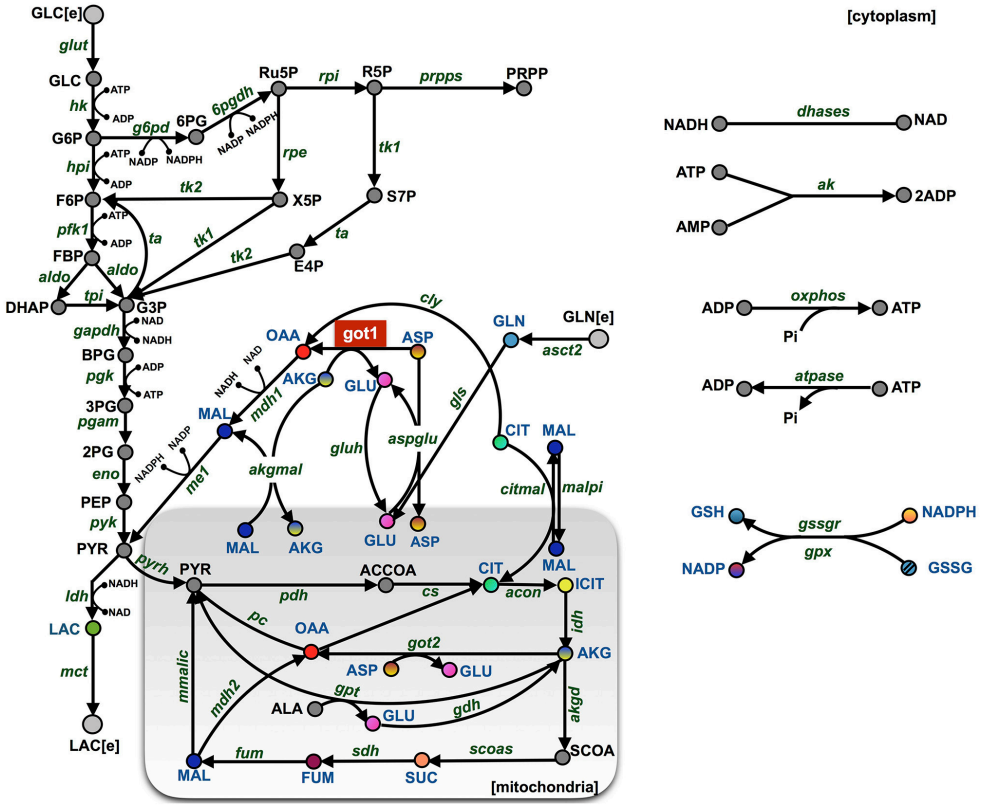

Figure 21: Metabolic reaction network of the PDAC mechanistic kinetic model [186]

Differently form the RBC MASS model, this model is fully detailed, with kinetic expressions of reaction fluxes that involve several allosteric interactions. The baseline model, adapted from literature, has a total of 372 parameters, which includes 71 reaction velocities. The model implementation is reported in MATLAB and SBML formats, provided as a supplementary material of the paper. Importing the SBML model in COPASI [69] and solving the system of ODEs, however, the computed time profiles of metabolite concentrations seem to contradict what is reported in the paper. More specifically, the main issue we encountered relates to the growth curves in figures 3 and 4 of the paper: neither in the paper nor in the supplementaries it is further specified how the "relative cell growth" is computed, and the abso-

lute cell number shows a early and rapid increase a the beginning of the simulation that can be hardly matched with the profiles shown in figures 3 and 4 in the paper [186]. Concerning this last point, and according to what is reported in the paper, the growth rate equation is defined as dependent on three metabolites, assuming a " Monod-type function":

$$\mu = \alpha_{atp}(\frac{ATP}{k_{ap} + ATP}) + \alpha_{glc}(\frac{Gl_{in}}{k_{gc} + Glc_{in}}) + \alpha_{gln}(\frac{Gln_{in}}{k_{gn} + Gln_{in}}) \quad (17)$$

The Monod equation is a mathematical model that links microbial growth rate to the concentration of a limiting nutrient in the growth medium.

$$\mu = \mu_{max}\frac{s}{(K_s + s)} \quad (18)$$

Observing Monod's equation, we could not interpret authors' choice to modify the equation replacing substrate concentrations in the media with intracellular concentrations. As glucose and glutamine start being upkaten by cells, their intracellular concentrations rise rapidly. If we consider just the portion of the equation that depends on one substrate, glucose for instance, we can see how the term

$$\alpha_{glc}(\frac{Glc_{in}}{k_{gc} + Glc_{in}})$$

would tend to increase as Glc_{in} increases, because k_{gc} (like k_{ap} and k_{gn}) has a value smaller than one. As the growth curve that best recapitulates the experimental growth profile is the logistic, one would expect the value of the growth rate to start at high values and progressively diminish to become zero at the plateau, and not a growth rate changing from low to high in the first phase. Also the meaning of the carrying capacity was not of easy interpretation. The differential equation that describes the changing in time of the number of cells is defined in the model as :

$$d_C = \mu * C * (1 - (C/kk)) - \mu_d * C$$

, where C stands for the cell number, kk for the carrying capacity and μ_d represents the death rate parameter. The carrying capacity is normally used to define the maximum size that the population can reach. Here, however, as the carrying capacity has no effect on the $\mu_d * C$ term, it can be the case that a steady state is reached at lower values that the carrying capacity. For these reasons, we decided to let the model display a growth curve that is not influenced by a kk parameter. In our model, the definition of μ and of $\frac{dC}{dt}$, were changed to the following:

$$\mu = \alpha_{glc}(\frac{Glc_{out}}{k_{glc} + Glc_{out}}) + \alpha_{gln}(\frac{Gln_{out}}{k_{gln} + Gln_{out}})$$

$$dC = \mu * C$$

The resulting code was run and was able to reproduce the temporal behavior of all system variables. The plots in figure 22 show the reproduced behavior of the PDAC kinetic model, with some components modified as we just mentioned.

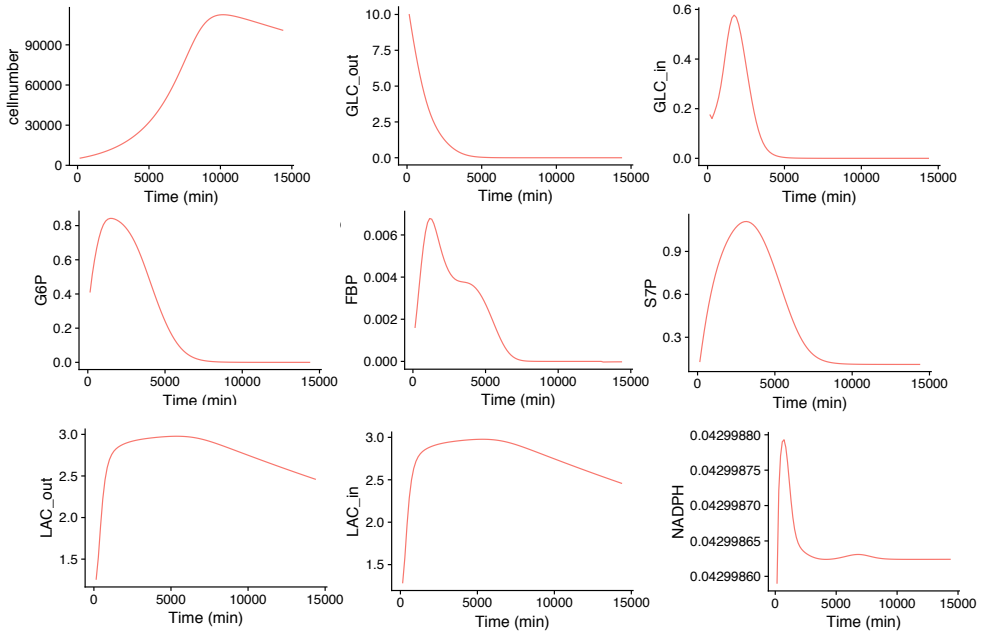

Figure 22: Dynamic behavior of intracellular metabolites obtained with the PDAC kinetic model

ORIGINAL MODEL RECAST TO NEW NCI-60 CELL LINE MODELS After this preliminary phase of model adjustments, we proceeded to test our method of data integration. Intriguingly, in section Methods of the paper, the authors explain, literally, how "reaction velocities are thought to distinguish metabolism across different cell types. Therefore, of the many kinetic parameters included in the reaction rate equations, only the reaction velocities were fit to the training data, and the other rate constants were held at their literature values".

In a process of identifiability analysis the authors assessed the correlation among the 71 reaction velocities for different initial conditions in the model. When the forward and reverse reaction velocities (V_f and V_r, respectively) for a particular reaction were shown to be highly correlated for multiple sets of initial conditions, only the V_f was considered for the data fitting procedure, while V_r was indirectly calculated from the value of V_f and the equilibrium constant of the reaction. The size of the parameter set that was considered for data fitting was thus shrunk from 71 to 59 parameters. These were fit to published experimental measurements consisting of the fold change in steady state concentrations of 14 intracellular metabolites upon GOT1 knockdown [217]. The data used for the fit was produced using 8988T cell line cultures.

Acknowledging these correlations present among v_{max} parameters, we proceeded recasting only the 59 maximal velocities that were not strongly correlated among them. The fact that several attempts of data integration within constraint-based models have been performed on NCI-60 cancer cell line

panel, motivated us to consider this panel to explore our approach and to compare it to the results of those integration processes.

Before we could perform the actual recast, we mapped the reactions defined in the paper with the reaction IDs listed in Recon 2 with the most similar stoichiometry. This task was performed manually, and was not devoid of obstacles. For each reaction included in the model, multiple reactions listed in Recon 2 shared the same reactants and products. Among these, we tended to select the ones that allowed forward and reverse fluxes. For the cases in which the choice would have been ambiguous, we used the reconstructions of NCI-60 cancer cell lines core metabolic pathways in [305] as a comparison and check.

Mapping genes to reactions

Once each reaction in the model was assigned a reaction in Recon2, for each reaction we retrieved the relative GPR rule. In this case, again, the conversion was performed replacing manually AND with $min()$ and OR with $sum()$.

All 59 low correlated v_{max} parameters were recast with the same procedure presented in chapter 3 and reviewed in this section.

For the FPKM of the reference condition we considered gene expression data of 8988T cell line, as data used for the PE in the original model were produced with this cell line. For the gene-expression data we downloaded publicly available FPKM data for 675 commonly used human cancer cell lines from https www.ebi.ac.ukarrayexpress with the accession EMTAB2706.

As a result, we obtained more than 675 models, one for each cell line, displaying different dynamic behaviors.

Performing the conversion by simple proportion we realized that kinetic parameters can become unreasonable big or small. If we consider that the original parameter set spans several orders of magnitude, specifically from $1e^{-10}$ to $1e^{+42}$, we can understand that, after the proportions are applied, this range becomes even larger. The biggest issue we had with these parameters was more practical than theoretical. When kinetic rate constants become so different, the resulting ODEs system becomes stiff, causing the simulations to slow down significantly or to stop. We should just mention here that problems of stiffness with numerical solutions represent in general an additional obstacle to the reproducibility of results present in the literature. In order to overcome the stiffness of the ODEs system we decided to define an upper bound for reaction velocities equal to $1e^{+15}$: all velocities greater that this value were adjusted to it. This can in principle reduce the differences among recast models. However just around 5 parameters over 59 had to be corrected to stick to this bound, thus we can be reasonably confident that we are not losing too much information with this procedure. Even if this was not the case, we believe that a limited computational cost needs to be a primary requirement for the utility of our approach. Hence, it was considered that the capability of obtaining more accurate simulations with tedious simulation times fell outside the goals of our method.

Dealing with ODEs stiffness

As an example of the type of outputs we were able to generate, the dynamic profiles for all 675 cell lines can be found in the appendix A.

The fact that the NCI-60 panel has been deeply and repeatedly studied offers the possibility to compare model predictions with many types of data. We report here a short list of some of the publicly available information that can be found on-line:

Publicly available data for NCI-60 cell lines

- Cell doubling times, from the NIH website, which inform us on the growth rate during the phase of exponential growth displayed by these cell lines *in-vitro*
- Exometabolomics data, reported in [72], which can inform us on extracellular metabolite concentrations. From these, in [72] uptake and release fluxes were also inferred
- Cell line specific biomass compositions [305]
- 13C fluxomics data for a few, widely used, cell lines [51, 127]
- Proteomics datasets, like the one at proteomics.wzw.tum.de/nci60/
- implementations and solutions of constraint-based cell line specific models, whose qualitative and quantitative solutions can be used as a source of comparison [305, 6, 279, 3, 210, 41, 291]

In addition to their abundance, these data are also easy to access, as they have already been processed several times in different studies.

Comparison with exometabolomics data

We tried to exploit these data in different ways to test the validity of our approach. In specific, we tried to match the data reported in [72] to recreate, *in silico*, the experimental conditions the authors used during the process of data acquisition. Starting from the information on the media in which cells were grown, congruent values for the initial conditions were defined. With a set of ODEs fully parameterized with the recast rate constants the behavior of the system could be simulated.

With the purpose to compare our results with the data in [72], we mimicked an exometabolomics analysis performing *in silico* sampling of metabolite concentrations at approximately 4-5 days from the start of the culture. Importantly, metabolomics data were available just for a small subgroup of metabolites, and only for a portion of 675 cell lines.

Quantitative comparison

In figures 23 we report a comparison between the dynamic profiles obtained with the system of ODEs and the metabolite concentrations measured experimentally. Exometabolomics data only refers to concentrations in the media, so in principle we could compare these data with the only three extracellular metabolites described in our model, namely glucose (GLC_out), glutamine (GLN_out) and lactate (LAC_out). Here, however, we decided to be less stringent and to anyway allow the comparison between exometabolomics data and the concentrations of the corresponding metabolite in our model, either this was intra or extracellular.

From the figures we can notice that the results are non-uniform. For some metabolites the simulated concentrations seem to reflect qualitatively

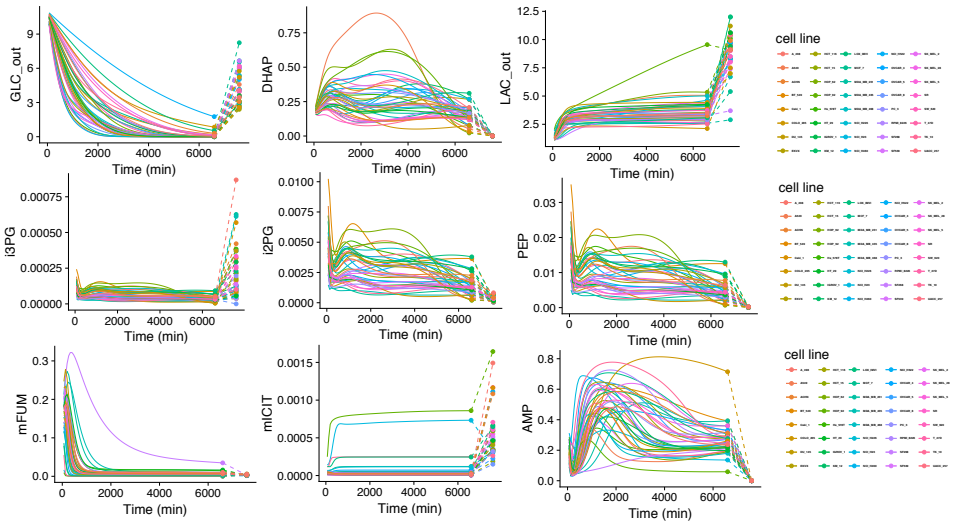

Figure 23: Comparison between the time behavior of the recast models and the concentration data retrieved from [72]. The dashed lines represent the match between the last time point in the simulation and the concentration data

with a good approximation the experimental measurements, for others the same is not true. Still it is can be noticed that for some metabolites, like lactate (LAC), mithocondrial citrate (mCIT), 2-Phosphoglycerate (2PG) and 3-Phosphoglycerate(3PG), some quantitative match can be observed. The range of simulated values and the range of experimental values, in fact, overlap.

Qualitative comparison

We then asked if, instead of evaluating the absolute values of the predicted metabolites, an analysis of their relative abundance could have been more informative. In this analysis then we compared concentrations at the end of the simulation with the measured values and checked whether we could find some analogies in the lower-to-higher ordering of these concentrations. Figure 24 shows just the final 20% of the simulated concentration profiles, compared with the experimental data, normalized and adapted to the range of simulated results. In this case, then, the absolute values of measured data looses its relevance. The dashed lines that connect predicted and experimental values are here meant to give a rapid and informative answer: if the ordering was conserved, we would want to see that lines do not cross one with each other. Here, again, the results show differences from metabolite to metabolite. For some subset of cell lines we can find that the higher-to-lower correspondence is maintained, however we could not find an overall consistent match between the measured and predicted concentrations.
We also tried to plot separately only the cell lines of a specific cancer (breast, colo-rectal, ovarian, brain), but we couldn't find a better match, so the plots are not reported here. As an additional comment on these plots, we should notice that the system is far from the steady state. We imagine an arguably better match could be obtained after a the stationary phase is reached.

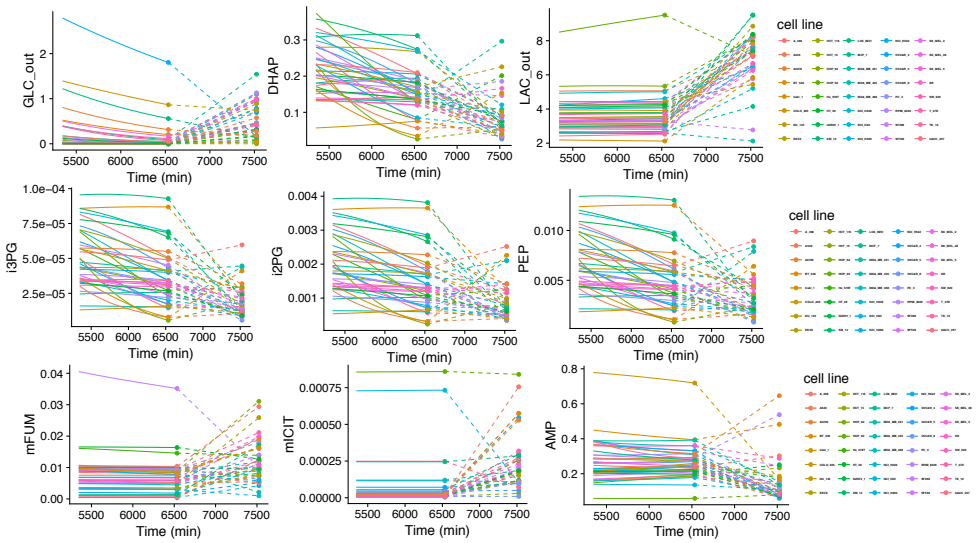

Figure 24: Comparison between the time behavior of the recast models and the concentration data retrieved from [72]. The dashed lines represent the match between the last time point in the simulation and the concentration data

Parameter Sweep Analysis

In the next step, we explored a way to improve the outcomes of our method. In doing so, a few considerations made us decide to disregard parameter estimation techniques. Firstly, for almost all the cell lines we considered, some experimental data like those reported in [217] and used to parameterize the original PDAC kinetic model are currently unavailable. Moreover, we sought to explore methods that are easier to implement, that do not require huge computational costs, and that could complement the recasting process rather than replace it. Thus, we asked if we could sensibly reach a better accuracy performing a parameter sweep analysis on a small subgroup of the recast parameters. In a parameter sweep, the values of one or more parameters are sampled randomly within a specific range, an ensemble of models is generated, and model outcomes are compared to the experimental results in order identify some combinations of parameters that could be better used to reproduce the data. With this goal in mind, we employed the Latin Hypercube Sampling method [205] to generate 100 combinations of the kinetic parameters associated with lactate, glucose and glutamine exchange reactions, over a range of $+/-5$ orders of magnitude. Due to the computational costs of the simulations, we limited the use of this sampling method to only 4 cell lines, namely A549, and breast cancer cell lines MCF$-$7, MDA_MB_231 and MDA_MB_468.

We ran the simulations and selected the parameterization producing the shortest distance from experimental data. In Figure 25 we can see how this additional procedure was able to slightly improve both the qualitative and quantitative predictions. Interestingly, if we look at the LAC subplot, we can

see that the higher-to-lower order is maintained.

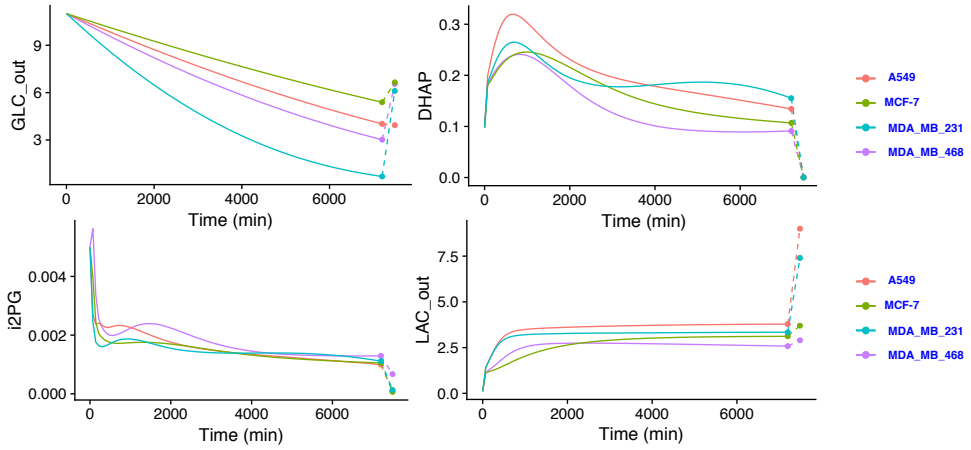

Figure 25: Comparison between the time behavior of the recast models for the 4 selected cell lines and the concentration data retrieved from [72]. The dashed lines represent the match between the last time point in the simulation and the concentration data

Potential applications of the presented approach: 1) provide qualitative information 2) support procedures of parameter estimation

Overall we can notice that the match between model outcomes and the experimental findings is loose. This stimulated us to go back and reconsider some of the concepts used in our approach. We then present here some considerations on the validity and utility of the method we proposed.

When, in the phase of model validation we judge the solidity of an approach, we should also acknowledge which of the currently available alternatives can perform similar tasks. If we are interested to a dynamic description of a system which, as it is often the case, has not been yet described by means of a kinetic model, we could imagine that any dynamic behavior could be possibly present. If we consider that the dynamics of the system is so highly undefined, we believe that this approach, which starts from a quantitative description of a reference condition and applies a quantitative transformation to describe a new condition, is still able to provide us some qualitative information of the uncharacterized condition of interest.
As we said, in the absence of time course-data, bottom-up strategies become the only plausible alternative to try to build kinetic models of cancer metabolism. One of their greatest values lies in their capability to gather high quantities of information collected along decades of investigation in biochemistry. Their most valuable characteristic, is, however also their most costly weakness. The high heterogeneity of these data, once integrated in a whole model, can in fact lead to possibly misleading results.

Considering all the factors we just mentioned, we believe the challenge of creating a kinetic model should be either abandoned, either embraced, together with all the uncertainty it is accompanied by. If we were to explore these approaches, and we faced the challenge of modeling bottom-up a sys-

tem we knew little about, we would start identifying the knowledge and data we can rely upon, and we would complement the lacking information using algorithms of parameter inference, as for the approach followed by the authors in [186]. If we were adopting a procedure analogous to theirs, that seeks parameter values starting from steady state data at a single reference state, our model would still be highly undetermined and so the confidence in inferred parameters would be low. In a context like cancer studies, where the information we can access is so little, we believe that every source of data could be effectively exploited, and so that there would be no reason to neglect gene expression data to focus solely on parameter estimation.

Moving one step further, we could speculate that gene expression data could facilitate some process of parameter estimation. When the model is very complex and the number of parameters to be estimated is high, in fact, the search space that is explored by parameter estimation algorithms becomes rugged. Parameter estimation tries to cope with these difficulties either with the use of global search algorithms, either repeating local search procedures from multiple, different starting points in the parameter space. In the scientific community global search algorithms are used less frequently, as they require that many meta-parameters are appropriately tuned. Considering this aspect, we believe that all PE problems in which local search approaches are used would benefit of a method, like the one we just proposed, as a way to identify a first draft parametrization that is later iteratively refined towards the optimum.

Integration of proteomics data

As an additional attempt to better interpret these results, we exploited the availability of protein expression data for NCI-60 cell lines (at proteomics.wzw.tum.de/nci60/) to perform the same exact recasting process. Proteomics quantifications are in fact normally believed to be more reliable than FPKMs for the inference of protein abundance. The process of data integration is however complicated by the fact that we did not have one protein for each gene symbol, and thus we did not have a protein expression value related to all the gene names that appear in gene-protein-reaction rules: for around 30% of the gene names no protein level was assigned. We then needed to check that the computation of reaction expression values was not drove to zero as a consequence of this. The outcomes of the comparison are presented in figures 26, 27, 28. The cell lines represented in the figures are the only ones for which both FPKMs and proteomics data were available.
From the plots we can notice that the behaviors are different, and that the curves generated with proteomics data display a higher diversity. For some metabolites, like ATP, it seems in fact that the computed dynamic behavior displays multiple steady states. Anyway, although tempting, if we consider the difficulties encountered in the GPRs mapping process, we believe that from these plots it would be inappropriate to reach any conclusion on the

better accuracy of one method with respect to the other. If we look at places like DHAP, FBP and G3P we can in fact suspect that the anomalous behavior of *SF539* and *SW_620* cell lines is the result of some error either present in the data, either introduced in the GPRs mapping process.

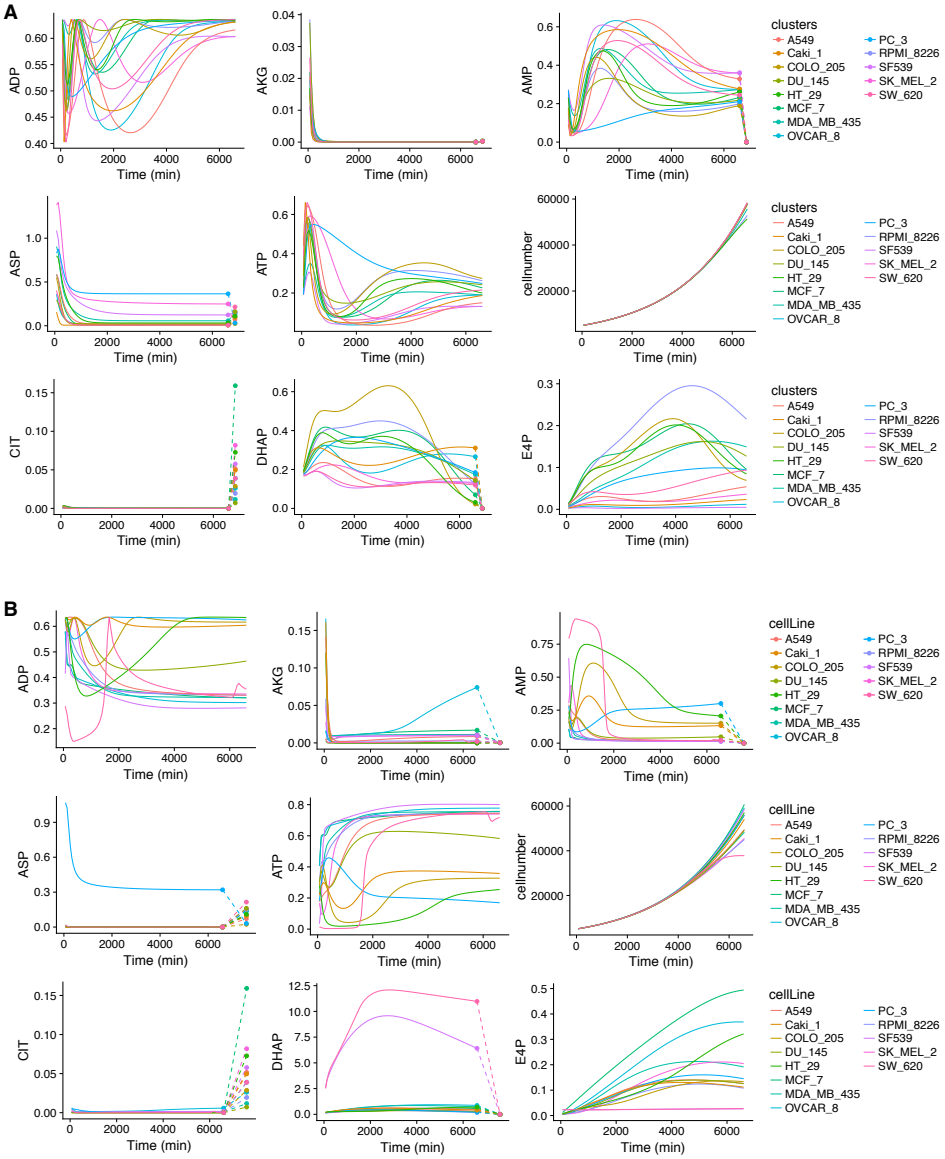

Figure 26: Behavior of cell line models obtained integrating FPKM values (A) and proteomics data (B)

As we have already discussed in chapter 2 other formalisms are available and amenable to be used to create models of cancer metabolism. Considering these results it could be argued that a wiser modeling choice should fall on a different approach, like the constraint-based. The detractors of constraint-based methods, on their side, believe that the tight dependence of

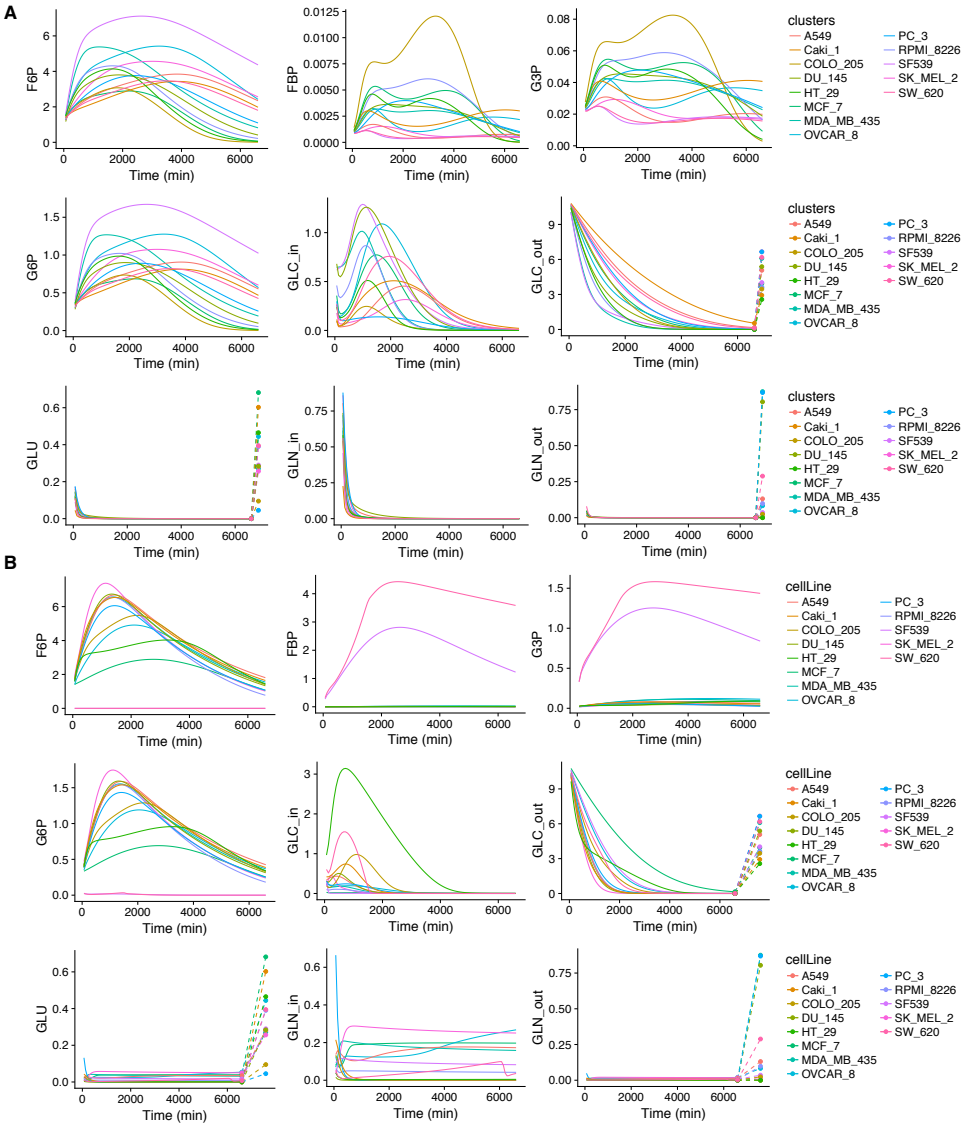

Figure 27: Behavior of cell line models obtained integrating FPKM values (A) and proteomics data (B)

FBA from the steady state assumption, and its disregard for metabolite concentrations are limitations that become invalidating when we try to study a highly complex and dynamic system like cancer. In between the two opposite positions, we believe that FBA and its extensions probably represent a reasonable compromise, but using constraint based models should anyway not preclude new attempts in the realm of kinetic models.

Our experiments suggest some observations on the effectiveness and accuracy of procedures that seek to integrate gene expression data into metabolic models. As we discussed in section 1.5 algorithms like mCADRE, INIT, GIMME, iMAT, PRIME all make use of quantifications of transcripts to de-

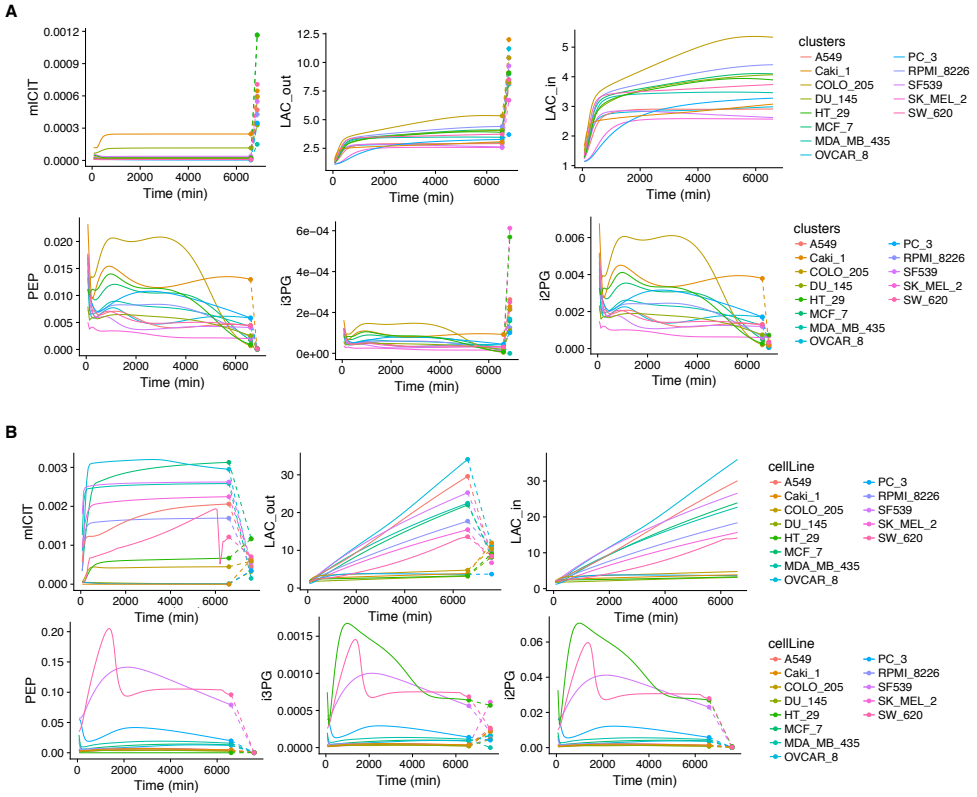

Figure 28: Behavior of cell line models obtained integrating FPKM values (A) and proteomics data (B)

fine which subset of all intracellular metabolic reactions are actually active in a cell/tissue. The reactions that, according to the gene expression profile, are considered inactive, are instead removed from the network.
The activity or inactivity of a reaction is called, with different rules according to the different algorithms, considering that a low expression of a transcript is most probably correlated to a low concentration of functional enzyme, and thus to a limited catalytic power. It can be seen, then, that the assumptions used by these methods very much resemble those that we have just specified. Besides the fact that each of these algorithms exploits different biological hypotheses, all of them assume, more or less explicitly, that the transcript abundance is a good predictor of the level of the functional protein actually present in a cell/tissue.

It is then interesting to notice how these similar theoretical considerations produce different model outcomes. When the aforementioned algorithms are used to create tissue or cell line specific genome scale models (like in [119, 147, 279, 3, 210, 41, 291]) different subsets of reactions are considered inactive and thus removed from the network. These differences however mostly concern reactions at the periphery of the network, while central metabolic pathways are normally retained in all models.
Some works [119, 147] have compared how these algorithms influenced the

resulting flux distribution of the model, computed with FBA. If the reconstruction of GS models is based only on gene or protein expression data and if reaction fluxes are not further constrained, or semi-constrained (following the same semantics proposed in [147]) with metabolomics data, still the various cell line models show significant differences among their steady state flux distributions. Moreover, these differences affect reaction fluxes both in peripheral areas of the network, where many more reactions are called inactive, both in the central metabolic pathways that all the different model retain.
Differently from these GS constraint-based models, the kinetic model we are using represents part of the core metabolic pathways, while all more peripheral reactions are neglected. As the network of central metabolic pathways is represented in both modeling approaches, it is interesting to try to see how these different approaches (kinetic vs constraint-based), together with the different strategies of gene expression data integration, describe the behavior of these common phenomena.

The majority of works in the literature consider that kinetic models, when compared to constraint-based models, display an expanded predictive power, despite the fact that they describe a smaller network. If we adjust the v_{max} parameters considering exclusively the assumption of the direct proportionality between transcripts and enzymes (which, again, is very similar to what is assumed by integration methods applied to constraint based models), we might than agree that, in view of the higher level of detail a kinetic model possesses, the behavior of the recast model would outperform a constraint based description of the system. If this is a fairly acceptable statement, it is then curious to witness the fact that all our recast kinetic models, even in the presence of changes of high magnitude in the value of the parameters, still display dynamic behaviors that tend to converge to the same steady state. In order to appreciate this, we need to observe the distribution of the fluxes, and not just the concentrations. In fact, if two models display the same fixed intracellular concentration at the steady state, this does not necessarily mean that their flux distributions are identical. For this reason we registered the time dependent behavior of the fluxes through the simulation. The values of the fluxes were scaled for the total cell volume and the time-dependent total number of cells. Figure 29 reports what was observed, specifically with integration of the proteomics data, which displayed the larger variability in the curves of the concentrations over time. While we can see that during the transient phase some of the lines(fluxes) tend to diverge, ultimately all of them converge to steady state values that are indeed very close.

This finding then would induce to reason, on one side, on the predictive power of bottom-up kinetic models, and on the other, on the actual accuracy of algorithms for the integration of gene expression data in constraint-based models. Indeed, also the authors of the work in [119], reach the conclusion that the transcriptome seems to be a modest predictor of metabolic fluxes. In conclusion, the predictive capabilities of the data integration approach

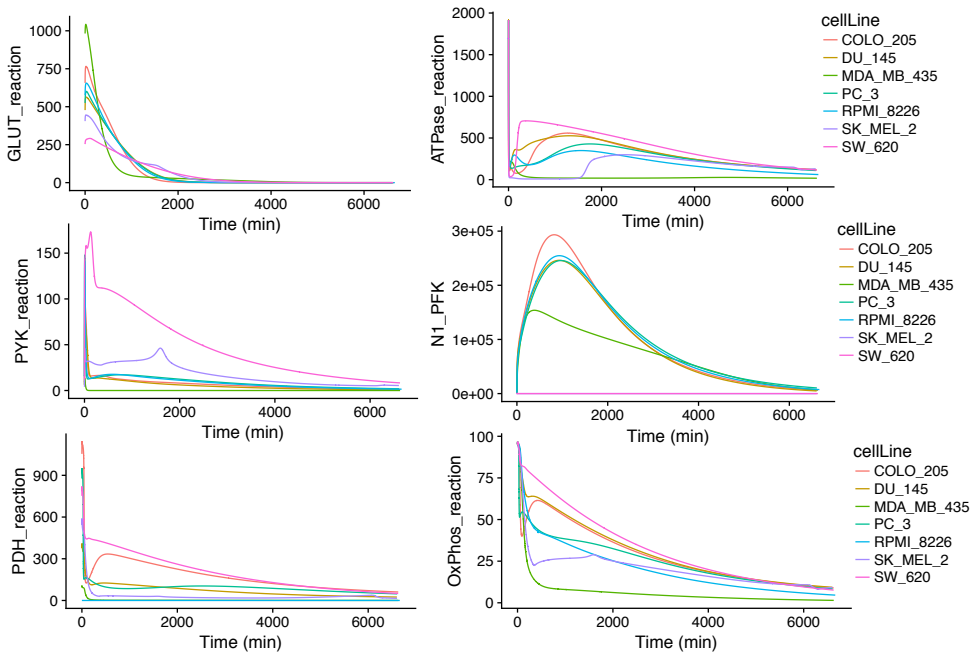

Figure 29: Variation in reaction fluxes across the simulation

we proposed should be more carefully assessed with the additional experiments we are planning. Our method, however, may offer a different point of observation and be used as a complementary tool to support predictions made with other approaches.

3 REPRESENTING METABOLIC HETEROGENEITY AT THE SINGLE CELL LEVEL

We show here how the Symmetric Stochastic Nets presented in chapter 3 and the procedure to recast parameters can be combined to represent intra-tumor metabolic heterogeneity at the single cell level. The gene expression values, in the form of FPKMs, that we have used so far are produced from pooled samples of cells. For each gene reported in the data, its FPKM quantification thus reflects the mean value across the overall sampled cell population. If we integrate these data, the model can just represent the average behavior of the population. If we were interested to grasp more subtle differences in the metabolic traits present in the population we could use the gene expression levels of a sub-portion of the whole tumor, or the FPKMs measured at a single cell level. Data of this kind are difficult and expensive to produce. However, thanks to the interest they have captured, their availability in online repositories is growing.

A dataset with single cell gene expression levels in breast cancer tissues was produced in a recent work [26]. The dataset contained 11 samples from 11 breast cancer patients. For each sample a variable number of single cells

were sequenced. Interestingly, in the supplementary material of the paper the authors report a classification of the cell types based on a computational approach that takes as input gene-expression inferred Copy Number Variants (CNV).

With these data in hand, we could build a population-based model representative of the in-vivo, patient-specific composition of cells in the tumor. Starting again from the PDAC model used so far, we used our recasting procedure to create an a new cell-specific model for each cell in the gene expression dataset. Following our intentions, we we did not want to simulate these single cell models independently, in parallel. Instead, the models were combined in a multi-scale framework, in which extracellular metabolites represent resources that are shared among all cells. The time dependent concentration profiles of extracellular variables, can thus account for the simultaneous processes of uptake and release happening at the level of all single cells. A scenario where some items are subdivided to follow some parallel and similar processes can be represented graphically taking advantage of the properties of Stochastic Symmetric Nets presented in section 3. If we draw an SPN describing the system of reactions inside one single cell, and in addition we define a color class, we obtain an SSN where each cell is associated to a specific color. Hence, all the transitions in the model can be described with transition instances, whose firing rate reflects the differences in the recast kinetic parameters in the population. Once the SSN was built, a complete system of ODEs could be derived and solved. The curves obtained for some of the metabolites in the system are shown in figures 30 31.

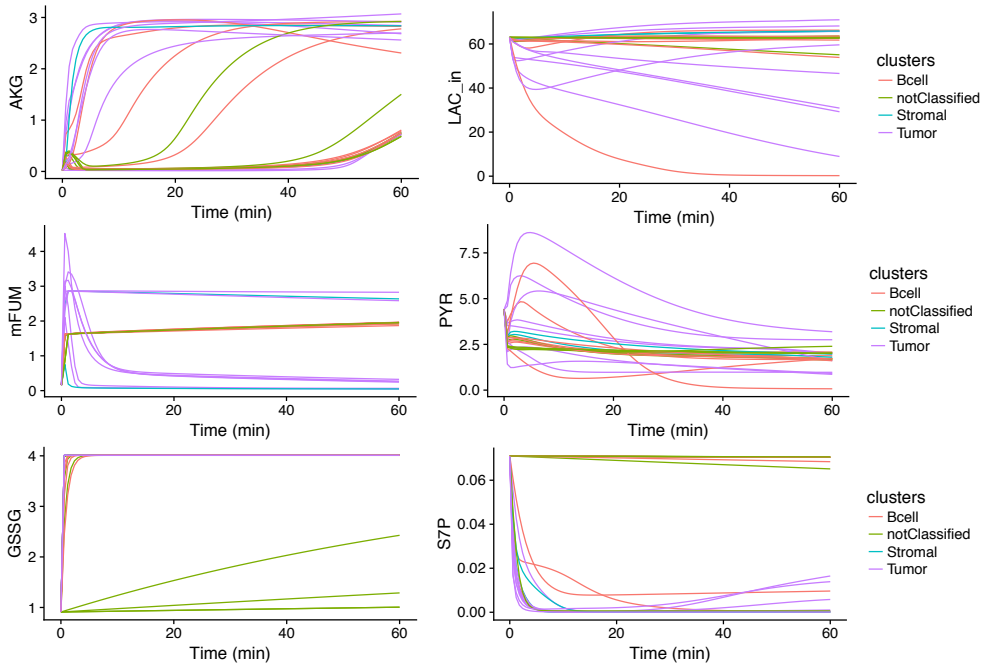

Figure 30: The dynamic behavior of intracellular metabolites concentrations shows differences among single cells

Looking at the figures, a few aspects can be visually noticed:

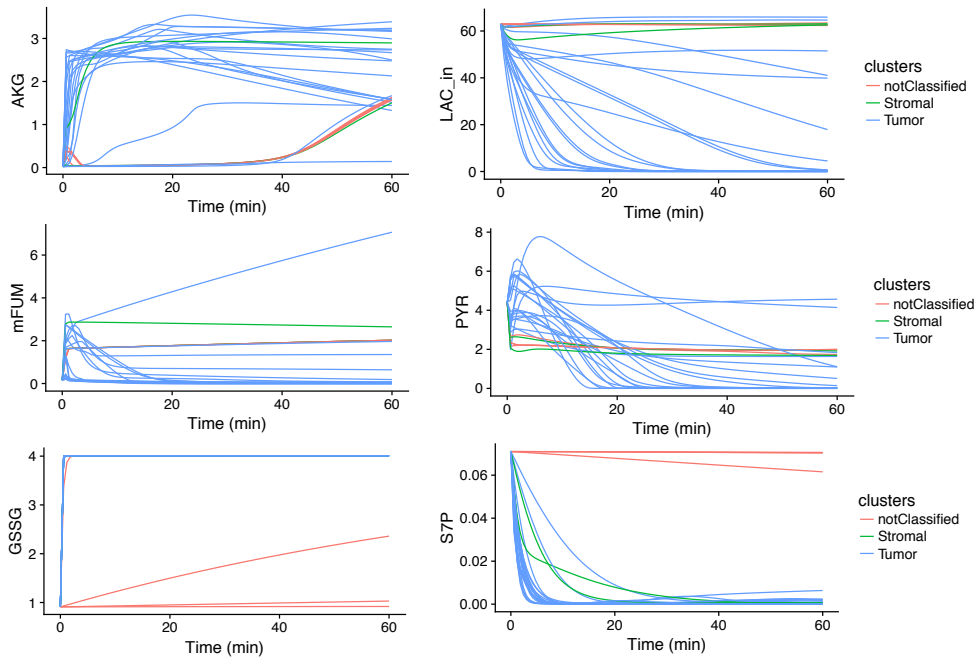

Figure 31: The dynamic behavior of intracellular metabolites concentrations shows differences among single cells

- A classification based on the trajectories able to resemble the classification proposed in [26] seems to be hardly predictive.
- The metabolites that are more central in the system, like LAC and PYR, do not allow one to discriminate any (two or more) clusters of different behavior.
- A few metabolites that are more marginal in the reaction network present bistabilities, like for alpha-Ketoglutarate (AKG), or clusters with radically different behaviors, like mitochondrial fumarate (mFUM), oxidized glutathione (GSSG) and sedoheptulose 7-phosphate (S7P)
- Intriguingly GSSG, a metabolite that, together with its reached form GSH, reflects accurately the redox state of the system, is able to isolate with high specificity the group of non-classified cells. We cannot trace any additional conclusion on the nature of these cells, however an interesting research question could aim to explore if these cells present any commonalities

Independently from the visual aspect of the figure, we decided to perform a classification to obtain a quantitative measurement of our ability to classify the cells with respect to their time course profiles. First of all we calculated the distances among all the curves, for all the cells, using Dynamic Time Warping as a distance measurement. We obtained a distance square matrix with as many entries as the number of cells. We then used the Weka software [60] using this matrix as input and

the algorithm J-48 as a classifier. In the case of the cells in figure 30 we obtained correctly classified instances 43.2432% vs 56.7568% incorrectly classified instances, while for the cells in figure 31 the correctly classified instances were 88.4615% vs 11.5385% incorrectly classified instances. Tables 3 and 4 report the confusion matrices as returned by Weka software.

classified as	Tcell	Bcell	Tumor	notClassified
Tcell	3	3	3	0
Bcell	3	3	3	0
Tumor	0	4	10	1
notClassified	1	2	1	0

Table 3: Confusion matrix output of the classification task performed on the data in figure 30

classified as	Tumor	notClassified	Stromal
Tumor	19	0	1
notClassified	0	4	0
Stromal	1	1	0

Table 4: Confusion matrix output of the classification task performed on the data in figure 31

5 CONCLUSIONS AND FUTURE PERSPECTIVES

In this thesis we have presented a summary of a 3-years work that focused on computational models of cancer metabolism. We started describing the difficulties and challenges that modelers are currently facing in this field and then we illustrated the computational methods we developed to try to overcome some of these obstacles. All our work was accompanied by the study of stochastic processes and their representation with Stochastic Petri nets (SPNs), a high level graphical formalism based on Petri Nets. Knowing that the behavior of a generic biochemical system can be described by stochastic processes like Continuous Time Markov Chains (CTMCs), we showed how these can be effectively represented with SPNs. The specific properties linked to both the structure and the behavior of SPN were continuously exploited for the methods we developed. More in specific, we saw that if, as customary, the fluid approximation is introduced and the behavior of the system is described with Ordinary Differential Equations (ODEs), then the full system of ODEs can be automatically derived from the SPN representation of the system under study.

MODEL INDETERMINATION In a first research thread we focused on a new approach to deal with biochemical models with lacking kinetic information. Although our premises and the methods we proposed can be referred to any metabolic system in which data is insufficient, we instantiated our work around the specific topic of modeling cancer metabolism. We thus described that for the scenario of human cancers, data availability has a crucial influence on the construction of bottom-up kinetic models of metabolism. We explained how data scarcity causes indetermination in the structure and the parameters of mechanism-based kinetic models. This indetermination is typically overcome with techniques of reverse engineering (RE) and parameter estimation (PE), which, however, inevitably fail when the space of feasible parameter values is not enough constrained by the data. Ultimately, all of these factors cause models to be poorly representative of the metabolic phenomena under study, and hence of difficult use.
Acknowledging this context, we developed a method which might be exploited in alternative or in addition to the classical RE and PE approaches. In our method we proposed to discriminate the biochemical events in the system, corresponding to reactions in a metabolic network and to transitions in an SPN, into two groups, according to their level of indetermination. We defined as determined reactions and determined transitions the reactions/transitions for which we assume to be able to retrieve, from the literature, accurate structure and kinetic parameters. On the other hand, we defined undetermined reactions and undetermined transitions those reactions/tran-

sitions for which we assume that the information available prevents their accurate mathematical description. While the velocity of undetermined reactions can be computed with a fully parameterized mathematic expression, we compute the fluxes of undetermined reactions with approximate rate laws, whose parameters are replaced by time varying coefficients. If, on one side, these coefficients completely loose the biological meaning associated to otherwise constant parameter values, they are here introduced as a way to condense into one specific value the influence that any unknown factor have on these reactions. We then propose that the empirical knowledge of the experimentalists allows us to define some boundary conditions for the coefficients as well as an objective function that generically describes the biological behavior we intend to reproduce. When all of these components have been defined, we showed how the behavior of the system can be reproduced exploiting iterative processes of optimization.
The procedure by which the method can be used for real-case applications was demonstrated in chapter 4, where we proposed to apply our approach to investigate specifically the behavior of three glycolytic enzymes, HK,PFK and PK. According to many evidences reported in the literature, these play a key role in metabolic alterations found in cancer metabolism, but at the same time their behavior in cancer has not been fully understood. In order to clarify the validity of the approach we proposed, as future works we are considering some additional experiments. The performances of any optimization process are highly dependent on the objective function, on the constraints that define the space of feasible solutions, and on the algorithm used to explore it. Depending on the specific model, objective functions with different biological meaning will be tested. We will also investigate which available biological data can be effectively used to adapt the model to the specific situation under study. Either data obtained with transcriptomics, proteomics and metabolomics, as well as reaction fluxes and thermodynamic information are all possible candidates. These data will provide a fundamental support to define the initial conditions of the model, to inspire the choice of the objective function and to set the constraints for the optimization. Finally the performances of different optimization algorithms will be tested.

DATA INTEGRATION In a second research thread we explored a method to integrate gene expression data in the form of Fragments per Kilobase Mapped Reads (FPKMs) to generate new condition-specific mechanistic models of metabolism. The approach bases its validity on the assumption that FPKMs, despite the occurrence of many post-transcriptional and post-translational processes, still retain some useful quantitative information regarding the abundance of the enzyme in the cell/tissue. The approach requires as input a kinetic model fully parametrized for a reference condition, and FPKM values for both the reference and the desired condition. Applying a simple proportion, the kinetic parameters that are directly related to enzyme concentrations, like v_{max} parameters, can be recast to represent the new condition.

As we presented in chapter 4 the procedure was used to convert a pancreatic cancer specific kinetic model to kinetic models representing, respectively, NCI-60 cell lines, breast cancer patients and breast cancer single cells. The results of the simulation were compared to available exometabolomics data (for NCI-60 cell lines), and were analyzed to assess their utility for classification tasks. The availability of proteomics data for a subgroup of the NCI-60 cell line panel allowed us to recast the parameters with protein levels instead of FPKMs, and to draw interesting comparisons between the resulting outputs. This research project is however still ongoing. For the near future, we plan to expand our analysis of the NCI-60 cell line models with a few additional experiments. In specific we are planning to:

- Assess how much our results are affected by the network reconstruction we choose to use (Recon 2 [241] vs 2.2 [233] vs 3D [18]), considering that the stoichiometry of the system as well as GPR rules show some differences.

- Use publicly available NCI-60 cell doubling times (available at https://dtp.cancer.gov/) to reproduce experimental growth curves, considering that the number of cells in time N(t) can be retrieved from the doubling time (dTime) using the following equations:

$$N(t) = N(0)e^{\mu * t}, \quad \mu = \frac{\ln(2)}{dTime}$$

  The curves obtained plotting N(t) can be compared with the curves produced by our models. We highlight the fact that, as in our model the growth rate varies with time ($\mu(t)$) the comparison among curves is easier and more effective than the comparison among growth rates.

- Use as a source of comparison the NCI-60 cell line specific genome scale constraint based models that can be found in the literature. These were originally built integrating gene expression data with different algorithms. We plan to try to compare the intracellular flux distributions calculated with these approaches with the reactions fluxes displayed by our models once the steady state is reached.
  The process of comparison between the two modeling approaches is not straightforward, and needs to be guided by a few considerations:
  - the size of our kinetic model is much smaller than these genome scale network reconstructions, so we should assume that the parameters in kinetic equations already take into consideration the effects of all the reactions that are excluded from the kinetic model.
  - FBA computes the flux distribution at the intracellular steady state. The intracellular steady state is characterized by null differentials for intracellular concentrations ($\frac{dm}{dt} = 0$), a constant flux through the biomass reaction, and thus by a constant growth rate. FBA then assumes that the cell population is in a phase of exponential growth. The plots of the original PDAC model [186]

instead do not directly show the intracellular steady state, as the concentrations of intracellular metabolites are not scaled by the total cell volume. Taking into consideration the average cell volume and the time-dependent number of cells, growing from the initial $n = 5000$, we will adjust the values of both intracellular fluxes and concentrations.

- Assess if our recast model can be used as a first draft parametrization that is later improved with local search PE algorithms that do not require excessive computational costs.
  The exometabolomics dataset produced in [72] was here used to try to validate the accuracy of our recasting approach. In a future step, instead, we will exploit it in a parameter estimation process, analogous to the one proposed in the original paper [186], where the simulated steady state metabolite concentrations will be fitted to the experimental values. Both global and local PE approaches will be used and the outcomes compared. Specifically, we will test if the recast parameters can be considered as a good starting point that helps local search approaches to correctly identify the global optimum.

REPRESENTING METABOLIC HETEROGENEITY In this thesis we showed how the distinctive features of Stochastic Symmetric nets (SSN) can be used to represent heterogeneity in metabolic systems at different levels. As a first example, we described that the introduction of color classes allows one to represent multiple enzyme isoforms with a much more compact description, that still retains all the information present in the system. If, for instance, we were willing to specify multiple isoforms for all the enzymes in a medium-scale metabolic network of reactions, we can understand that a SSN representation, from which the complete ODEs system can be directly derived, can be of great help.
As a further application of SSNs, we moved our focus to the heterogeneity of metabolic traits among single cells in a population. Importantly, in our examples we considered a virtual tumor composed of cells with different metabolic characteristics that are located in a common environment and share the same pool of nutrients. In such a situation, each different cellular metabolic phenotype can be assigned a color that identifies a phenotype-specific set of kinetic parameters. The overall virtual tumor can thus be represented by one single network of metabolic reactions with a color class that retains the heterogeneity of parameters, and hence of behaviors, of the system. The majority of our research efforts are directed, and in the future will most likely be directed, on the study of intra-population metabolic heterogeneity, both in the context of cancer and of microbial communities. At the present we are working on an agent-based multi scale model that describes how the composition of a population of cells evolves depending on the metabolic traits that it hosts. A graphical scheme of its structure can be found in the appendix A. We believe that these types of framework will

provide a powerful tool to investigate eco-evolutionary metabolic dynamics in multicellular systems.

With this goal in mind, the predictions of these frameworks can be reinforced by approaches of single cell data integration like the one we just presented. Other approaches proposed recently [32] seek to integrate single cell gene expression data into constraint-based models. Analogously to the comparison between kinetic and constraint-based models of the NCI-60 cell lines, we are then planning to compare the outcomes of kinetic and constraint-based models of a tumor that is heterogeneous in its metabolic aspects. Using the same RNA-Seq datasets, the same network reconstruction(Recon 2.2 [233]) and the same rules of integration, we will compare the steady state analyses among the results obtained with the two approaches. Overall, we believe that this type of comparison can help uncover strengths and weaknesses of the different modeling approaches and thus to support modelers in the choice of the modeling formalism that best suits their research goals.

A APPENDIX

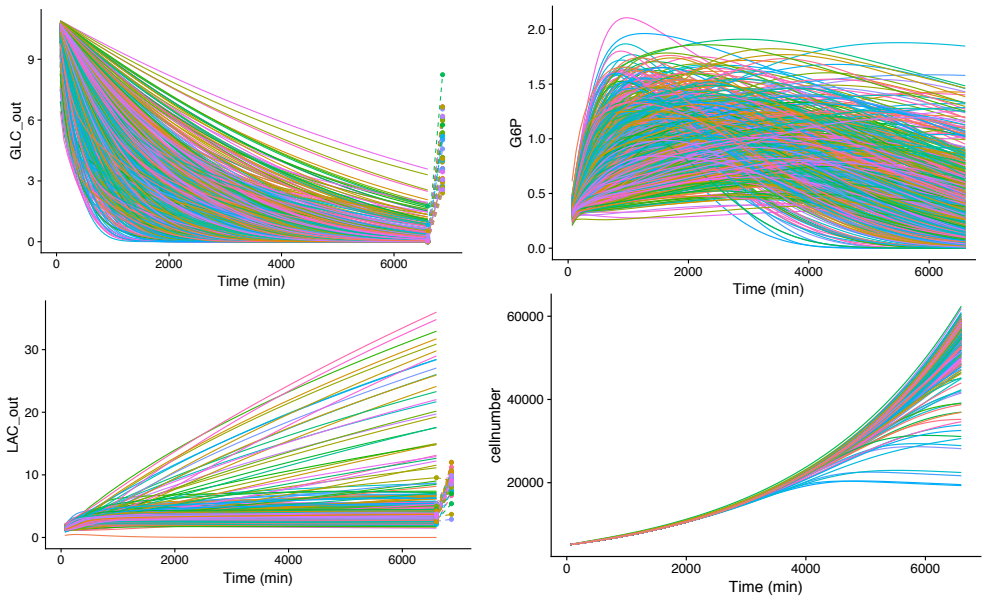

Figure 32: Outcomes of all 675 recast cell line models. Each cell line is plotted with a different color. The dashed lines represent the match between the last time point in the simulation and the concentration data retrieved from [72]. As one can notice, metabolite concentrations were available just for some a subgroup of 675 cell lines. The plots show how the variability in glucose consumption is reflected on variable growth profiles.

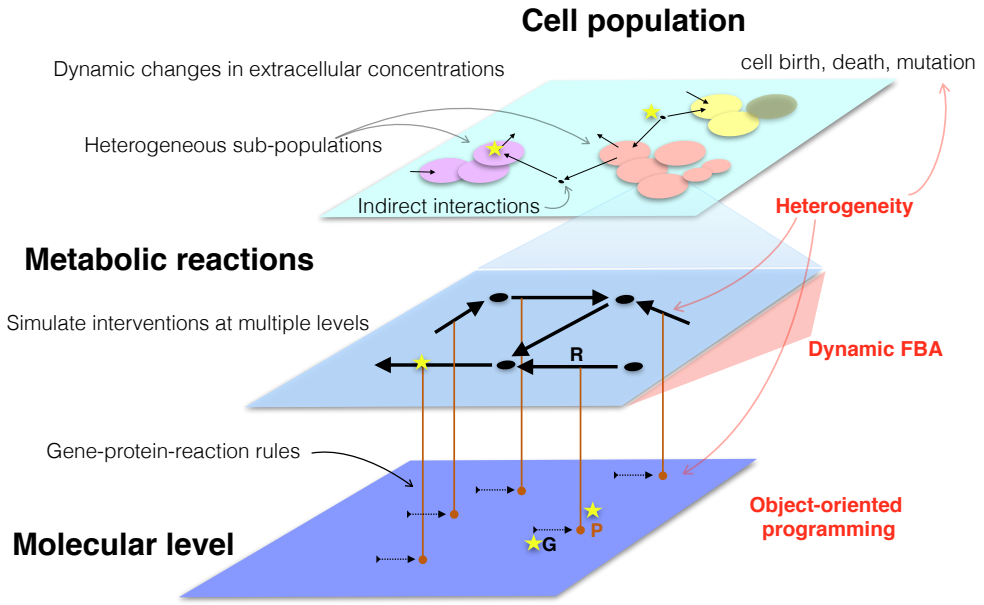

Figure 33: Graphical scheme of a multi-scale agent-based model representing intra-population metabolic heterogeneity. In this modeling framework we are interested to understand how the composition of the population over time is conditioned by metabolic processes happening both at the intracellular level and in the extracellular space. A multi-scale model, in specific, allows one to link the events occurring at higher scales to the molecular characteristics of each cell in the population, and thus to test different interventions at different levels

BIBLIOGRAPHY

[1] Appendix A: Review of Nonlinear Optimization. In *Learning from Data*, pages 507–513. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[2] Rasmus Agren, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS computational biology*, 8(5):e1002518, 2012.

[3] Rasmus Agren, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, 8(5), 2012.

[4] Tero Aittokallio and Benno Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, 7(3):243–255, 2006.

[5] Dimitrios Anastasiou. Tumour microenvironment factors shaping the cancer metabolism landscape. *British journal of cancer*, 116(3):277, 2017.

[6] Maike K Aurich, Ronan M T Fleming, and Ines Thiele. A systems approach reveals distinct metabolic strategies among the NCI-60 cancer cell lines. *PLoS computational biology*, 13(8):e1005698, aug 2017.

[7] J. Babar, M. Beccuti, S. Donatelli, and A. S. Miner. Greatspn enhanced with decision diagram data structures. In *Proceedings of 31st Int. Conf. of Applications and Theory of Petri Nets*, pages 308–317. IEEE Computer Society, June 2010.

[8] Samuel Bandara, Johannes P. Schlder, Roland Eils, Hans Georg Bock, and Tobias Meyer. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Computational Biology*, 5(11), 2009.

[9] Brandon E. Barker, Narayanan Sadagopan, Yiping Wang, Kieran Smallbone, Christopher R. Myers, Hongwei Xi, Jason W. Locasale, and Zhenglong Gu. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. *Computational Biology and Chemistry*, 59:98 – 112, 2015. Advances in Systems Biology.

[10] Scott A Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.

[11] Aharon Ben-Tal and Michael Zibulevsky. Penalty/barrier multiplier methods for convex programming problems. *SIAM J. on Optimization*, 7(2):347–366, February 1997.

[12] Bryson D. Bennett, Elizabeth H. Kimball, Melissa Gao, Robin Osterhout, Stephen J. Van Dien, and Joshua D. Rabinowit. Absolute metabolite concentrations and implied enzyme active site occupancy in escherichia coli. *Nature Chemical Biology*, 5(8):593–599, 2009.

[13] Sara Berthoumieux, Daniel Kahn, Hidde de Jong, and Eugenio Cinquemani. Structural and practical identifiability of approximate metabolic network models. *IFAC Proceedings Volumes*, 45(16):1719–1724, jul 2012.

[14] Anna Blazier and Jason Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*, 3:299, 2012.

[15] Brett a Boghigian, Hai Shi, Kyongbum Lee, and Blaine a Pfeifer. Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design. *BMC systems biology*, 4:49, 2010.

[16] Aarash Bordbar, Douglas Mccloskey, Daniel C Zielinski, Nikolaus Sonnenschein, Neema Jamshidi, Bernhard O Palsson Correspondence, and Bernhard O Palsson. Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *Cell Systems*, 1:283–292, 2015.

[17] Uwe Borgmann, Thomas W. Moon, and Keith J. Laidler. Molecular Kinetics of Beef Heart Lactate Dehydrogenase. *J . Biol. Chem. Annu. Rev. Bio- Methods Enzymol. J . Biol. Davis, B. J . Ann. N . Y. Acad. Sci. Science J. Arch. Biochem*, 13(25):235–258, 1974.

[18] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, et al. Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272, 2018.

[19] Sascha Bulik, Sergio Grimbs, Carola Huthmacher, Joachim Selbig, and Hermann G. Holzhutter. Kinetic hybrid models composed of mechanistic and simplified enzymatic rate laws - a promising method for speeding up the kinetic modelling of complex metabolic networks. *FEBS Journal*, 276(2):410–424, 2009.

[20] Sascha Bulik, Sergio Grimbs, Carola Huthmacher, Joachim Selbig, and Hermann G Holzhütter. Kinetic hybrid models composed of mechanistic and simplified enzymatic rate laws–a promising method for speeding up the kinetic modelling of complex metabolic networks. *The FEBS journal*, 276(2):410–424, 2009.

[21] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):D459–D471, jan 2014.

[22] P Cazzaniga, C. Daminani, D Besozzi, R. Colombo, M.S. Nobile, D. Gaglio, D. Pescini, S Molinari, G. Mauri, L. Alberghina, and M. Vanoni. Computational strategies for a system-level understanding of metabolism. *Metabolites*, 4:1034–1087, 2014.

[23] Paolo Cazzaniga, Chiara Damiani, Daniela Besozzi, Riccardo Colombo, Marco S Nobile, Daniela Gaglio, Dario Pescini, Sara Molinari, Giancarlo Mauri, Lilia Alberghina, and Marco Vanoni. Computational Strategies for a System-Level Understandingof Metabolism. (August):1034–1087, 2014.

[24] Christophe Chassagnole, Naruemol Noisommit-rizzi, Joachim W Schmid, Klaus Mauch, and Matthias Reuss. Dynamic Modeling of the Central Carbon Metabolism of Escherichia coli. 2002.

[25] G. Chiola, C. Dutheillet, G. Franceschinis, and S. Haddad. A symbolic reachability graph for coloured Petri nets. *Theoretical Computer Science B (Logic, semantics and theory of programming)*, 176(1-2):39–65, 1997.

[26] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8:15081, 2017.

[27] Marcin Cieślik and Arul M. Chinnaiyan. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93–109, dec 2017.

[28] Angela Cintolesi, James M Clomburg, Venetia Rigou, Kyriacos Zygourakis, and Ramon Gonzalez. Quantitative analysis of the fermentative metabolism of glycerol in escherichia coli. *Biotechnology and Bioengineering*, 109(1):187–198, 2012.

[29] Rafael S Costa, Andras Hartmann, and Susana Vinga. Kinetic modeling of cell metabolism for microbial production. 219:126–141, 2016.

[30] Rafael S Costa, André Veríssimo, and Susana Vinga. Ki MoSys: a web-based repository of experimental data for KInetic MOdels of biological SYStems. *BMC Systems Biology*, 8(1):85, dec 2014.

[31] Chiara Damiani, Marzia Di Filippo, Dario Pescini, Davide Maspero, Riccardo Colombo, and Giancarlo Mauri. popfba: tackling intratumour heterogeneity with flux balance analysis. *Bioinformatics*, 33(14):i311–i318, 2017.

[32] Chiara Damiani, Davide Maspero, Marzia Di Filippo, Riccardo Colombo, Dario Pescini, Alex Graudenzi, Hans Victor Westerhof, Lilia Alberghina, Marco Ercole Vanoni, and Giancarlo Mauri. Integration of single-cell rna-seq data into metabolic models to characterize tumour cell populations. *bioRxiv*, page 256644, 2018.

[33] Chiara Damiani, Davide Maspero, Marzia Di Filippo, Riccardo Colombo, Dario Pescini, Alex Graudenzi, Hans Victor Westerhof, Lilia Alberghina, Marco Ercole Vanoni, and Giancarlo Mauri. Integration of single-cell RNA-seq data into metabolic models to characterize tumour cell populations. *bioRxiv*, page 256644, jan 2018.

[34] Gundián M de Hijas-Liste, Edda Klipp, Eva Balsa-Canto, and Julio R Banga. Global dynamic optimization approach to predict activation in metabolic pathways. *BMC systems biology*, 8:1, 2014.

[35] Aurelio A de Los Reyes V, Eunok Jung, and Yangjin Kim. Optimal control strategies of eradicating invisible glioblastoma cells after conventional surgery. *Journal of the Royal Society, Interface / the Royal Society*, 12(106):20141392–, 2015.

[36] R. J. DeBerardinis and N. S. Chandel. Fundamentals of cancer metabolism. *Science Advances*, 2(5):e1600200–e1600200, may 2016.

[37] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(Database issue):D344–50, jan 2008.

[38] Rodrigo Diaz-Ruiz, Michel Rigoulet, and Anne Devin. The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1807(6):568–576, jun 2011.

[39] Rodrigo Diaz-Ruiz, Michel Rigoulet, and Anne Devin. The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1807(6):568–576, jun 2011.

[40] A Doi, S Fujita, H Matsuno, M Nagasaki, and S Miyano. Constructing biological pathway models with hybrid functional petri nets. *Studies in health technology and informatics*, 162:92–112, 2011.

[41] Sonia C Dolfi, Leo Li-Ying Chan, Jean Qiu, Philip M Tedeschi, Joseph R Bertino, Kim M Hirshfield, Zoltán N Oltvai, and Alexei Vazquez. The

metabolic demands of cancer cells are coupled to their size and protein synthesis rates. *Cancer & metabolism*, 1(1):20, nov 2013.

[42] Bin Du, Daniel C Zielinski, Erol S Kavvas, Andreas Dräger, Justin Tan, Zhen Zhang, Kayla E Ruggiero, Garri A Arzumanyan, and Bernhard O Palsson. Evaluation of rate law approximations in bottom-up kinetic models of metabolism. *BMC Systems Biology*, pages 1–15, 2016.

[43] G A Dunaway, T P Kasten, T Sebo, and R Trapp. Analysis of the phosphofructokinase subunits and isoenzymes in human tissues. *The Biochemical journal*, 251(3):677–83, 1988.

[44] Jing Fan, Jurre J Kamphorst, Robin Mathew, Michelle K Chung, Eileen White, Tomer Shlomi, and Joshua D Rabinowitz. Glutamine-driven oxidative phosphorylation is a major atp source in transformed mammalian cells in both normoxia and hypoxia. *Molecular systems biology*, 9(1):712, 2013.

[45] David A Fell and Andreas Wagner. The small world of metabolism. *Nature biotechnology*, 18(11):1121, 2000.

[46] W. Feller. *An Introduction to Probability Theory*, volume 1. Wiley, New York, NY, 1968.

[47] Ciarán P Fisher, Nicholas J Plant, J Bernadette Moore, and Andrzej M Kierzek. QSSPN: dynamic simulation of molecular interaction networks describing gene regulation, signalling and whole-cell metabolism in human cells. *Bioinformatics*, 29:3181–90, 2013.

[48] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(501):501, 2011.

[49] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7:501, jun 2011.

[50] L. L. Fonseca, C. Sanchez, H. Santos, and E. O. Voit. Complex coordination of multi-scale cellular responses to environmental stress. *Molecular BioSystems*, 7(3):731–741, 2011.

[51] Neil S. Forbes, Adam L. Meadows, Douglas S. Clark, and Harvey W. Blanch. Estradiol stimulates the biosynthetic pathways of breast cancer cells: Detection by metabolic flux analysis. *Metabolic Engineering*, 8(6):639–652, nov 2006.

[52] Moles C. G., Mendes P., and Banga J. R. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467–2474, 2003.

[53] Yash Garg and Silvestro Roberto Poccia. On the effectiveness of distance measures for similarity search in multi-variate sensory data: Effectiveness of distance measures for similarity search. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 489–493. ACM, 2017.

[54] Francesco Gatto, Heike Miess, Almut Schulze, and Jens Nielsen. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Scientific Reports*, 5(JANUARY):10738, 2015.

[55] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

[56] Toni Giorgino et al. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31(7):1–24, 2009.

[57] Ramon Grima and Santiago Schnell. How reaction kinetics with time-dependent rate coefficients differs from generalized mass action. *ChemPhysChem*, 7(7):1422–1424, 2006.

[58] H Gutfreund, R Cantwell, C H McMurray, R S Criddle, and G Hathaway. The kinetics of the reversible inhibition of heart lactate dehydrogenase through the formation of the enzyme-oxidized nicotinamide-adenine dinucleotide-pyruvate compounds. *The Biochemical journal*, 106(3):683–7, 1968.

[59] S. Haddad and J.M. Couvreur. Towards a General and Powerful Computation of Flows for Parametrized Coloured Nets. In *Proc. 9th Europ. Workshop on Application and Theory of Petri Nets*, Venezia, Italy, June 1988.

[60] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[61] William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, Christopher J. Marx, and Daniel Segrè. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7(4):1104–1115, 2014.

[62] M. Vander Heiden. Targeting cancer metabolism: a therapeutic window opens. *Nature Reviews. Drug discovery*, 10(9):671–684, 2011.

[63] Monika Heiner, David Gilbert, and Robin Donaldson. Petri nets for systems and synthetic biology. In *International School on Formal Methods for the Design of Computer, Communication and Software Systems*, pages 215–264. Springer, 2008.

[64] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N. Mendoza, Anne Richelle, Almut Heinken, Hulda S. Haraldsdóttir, Jacek Wachowiak, Sarah M. Keating, Vanja Vlasov, Stefania Magnusdóttir, Chiam Yu Ng, German Preciat, Alise Žagare, Siu H. J. Chan, Maike K. Aurich, Catherine M. Clancy, Jennifer Modamio, John T. Sauls, Alberto Noronha, Aarash Bordbar, Benjamin Cousins, Diana C. El Assal, Luis V. Valcarcel, Iñigo Apaolaza, Susan Ghaderi, Masoud Ahookhosh, Marouen Ben Guebila, Andrejs Kostromins, Nicolas Sompairac, Hoai M. Le, Ding Ma, Yuekai Sun, Lin Wang, James T. Yurkovich, Miguel A. P. Oliveira, Phan T. Vuong, Lemmer P. El Assal, Inna Kuperstein, Andrei Zinovyev, H. Scott Hinton, William A. Bryant, Francisco J. Aragón Artacho, Francisco J. Planes, Egils Stalidzans, Alejandro Maass, Santosh Vempala, Michael Hucka, Michael A. Saunders, Costas D. Maranas, Nathan E. Lewis, Thomas Sauter, Bernhard Ø. Palsson, Ines Thiele, and Ronan M. T. Fleming. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. oct 2017.

[65] Diana M. Hendrickx, Margriet M. W. B. Hendriks, Paul H. C. Eilers, Age K. Smilde, and Huub C. J. Hoefsloot. Reverse engineering of metabolic networks, a critical assessment. *Molecular BioSystems*, 7:511–520, 2011.

[66] Anique Herling, Matthias König, Sascha Bulik, and Hermann-Georg Holzhütter. Enzymatic features of the glucose metabolism in tumor cells. *The FEBS journal*, 278(14):2436–2459, 2011.

[67] J.L. Hjersted and M.A. Henson. Steady-state and dynamic flux balance analysis of ethanol production by Saccharomyces cerevisiae. *IET Systems Biology*, 3(3):167–179, may 2009.

[68] Nguyen Hoang Son, Rafael Costa Isabel Maria de, and Rafael Costa. Exploring optimal objective functions and additional constraints for flux prediction in genome-scale models Examination Committee. 2014.

[69] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. Copasi—a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[70] P. Huber, A.M. Jensen, L.O. Jepsen, and K. Jensen. Towards reachability trees for high-level Petri nets. In G. Rozenberg, editor, *Advances on Petri Nets '84*, volume 188 of *LNCS*, pages 215–233. Springer Verlag, 1984.

[71] Esther Imperlini, Lucia Santorelli, Stefania Orrù, Emanuela Scolamiero, Margherita Ruoppolo, and Marianna Caterino. Mass spectrometry-based metabolomic and proteomic strategies in organic acidemias. *BioMed research international*, 2016, 2016.

[72] Mohit Jain, Roland Nilsson, Sonia Sharma, Nikhil Madhusudhan, Toshimori Kitami, Amanda L Souza, Ran Kafri, Marc W Kirschner, Clary B Clish, and Vamsi K Mootha. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science (New York, N.Y.)*, 336(6084):1040–4, may 2012.

[73] Neema Jamshidi and B. Palsson. Mass action stoichiometric simulation models: Incorporating kinetics and regulation into stoichiometric models. *Biophysical Journal*, 98(2):175–185, 2010.

[74] Neema Jamshidi and Bernhard Ø. Palsson. Mass action stoichiometric simulation models: Incorporating kinetics and regulatoin into stoichiometric models. *Biophysical Journal*, 98(2):175–185, 2010.

[75] Livnat Jerby, Tomer Shlomi, and Eytan Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology*, 6(1):401, 2010.

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

[77] T. K. Kar and Ashim Batabyal. Stability analysis and optimal control of an SIR epidemic model with vaccination. *BioSystems*, 104(2-3):127–135, 2011.

[78] Soetaert [aut] Karline, cre] Thomas, Petzoldt [aut, Setzer [aut] R. Woodrow, and odepack authors [cph]. Solvers for Initial Value Problems of Differential Equations (ODE, DAE, DDE). https://cran.r-project.org/web/packages/deSolve/deSolve.pdf, 2016.

[79] Mark A Keibler, Thomas M Wasylenko, Joanne K Kelleher, Othon Iliopoulos, Matthew G Vander Heiden, and Gregory Stephanopoulos. Metabolic requirements for cancer cell proliferation. *Cancer & Metabolism*, pages 1–16, 2016.

[80] Andreas Keller, Petra Leidinger, Anne Borries, Anke Wendschlag, Frank Wucherpfennig, Matthias Scheffler, Hanno Huwer, Hans-Peter Lenhof, and Eckart Meese. mirnas in lung cancer-studying complex fingerprints in patient's blood cells by microarray experiments. *BMC cancer*, 9(1):1, 2009.

[81] Ruchir A. Khandelwal, Brett G. Olivier, Wilfred F. M. Röling, Bas Teusink, and Frank J. Bruggeman. Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE*, 8(5):e64567, may 2013.

[82] Tahmineh Khazaei, Alison P McGuigan, and Radhakrishnan Mahadevan. Ensemble modeling of cancer metabolism. *Frontiers in physiology*, 3:135, 2012.

[83] Hoon Kim, John Watkinson, and Dimitris Anastassiou. Biomarker discovery using statistically significant gene sets. *Journal of computational biology : a journal of computational molecular cell biology*, 18(10):1329–38, oct 2011.

[84] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[85] Steffen Klamt, Georg Regensburger, Matthias P Gerstl, Christian Jungreuthmayer, Stefan Schuster, Radhakrishnan Mahadevan, Jürgen Zanghellini, and Stefan Müller. From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. *PLoS computational biology*, 13(4):e1005409, apr 2017.

[86] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[87] Willem H. Koppenol, Patricia L. Bounds, and Chi V. Dang. Otto Warburg's contributions to current concepts of cancer metabolism. *Nature Reviews Cancer*, 11(5):325–337, may 2011.

[88] Michael R. Kosorok and Shuangge Ma. Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *Ann. Statist.*, 35(4):1456–1486, 08 2007.

[89] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[90] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[91] Antonija Kreso, Catherine A O'Brien, Peter van Galen, Olga I Gan, Faiyaz Notta, Andrew MK Brown, Karen Ng, Jing Ma, Erno Wienholds, Cyrille Dunant, et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science*, 339(6119):543–548, 2013.

[92] Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, 28(5):1–26, 2008.

[93] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, page gkw377, 2016.

[94] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L.

Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

[95] Robert J Kurman and Ie-Ming Shih. The origin and pathogenesis of epithelial ovarian cancer-a proposed unifying theory. *The American journal of surgical pathology*, 34(3):433, 2010.

[96] Vladimir Kuznetsov, Hwee Kuan Lee, Sebastian Maurer-Stroh, Maria Judit Molnár, Sandor Pongor, Birgit Eisenhaber, and Frank Eisenhaber. How bioinformatics influences health informatics: usage of biomolecular sequences, expression profiles and automated microscopic image analyses for clinical needs and public health. *Health Information Science and Systems*, 1(1):1–18, 2013.

[97] R.-S. Wang L. Chen and X.-S. Zhang. *Biomolecular Networks: Methods and Applications in Systems Biology*. Wiley, New York, NY, 2009.

[98] Harri Lähdesmäki, Llya Shmulevich, Valerie Dunmire, Olli Yli-Harja, and Wei Zhang. In silico microdissection of microarray data from heterogeneous cell populations. *BMC bioinformatics*, 6:54, 2005.

[99] Anna Kane Laird. Dynamics of tumour growth. *British journal of cancer*, 18(3):490, 1964.

[100] K. Lange. *Optimization*. Springer, New York, NY, second edition, 2013.

[101] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[102] Jonathan A Ledermann, Christian Marth, Mark S Carey, Michael Birrer, David DL Bowtell, Stan Kaye, Iain McNeish, Amit Oza, Giovanni Scambia, Gordon Rustin, et al. Role of molecular agents and targeted therapy in clinical trials for women with ovarian cancer. *International Journal of Gynecological Cancer*, 21(4):763–770, 2011.

[103] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217, nov 2008.

[104] Y Lee, A Miron, R Drapkin, MR Nucci, F Medeiros, A Saleemuddin, J Garber, C Birch, H Mou, RW Gordon, et al. A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *The Journal of pathology*, 211(1):26–35, 2007.

[105] Bei Li, Lingxiao Jiang, Qifeng Song, Junxin Yang, Zeliang Chen, Zhaobiao Guo, Dongsheng Zhou, Zongmin Du, Yajun Song, Jin Wang, et al.

Protein microarray for profiling antibody responses to yersinia pestis live vaccine. *Infection and immunity*, 73(6):3734–3739, 2005.

[106] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie I Stefan, Jacky L Snoep, Michael Hucka, Nicolas Le Novère, and Camille Laibe. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(1):92, jun 2010.

[107] Pu Li and Quoc Dong Vu. Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC systems biology*, 7:91, sep 2013.

[108] Shuzhao Li, Nadine Rouphael, Sai Duraisingham, Sandra Romero-Steiner, Scott Presnell, Carl Davis, Daniel S Schmidt, Scott E Johnson, Andrea Milton, Gowrisankar Rajam, Sudhir Kasturi, George M Carlone, Charlie Quinn, Damien Chaussabel, A Karolina Palucka, Mark J Mulligan, Rafi Ahmed, David S Stephens, Helder I Nakaya, and Bali Pulendran. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature immunology*, 15(2):195–204, feb 2014.

[109] Xiaobo Li, Xue Gong, Xiaoning Peng, and Sihua Peng. Ssicp: A new svm based recursive feature elimination algorithm for multiclass cancer classification. *International Journal of Multimedial Ubiquitous Engineering*, 2014.

[110] J.G. Liao and Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951, 2007.

[111] JG Liao and Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951, 2007.

[112] Wolfram Liebermeister. Elasticity sampling links thermodynamics to metabolic control. *arXiv preprint arXiv:1309.0267*, pages 1–51, 2013.

[113] Xin Liu and Mahesan Niranjan. State and parameter estimation of the heat shock response system using kalman and particle filters. *Bioinformatics*, 28(11):1501–1507, 2012.

[114] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165:535–550, 2016.

[115] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3):535–550, 2016.

[116] Jason G Lomnitz and Michael A Savageau. Design Space Toolbox V2: Automated Software Enabling a Novel Phenotype-Centric Modeling Strategy for Natural and Synthetic Biological Systems. *Frontiers in genetics*, 7:118, 2016.

[117] Stefano Lonardi, Denisa Duma, Matthew Alpert, Francesca Cordero, Marco Beccuti, Prasanna R Bhat, Yonghui Wu, Gianfranco Ciardo, Burair Alsaihati, Yaqin Ma, et al. Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS Comput Biol*, 9(4):e1003010, 2013.

[118] Doriane Lorendeau, Stefan Christen, Gianmarco Rinaldi, and Sarah-Maria Fendt. Metabolic control of signaling pathways and metabolic auto-regulation.

[119] Daniel Machado and Markus Herrgård. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Computational Biology*, 10(4):e1003580, apr 2014.

[120] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophysical journal*, 83(September):1331–1340, 2002.

[121] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mRNA and protein in complex biological samples. 2009.

[122] Thomas Robert Maltus. *An essay on the principle of population*, volume 1. Cosimo, Inc., 2006.

[123] Ubaldo E. Martinez-Outschoorn, Maria Peiris-Pagés, Richard G. Pestell, Federica Sotgia, and Michael P. Lisanti. Cancer metabolism: a therapeutic perspective. *Nature Reviews Clinical Oncology*, 14(1):11–31, jan 2017.

[124] Ubaldo E. Martinez-Outschoorn, Maria Peiris-Pagés, Richard G. Pestell, Federica Sotgia, and Michael P. Lisanti. Cancer metabolism: a therapeutic perspective. *Nature Reviews Clinical Oncology*, May 2016.

[125] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[126] Lauren MF Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–935, 2006.

[127] Christian M. Metallo, Jason L. Walther, and Gregory Stephanopoulos. Evaluation of 13C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of Biotechnology*, 144(3):167–174, nov 2009.

[128] Evdokia Michalopoulou, Vinay Bulusu, and Jurre J Kamphorst. Metabolic scavenging by cancer cells: when the going gets tough, the tough keep eating. *British journal of cancer*, 115(6):635, 2016.

[129] Ljubisa Miskovic and Vassily Hatzimanikatis. Production of biofuels and biochemicals: in need of an oracle. *Trends in biotechnology*, 28(8):391–397, 2010.

[130] Ljubiša Miškoviü and Vassily Hatzimanikatis. Modeling of uncertainties in biochemical reactions. *Systems Biotechnology Biotechnology and Bioengineering*.

[131] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[132] Carmen G. Moles, Pedro Mendes, and Julio R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467—2474, 2003.

[133] M. K. Molloy. Performance analysis using stochastic petri nets. *IEEE Transactions on Computers*, 31(9):913–917, 1982.

[134] Yasuo Mori, Hiroki Nagse, Hiroshi Ando, Akira Horii, Shigetoshi Ichii, Shuichi Nakatsuru, Takahisa Aoki, Yoshio Miki, Takesada Mori, and Yusuke Nakamura. Somatic mutations of the apc gene in colorectal tumors: mutation cluster region in the apc gene. *Human molecular genetics*, 1(4):229–233, 1992.

[135] Bhanu Chandra Mulukutla, Andrew Yongky, Prodromos Daoutidis, and Wei-Shou Hu. Bistability in glycolysis pathway as a physiological switch in energy metabolism. *PLoS ONE*, 9(6):1–12, June 2014.

[136] Bhanu Chandra Mulukutla, Andrew Yongky, Prodromos Daoutidis, and Wei-Shou Hu. Bistability in glycolysis pathway as a physiological switch in energy metabolism. *PloS one*, 9(6):e98756, 2014.

[137] Erica C. Nakajima and Bennett Van Houten. Metabolic symbiosis in cancer: Refocusing the Warburg lens. *Molecular Carcinogenesis*, 52(5):329–337, 2013.

[138] Helder I Nakaya, Jens Wrammert, Eva K Lee, Luigi Racioppi, Stephanie Marie-Kunze, W Nicholas Haining, Anthony R Means, Sudhir P Kasturi, Nooruddin Khan, Gui-Mei Li, et al. Systems biology of seasonal influenza vaccination in humans. *Nature immunology*, 2011.

[139] Helder I Nakaya, Jens Wrammert, Eva K Lee, Luigi Racioppi, Stephanie Marie-Kunze, W Nicholas Haining, Anthony R Means, Sudhir P Kasturi, Nooruddin Khan, Gui-Mei Li, Megan McCausland, Vibhu Kanchan, Kenneth E Kokko, Shuzhao Li, Rivka Elbein, Aneesh K Mehta, Alan Aderem, Kanta Subbarao, Rafi Ahmed, and Bali Pulendran. Systems biology of vaccination for seasonal influenza in humans. *Nature immunology*, 12(8):786–95, aug 2011.

[140] Charlotte KY Ng, Susanna L Cooke, Kevin Howe, Scott Newman, Jian Xian, Jillian Temple, Elizabeth M Batty, Jessica Pole, Simon P Langdon, Paul AW Edwards, et al. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *The Journal of pathology*, 226(5):703–712, 2012.

[141] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[142] Gerlinde Obermoser, Scott Presnell, Kelly Domico, Hui Xu, Yuanyuan Wang, Esperanza Anguiano, LuAnn Thompson-Snipes, Rajaram Ranganathan, Brad Zeitner, Anna Bjork, David Anderson, Cate Speake, Emily Ruchaud, Jason Skinner, Laia Alsina, Mamta Sharma, Helene Dutartre, Alma Cepika, Elisabeth Israelsson, Phuong Nguyen, Quynh Anh Nguyen, A. Carson Harrod, Sandra M. Zurawski, Virginia Pascual, Hideki Ueno, Gerald T. Nepom, Charlie Quinn, Derek Blankenship, Karolina Palucka, Jacques Banchereau, and Damien Chaussabel. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, 38(4):831–844, 2013.

[143] Gerlinde Obermoser, Scott Presnell, Kelly Domico, Hui Xu, Yuanyuan Wang, Esperanza Anguiano, Luann Thompson-Snipes, Rajaram Ranganathan, Brad Zeitner, Anna Bjork, David Anderson, Cate Speake, Emily Ruchaud, Jason Skinner, Laia Alsina, Mamta Sharma, Helene Dutartre, Alma Cepika, Elisabeth Israelsson, Phuong Nguyen, Quynh-Anh Nguyen, a Carson Harrod, Sandra M Zurawski, Virginia Pascual, Hideki Ueno, Gerald T Nepom, Charlie Quinn, Derek Blankenship, Karolina Palucka, Jacques Banchereau, and Damien Chaussabel. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, 38(4):831–44, apr 2013.

[144] A. Ocone, A.J. Millar, and G. Sanguinetti. Hybrid regulatory models: A statistically tractable approach to model regulatory network dynamics. *Bioinformatics*, 29(7):910–916, 2013.

[145] Roberto Olivares-Hernández, Sergio Bordel, and Jens Nielsen. Codon usage variability determines the correlation between proteome and transcriptome fold changes. *BMC systems biology*, 5:33, feb 2011.

[146] Brett G Olivier and Jacky L Snoep. Web-based kinetic modelling using JWS Online. *BIOINFORMATICS APPLICATIONS NOTE*, 20(13):2143–214410, 2004.

[147] Sjoerd Opdam, Anne Richelle, Benjamin Kellman, Shanzhong Li, Daniel C. Zielinski, and Nathan E. Lewis. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Systems*, 4(3):318–329.e6, mar 2017.

[148] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28:245–248, March 2010.

[149] Alireza Osareh and Bita Shadgar. Machine learning techniques to diagnose breast cancer. In *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, pages 114–120. IEEE, 2010.

[150] Matthias OTTO, Reinhart HEINRICH, Gisela JACOBASCH, and Samuel RAPOPORT. A Mathematical Model for the Influence of Anionic Effectors on the Phosphofructokinase from Rat Erythrocytes. *European Journal of Biochemistry*, 74(2):413–420, 1977.

[151] N Ouchi, J L Parker, J J Lugus, and K Walsh. Adipokines in inflammation and metabolic disease. *Nat Rev Immunol*, 11(2):85–97, 2011.

[152] Diego Antonio Oyarz. A control-theoretic approach to dynamic optimization of metabolic networks. (February):130, 2010.

[153] Wei Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.

[154] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, 39(Database):D1002–D1004, jan 2011.

[155] Beresford N Parlett. *The symmetric eigenvalue problem*, volume 7. SIAM, 1980.

[156] Stephanos Pavlides, Diana Whitaker-Menezes, Remedios Castello-Cros, Neal Flomenberg, Agnieszka K. Witkiewicz, Philippe G. Frank, Mathew C. Casimiro, Chenguang Wang, Paolo Fortina, Sankar Addya, Richard G. Pestell, Ubaldo E. Martinez-Outschoorn, Federica Sotgia, and Michael P. Lisanti. The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell Cycle*, 8(23):3984–4001, dec 2009.

[157] Paul Pavlidis, Jie Qin, Victoria Arango, John J. Mann, and Etienne Sibille. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex, 2004.

[158] Jie Peng and Hans-Georg Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, pages 1056–1077, 2008.

[159] Meir Perez and Tshilidzi Marwala. Microarray data feature selection using hybrid genetic algorithm simulated annealing. In *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–5. IEEE, 2012.

[160] Jeffrey P. Perley, Judith Mikolajczak, Marietta L. Harrison, Gregery T. Buzzard, and Ann E. Rundell. Multiple Model-Informed Open-Loop Control of Uncertain Intracellular Signaling Dynamics. *PLoS Computational Biology*, 10(4), 2014.

[161] James L Peterson. Petri net theory and the modeling of systems. 1981.

[162] Carl Adam Petri. Technical report no. radc-tr-65-377. 1966.

[163] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *ArXiv e-prints*, February 2016.

[164] Roger Pique-Regi and Antonio Ortega. Block diagonal linear discriminant analysis with sequential embedded feature selection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[165] Mariagrazia Pizza, Vincenzo Scarlato, Vega Masignani, Marzia Monica Giuliani, Beatrice Aricò, Maurizio Comanducci, Gary T. Jennings, Lucia Baldi, Erika Bartolini, Barbara Capecchi, Cesira L. Galeotti, Enrico Luzzi, Roberto Manetti, Elisa Marchetti, Marirosa Mora, Sandra Nuti, Giulio Ratti, Laura Santini, Silvana Savino, Maria Scarselli, Elisa Storni, Peijun Zuo, Michael Broeker, Erika Hundt, Bernard Knapp, Eric Blair, Tanya Mason, Hervé Tettelin, Derek W. Hood, Alex C. Jeffries, Nigel J. Saunders, Dan M. Granoff, J. Craig Venter, E. Richard Moxon, Guido Grandi, and Rino Rappuoli. Identification of vaccine candidates against serogroup b meningococcus by whole-genome sequencing. *Science*, 287(5459):1816–1820, 2000.

[166] L. Popova-Zeugmann. *Time and Petri nets*. Springer, Heidelberg, GE, 2013.

[167] Bali Pulendran and Rafi Ahmed. Immunological mechanisms of vaccination. *Nature immunology*, 12(6):509–517, 2011.

[168] Bali Pulendran, Shuzhao Li, and Helder I. Nakaya. Systems vaccinology, 2010.

[169] Biao Qin, Yuni Xia, Sunil Prabhakar, and Yicheng Tu. A rule-based classification algorithm for uncertain data. In *2009 IEEE 25th International Conference on Data Engineering*, pages 1633–1640. IEEE, 2009.

[170] M. Quach, N. Brunel, and F. D'alché-Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.

[171] Troy D Querec, Rama S Akondy, Eva K Lee, Weiping Cao, Helder I Nakaya, Dirk Teuwen, Ali Pirani, Kim Gernert, Jiusheng Deng, Bruz Marzolf, Kathleen Kennedy, Haiyan Wu, Soumaya Bennouna, Herold Oluoch, Joseph Miller, Ricardo Z Vencio, Mark Mulligan, Alan Aderem, Rafi Ahmed, and Bali Pulendran. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature immunology*, 10(1):116–25, jan 2009.

[172] Chang F Quo, Richard a Moffitt, Alfred H Merrill, and May D Wang. Adaptive control model reveals systematic feedback and key molecules in metabolic pathway regulation. *Journal of computational biology*, 18(2):169–82, 2011.

[173] Chang F. Quo and May D. Wang. Biological interpretation of model-reference adaptive control in a mass action kinetics metabolic pathway model. *Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011*, pages 265–268, 2011.

[174] Mahadevan Radhakrishnan, Jeremy S. Edwards, and Francis J. Doyle. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical Journal 83.3 (2002)*, 83(3):1331–1340, 2002.

[175] FA Raja, N Chopra, and JA Ledermann. Optimal first-line treatment in ovarian cancer. *Annals of Oncology*, 23(suppl 10):x118–x127, 2012.

[176] Ramprasad Ramakrishna, Jeremy S Edwards, Andrew McCulloch, and Bernhard O Palsson. Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 280(3):R695–R704, 2001.

[177] Sridhar Ramaswamy, Ken N Ross, Eric S Lander, and Todd R Golub. A molecular signature of metastasis in primary solid tumors. *Nature genetics*, 33(1):49–54, 2003.

[178] JO Ramsay and BW Silverman. Functional data analysis.

[179] Rino Rappuoli and Alan Aderem. A 2020 vision for vaccines against HIV, tuberculosis and malaria. *Nature*, 473(7348):463–469, 2011.

[180] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[181] Florian Reisinger, Noemi Del-Toro, Tobias Ternent, Henning Hermjakob, and Juan Antonio Vizcaíno. Introducing the PRIDE Archive RESTful web services. *Nucleic acids research*, 43(W1):W599–604, jul 2015.

[182] Osbaldo Resendis-Antonio, Alberto Checa, and Sergio Encarnación. Modeling core metabolism in cancer cells: Surveying the topology underlying the warburg effect. *PLoS ONE*, 5(8), 2010.

[183] John A Rice and Bernard W Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 233–243, 1991.

[184] Lior Rokach and Oded Maimon. *Data mining with decision trees: theory and applications*. World scientific, 2014.

[185] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[186] Mahua Roy and Stacey D Finley. Computational Model Predicts the Effects of Targeting Cellular Metabolism in Pancreatic Cancer. *Frontiers in physiology*, 8:217, 2017.

[187] Sandip Roy, Terry F. McElwain, and Yan Wan. A network control theory approach to modeling and optimal control of zoonoses: Case study of brucellosis transmission in Sub-Saharan Africa. *PLoS Neglected Tropical Diseases*, 5(10), 2011.

[188] Charles M. Rudin, Erika Avila-Tang, Curtis C. Harris, James G. Herman, Fred R. Hirsch, William Pao, Ann G. Schwartz, Kirsi H. Vahakangas, and Jonathan M. Samet. Lung cancer in never smokers: Molecular profiles and therapeutic implications. *Clinical Cancer Research*, 15(18):5646–5661, 2009.

[189] Roberto Ruiz, José C Riquelme, and Jesús S Aguilar-Ruiz. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39(12):2383–2392, 2006.

[190] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

[191] Pedro A Saa and Lars K Nielsen. Construction of feasible and accurate kinetic models of metabolism: A bayesian approach. *Scientific reports*, 6:29635, 2016.

[192] Pedro A. Saa and Lars K. Nielsen. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnology Advances*, 35(8):981–1003, 2017.

[193] Pedro A Saa and Lars K Nielsen. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnology advances*, 2017.

[194] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, oct 2007.

[195] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[196] Y Sasaki, D Darmochwal-Kolarz, D Suzuki, M Sakai, M Ito, T Shima, A Shiozaki, J Rolinski, and S Saito. Proportion of peripheral blood and decidual CD4(+) CD25(bright) regulatory T cells in pre-eclampsia. *Clinical and experimental immunology*, 149(1):139–45, 2007.

[197] Jaya M. Satagopan and Katherine S. Panageas. A statistical perspective on gene expression data analysis. *Statistics in Medicine*, 22(3):481–499, 2 2003.

[198] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(Database):D38–D51, jan 2011.

[199] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, apr 2010.

[200] Santiago Schnell. Validity of the michaelis–menten equation–steady-state or reactant stationary assumption: that is the question. *The FEBS journal*, 281(2):464–472, 2014.

[201] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[202] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(Database issue):D431–3, jan 2004.

[203] Roland F Schwarz, Charlotte KY Ng, Susanna L Cooke, Scott Newman, Jillian Temple, Anna M Piskorz, Davina Gale, Karen Sayal, Muhammed Murtaza, Peter J Baldwin, et al. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med*, 12(2):e1001789, 2015.

[204] Clare Scott, Marc A Becker, Paul Haluska, and Goli Samimi. Patient-derived xenograft models to improve targeted therapy in epithelial ovarian cancer treatment. *Frontiers in oncology*, 3:295, 2013.

[205] Susan K Seaholm, Eugene Ackerman, and Shu-Chen Wu. Latin hypercube sampling and the sensitivity analysis of a monte carlo epidemic model. *International journal of bio-medical computing*, 23(1):97–112, 1988.

[206] Todd Senn, Stanley L Hazen, and WH Wilson Tang. Translating metabolomics to cardiovascular biomarkers. *Progress in cardiovascular diseases*, 55(1):70–76, 2012.

[207] Sohrab P Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, 2012.

[208] Lanlan Shen, Minoru Toyota, Yutaka Kondo, E Lin, Li Zhang, Yi Guo, Natalie Supunpong Hernandez, Xinli Chen, Saira Ahmed, Kazuo Konishi, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proceedings of the National Academy of Sciences*, 104(47):18654–18659, 2007.

[209] Lingyan Sheng, Roger Pique-Regi, Shahab Asgharzadeh, and Antonio Ortega. Microarray classification using block diagonal linear discriminant analysis with embedded feature selection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1757–1760. IEEE, 2009.

[210] Tomer Shlomi, Tomer Benyamini, Eyal Gottlieb, Roded Sharan, and Eytan Ruppin. Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect. *PLoS Computational Biology*, 7(3):e1002018, mar 2011.

[211] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, 31(1):1–31, 2017.

[212] M. A. Singer and S. Lindquist. Thermotolerance in saccharomyces cerevisiae: the yin and yang of trehalose. *Trends in Biotechnology*, 16(11):460–468, 1998.

[213] Despina Siolas and Gregory J Hannon. Patient-derived tumor xenografts: transforming clinical samples into mouse models. *Cancer research*, 73(17):5315–5319, 2013.

[214] M Slawski, M Daumer, and A-L Boulesteix. CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC bioinformatics*, 9:439, jan 2008.

[215] Kieran Smallbone and Pedro Mendes. Large-Scale Metabolic Models: From Reconstruction to Differential Equations. *Industrial Biotechnology*, 9(4):179–184, 2013.

[216] Kieran Smallbone, Hanan L. Messiha, Kathleen M. Carroll, Catherine L. Winder, Naglis Malys, Warwick B. Dunn, Ettore Murabito, Neil Swainston, Joseph O. Dada, Farid Khan, Pinar Pir, Evangelos Simeonidis, Irena Spasić, Jill Wishart, Dieter Weichart, Neil W. Hayes, Daniel Jameson, David S. Broomhead, Stephen G. Oliver, Simon J. Gaskell, John E G McCarthy, Norman W. Paton, Hans V. Westerhoff, Douglas B. Kell, and Pedro Mendes. A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS Letters*, 587:2832–2841, 2013.

[217] J Son, C A Lyssiotis, H Ying, X Wang, S Hua, M Ligorio, R M Perera, C R Ferrone, E Mullarky, N Shyh-Chang, Y Kang, J B Fleming, N Bardeesy, J M Asara, M C Haigis, R A DePinho, L C Cantley, and A C Kimmelman. Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. *Nature Letter*, 496:101–105, apr 2013.

[218] Hyun-seob Song, William R Cannon, Alexander S Beliaev, and Allan Konopka. Mathematical Modeling of Microbial Community Dynamics: A Methodological Review. pages 711–752, 2014.

[219] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[220] G A Soto-Pena, A L Luna, L Acosta-Saavedra, P Conde, L Lopez-Carrillo, M E Cebrian, M Bastida, E S Calderon-Aranda, and L Vega. Assessment of lymphocyte subpopulations and cytokine secretion in children exposed to arsenic. *FASEB J*, 20(6):779–781, 2006.

[221] Andrea Sottoriva, Haeyoun Kang, Zhicheng Ma, Trevor A Graham, Matthew P Salomon, Junsong Zhao, Paul Marjoram, Kimberly Siegmund, Michael F Press, Darryl Shibata, et al. A big bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209–216, 2015.

[222] Joan G Staniswalis and J Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

[223] Jörg Stelling. Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, 7(5):513–518, oct 2004.

[224] Ralf Steuer, Thilo Gross, Joachim Selbig, and Bernd Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103(32):11868–11873, 2006.

[225] Ralf Steuer, Thilo Gross, Joachim Selbig, and Bernd Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):11868–73, aug 2006.

[226] Sergey Stolyar, Steve Van Dien, Kristina Linnea Hillesland, Nicolas Pinel, Thomas J Lie, John A Leigh, and David A Stahl. Metabolic modeling of a mutualistic microbial community. *Molecular systems biology*, 3:92, 2007.

[227] Michael R Stratton. Exploring the genomes of cancer cells: progress and promise. *science*, 331(6024):1553–1558, 2011.

[228] Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. Gsea-p: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23):3251–3253, 2007.

[229] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[230] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

[231] Jyothi Subramanian and Richard Simon. Overfitting in prediction models - Is it a problem only in high dimensions? *Contemporary Clinical Trials*, 36(2):636–641, 2013.

[232] Patrick F. Suthers, Anthony P. Burgard, Madhukar S. Dasika, Farnaz Nowroozi, Stephen Van Dien, Jay D. Keasling, and Costas D. Maranas. Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metabolic Engineering*, 9(5-6):387–405, sep 2007.

[233] Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D Dobson, Judy Brewer, Michael Hanscho, Daniel C Zielinski, Kok Siong Ang, Natalie J Gardiner, Jahir M Gutierrez, et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):109, 2016.

[234] Gubian Sylvain, Xiang Yang, Suomela Brian, Hoeng Julia, and PMP SA. R Functions for Generalized Simulated Annealing. https://cran.r-project.org/web/packages/GenSA/GenSA.pdf, 2016.

[235] Jinying Tan and Xiufen Zou. Optimal Control Strategy for Abnormal Innate Immune Response. 2015, 2015.

[236] Yan Tan, Pablo Tamayo, Helder Nakaya, Bali Pulendran, Jill P. Mesirov, and W. Nicholas Haining. Gene signatures related to B-cell proliferation predict influenza vaccine-induced antibody response. *European Journal of Immunology*, 44(1):285–295, 2014.

[237] E Ke Tang, Ponnuthurai N Suganthan, and Xin Yao. Gene selection algorithms for microarray data based on least squares support vector machine. *BMC bioinformatics*, 7(1):1, 2006.

[238] John J Tentler, Aik Choon Tan, Colin D Weekes, Antonio Jimeno, Stephen Leong, Todd M Pitts, John J Arcaroli, Wells A Messersmith, and S Gail Eckhardt. Patient-derived tumour xenografts as models for oncology drug development. *Nature reviews Clinical oncology*, 9(6):338–350, 2012.

[239] F O R The, O F T H E Mechanics, and O F T H E Human. A mathematical model for the. 3(February):179–186, 1970.

[240] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93, 2010.

[241] Ines Thiele, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, Stefan G Thorleifsson, Rasmus Agren, Sergio Bordel, Arvind K Chavali, Paul Dobson, Warwick B Dunn, Lukas Endler, David Hala, Michael Hucka, Duncan Hull, Daniel Jameson, Neema Jamshidi, Jon J Jonsson, Nick Juty, Sarah Keating, Intawat Nookaew, Naglis Malys, Alexander Mazein, Jason A Papin, Nathan D Price, Evgeni Selkov, Martin I Sigurdsson, Evangelos Simeonidis, Nikolaus Sonnenschein, Kieran Smallbone, Anatoly Sorokin, Johannes H G M Van Beek, Dieter Weichart, Igor Goryanin, Jens Nielsen, Hans V Westerhoff, Douglas B Kell, and Pedro Mendes. A community-driven global reconstruction of human metabolism. 31(5), 2013.

[242] Philipp Thomas, Arthur V Straube, and Ramon Grima. Communication: Limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks. *THE JOURNAL OF CHEMICAL PHYSICS*, 135, 2011.

[243] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant

pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–9, 2005.

[244] Q. Tian, N. D. Price, and L. Hood. Systems cancer medicine: Towards realization of predictive, preventive, personalized and participatory (P4) medicine. In *Journal of Internal Medicine*, volume 271, pages 111–121, 2012.

[245] S. H. Toh, P. Prathipati, E. Motakis, C. K. Kwoh, S. P. Yenamandra, and V. A. Kuznetsov. A robust tool for discriminative analysis and feature selection in paired samples impacts the identification of the genes essential for reprogramming lung tissue to adenocarcinoma. *BMC Genomics*, 12, 2011.

[246] Laura Tolosi and Thomas Lengauer. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[247] Alicia A Tone, Melissa K McConechy, Winnie Yang, Jiarui Ding, Stephen Yip, Esther Kong, Kwong-Kwok Wong, David M Gershenson, Helen Mackay, Sohrab Shah, et al. Intratumoral heterogeneity in a minority of ovarian low-grade serous carcinomas. *BMC cancer*, 14(1):1, 2014.

[248] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[249] Niccoì O Totis, Laura Follia, Chiara Riganti, Francesco Novelli, Francesca Cordero, and Marco Beccuti. Overcoming the lack of kinetic information in biochemical reactions networks.

[250] Niccolò Totis, Marco Beccuti, Francesca Cordero, Laura Follia, Chiara Riganti, and Francesco Novelli. Dealing with indetermination in biochemical networks. *Giardini Naxos*, pages 123–4567, 2016.

[251] Linh M Tran, Matthew L Rizk, and James C Liao. Ensemble modeling of metabolic networks. *Biophysical journal*, 95(12):5606–5617, 2008.

[252] Linh M Tran, Matthew L Rizk, and James C Liao. Ensemble modeling of metabolic networks. *Biophysical journal*, 95(12):5606–5617, 2008.

[253] Eugenia Trushina and Michelle M Mielke. Recent advances in the application of metabolomics to alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(8):1232–1239, 2014.

[254] John S Tsang. Utilizing population variation, vaccination, and systems biology to study human immunology. *Trends in immunology*, 36(8):479–493, 2015.

[255] John S Tsang, Pamela L Schwartzberg, Yuri Kotliarov, Angelique Biancotto, Zhi Xie, Ronald N Germain, Ena Wang, Matthew J Olnes, Manikandan Narayanan, Hana Golding, et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell*, 157(2):499–513, 2014.

[256] John S. Tsang, Pamela L. Schwartzberg, Yuri Kotliarov, Angelique Biancotto, Zhi Xie, Ronald N. Germain, Ena Wang, Matthew J. Olnes, Manikandan Narayanan, Hana Golding, Susan Moir, Howard B. Dickler, Shira Perl, and Foo Cheung. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell*, 157(2):499–513, apr 2014.

[257] Kai-Yuen Tso, Sau Dan Lee, Kwok-Wai Lo, and Kevin Y Yip. Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? *BMC genomics*, 15(1):1, 2014.

[258] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, jan 2015.

[259] Maryanne T Vahey, Zhining Wang, Kent E Kester, James Cummings, D Gray Heppner, Martin E Nau, Opokua Ofori-Anyinam, Joe Cohen, Thierry Coche, W Ripley Ballou, and Christian F Ockenhouse. Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *The Journal of infectious diseases*, 201(4):580–9, feb 2010.

[260] Marc M Van Hulle. Self-organizing maps. In *Handbook of Natural Computing*, pages 585–622. Springer, 2012.

[261] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

[262] M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science*, 324(5930):1029–1033, may 2009.

[263] Matthew G Vander Heiden and Ralph J DeBerardinis. Understanding the Intersections between Metabolism and Cancer Biology. *Cell*, 168(4):657–669, 2017.

[264] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu. Gene expression data classification using support vector machine and mutual information-based gene selection. *Procedia Computer Science*, 47:13 – 21, 2015.

[265] A Varma and B O Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental microbiology*, 60(10):3724–31, oct 1994.

[266] Alexei Vazquez, Jiangxia Liu, Yi Zhou, and Zoltán N Oltvai. Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. *BMC systems biology*, 4:58, may 2010.

[267] Alexei Vazquez, Elke K Markert, and Zoltán N Oltvai. Serine biosynthesis with one carbon catabolism and the glycine cleavage system represents a novel pathway for atp generation. *PloS one*, 6(11):e25881, 2011.

[268] Alexei Vazquez, Elke K. Markert, and Zoltán N. Oltvai. Serine biosynthesis with one carbon catabolism and the glycine cleavage system represents a novel pathway for ATP generation. *PLoS ONE*, 6:3–7, 2011.

[269] Kathleen A Vermeersch and Mark P Styczynski. Applications of metabolomics in cancer research. *Journal of carcinogenesis*, 12, 2013.

[270] Veronica Vinciotti, Allan Tucker, Paul Kellam, and Xiaohui Liu. Robust Selection of Predictive Genes via a Simple Classifier. 5(1):1–11, jan 2006.

[271] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[272] Eberhard Voit, Ana R. Neves, and H. Santos. The intricate side of systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9452–9457, 2006.

[273] Eberhard O. Voit. *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*. Cambridge University Press, Cambridge, New York, 2000.

[274] Eberhard O. Voit. Biochemical Systems Theory: A Review. *ISRN Biomathematics*, 2013:1–53, jan 2013.

[275] Liqing Wang, Inanç Birol, and Vassily Hatzimanikatis. Metabolic control analysis under uncertainty: Framework development and case studies. *Biophysical Journal*, 87(6):3750–3763, 2004.

[276] Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E. Meijer-Van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, and John A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, 2005.

[277] Yu Wang, Igor V Tetko, Mark A Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W Mewes. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational biology and chemistry*, 29(1):37–46, 2005.

[278] Yuliang Wang, James A Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. *BMC systems biology*, 6(1):153, 2012.

[279] Yuliang Wang, James a Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC systems biology*, 6:153, 2012.

[280] Klaus-Wolfgang WENZEL, Boris I KURGANOV, Gerolf ZIMMERMANN, Victor A YAKOVLEV, Wolfgang SCHELLENBERGER, and Eberhard HOFMANN. Self-association of human erythrocyte phosphofructokinase. *The FEBS Journal*, 61(1):181–190, 1976.

[281] Hans V Westerhoff and Yi-Der Chen. How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. *Eur. J. Biochem*, 142:425–430, 1984.

[282] J Wilpon and L Rabiner. A modified k-means clustering algorithm for use in isolated work recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):587–594, 1985.

[283] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[284] Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, Olga Krebs, Meik Bittkowski, Elina Wetsch, Isabel Rojas, and Wolfgang Müller. SABIO-RK–database for biochemical reaction kinetics. *Nucleic acids research*, 40(Database issue):D790–6, jan 2012.

[285] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[286] Lucy R. Yates and Peter J. Campbell. Evolution of the cancer genome. *Nat Rev Genet*, 13:795–806, 2012.

[287] Lucy R Yates and Peter J Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806, 2012.

[288] Ka Yee Yeung-Rhee and Roger E Bumgarner. Multiclass classification of microarray data with repeated measurements. *Genome Biology*, 4(12), 2003.

[289] Keren Yizhak, Barbara Chaneton, Eyal Gottlieb, and Eytan Ruppin. Modeling cancer metabolism on a genome scale. *Molecular systems biology*, 11(6):817, jun 2015.

[290] Keren Yizhak, Edoardo Gaude, Sylvia Le Dévédec, Yedael Y Waldman, Gideon Y Stein, Bob van de Water, Christian Frezza, and Eytan Ruppin. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife*, 3:1–23, 2014.

[291] Keren Yizhak, Edoardo Gaude, Sylvia Le Dévédec, Yedael Y Waldman, Gideon Y Stein, Bob van de Water, Christian Frezza, and Eytan Ruppin. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife*, 3, nov 2014.

[292] Jamey D. Young. Learning from the steersman: A natural history of cybernetic models. *Industrial & Engineering Chemistry Research*, 54(42):10162–10169, 2015.

[293] Jamey D Young, Kristene L Henne, John A Morgan, Allan E Konopka, and Doraiswami Ramkrishna. Integrating cybernetic modeling with pathway analysis provides a dynamic, systems-level description of metabolic control. *Biotechnology and bioengineering*, 100(3):542–559, 2008.

[294] Jamey D Young and Doraiswami Ramkrishna. On the Matching and Proportional Laws of Cybernetic Models. pages 83–99, 2007.

[295] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *International Conference on Machine Learning (ICML)*, pages 1–8, 2003.

[296] Min Yu, Aditya Bardia, Ben S Wittner, Shannon L Stott, Malgorzata E Smas, David T Ting, Steven J Isakoff, Jordan C Ciciliano, Marissa N Wells, Ajay M Shah, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *science*, 339(6119):580–584, 2013.

[297] T. Yu, C. M. Lloyd, D. P. Nickerson, M. T. Cooling, A. K. Miller, A. Garny, J. R. Terkildsen, J. Lawson, R. D. Britten, P. J. Hunter, and P. M. F. Nielsen. The Physiome Model Repository 2. *Bioinformatics*, 27(5):743–744, mar 2011.

[298] Katsuyuki Yugi, Yoichi Nakayama, Ayako Kinoshita, and Masaru Tomita. Hybrid dynamic/static method for large-scale simulation of metabolism. *Theoretical Biology and Medical Modelling*, 2(1):42, 2005.

[299] Katsuyuki Yugi, Yoichi Nakayama, Ayako Kinoshita, and Masaru Tomita. Hybrid dynamic/static method for large-scale simulation of metabolism. *Theoretical biology & medical modelling*, 2:42, 2005.

[300] Jürgen Zanghellini, David E. Ruckerbauer, Michael Hanscho, and Christian Jungreuthmayer. Elementary flux modes in a nutshell: Properties, calculation and applications. *Biotechnology Journal*, 8(9):1009–1016, 2013.

[301] Zena M. Hira and Duncan F. Gillies. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015(1), 2015.

[302] Michalis Zervakis, Michalis E Blazadonakis, Georgia Tsiliki, Vasiliki Danilatou, Manolis Tsiknakis, and Dimitris Kafetzopoulos. Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC bioinformatics*, 10:53, jan 2009.

[303] Virginia Zewe and HJ Fromm. Kinetic studies of rabbit muscle lactate dehydrogenase. *Biochemistry*, 4(4):782–792, 1965.

[304] Boxuan Simen Zhao, Ian A. Roundtree, and Chuan He. Post-transcriptional gene regulation by mRNA modifications. *Nature Reviews Molecular Cell Biology*, 18(1):31–42, nov 2016.

[305] Daniel C Zielinski, Neema Jamshidi, Austin J Corbett, Aarash Bordbar, Alex Thomas, and Bernhard O Palsson. Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. *Scientific reports*, pages 1–31, 2017.

[306] Ali R Zomorrodi and Costas D Maranas. Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology*, 8(2):e1002363, 2012.

[307] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010.