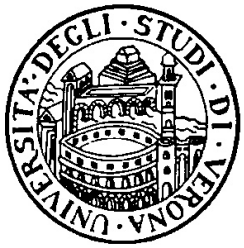

UNIVERSITY OF VERONA
DEPARTMENT OF BIOTECHNOLOGY

PHD SCHOOL
PHD PROGRAM IN BIOTECHNOLOGY
CYCLE XXXVII

WITH THE FINANCIAL CONTRIBUTION OF
UNIVERSITY OF VERONA

Unraveling strain-specific metabolic diversity using genome-scale models: development of computational methods and applications

S.S.D. BIOS-08/A

Coordinator: Prof. Flavia Guzzo

Tutor: Prof. Nicola Vitulo

Co-tutor: Prof. Giovanna Felis

Student: Gioele Lazzari

UNIVERSITY OF VERONA
DEPARTMENT OF BIOTECHNOLOGY

PHD SCHOOL
PHD PROGRAM IN BIOTECHNOLOGY
CYCLE XXXVII

WITH THE FINANCIAL CONTRIBUTION OF
UNIVERSITY OF VERONA

Unraveling strain-specific metabolic diversity using genome-scale models: development of computational methods and applications

S.S.D. BIOS-08/A

Coordinator: Prof. Flavia Guzzo

Tutor: Prof. Nicola Vitulo

Co-tutor: Prof. Giovanna Felis

Student: Gioele Lazzari

Unraveling strain-specific metabolic diversity using genome-scale models: development of computational methods and applications.
Gioele Lazzari. PhD Thesis. Verona, 6th December 2024.

“Nel mezzo del cammin di nostra vita
mi ritrovai per una selva oscura,
ché la diritta via era smarrita.”

“Midway upon the journey of our life
I found myself in a forest dark,
for the straightforward pathway had been lost.”

Dante Alighieri, Inferno, 1321
(translated by H. F. Cary, 1814)

Sommario 🇮🇹

Questa tesi di dottorato riguarda lo studio, lo sviluppo e l'applicazione di metodi computazionali per la ricostruzione e l'analisi di modelli metabolici a scala genomica. Nello specifico, questa tesi si focalizza sui metodi allo stato dell'arte per la modellizzazione simultanea di più ceppi batterici della stessa specie o di specie filogeneticamente vicine. Ciascun modello racchiude il potenziale metabolico di un ceppo, e può essere usato per simulare la crescita cellulare in diverse condizioni, permettendo di ottenere predizioni di fenotipo basate sul genoma.

Qui viene riportato lo sviluppo di un nuovo metodo di ricostruzione multi-ceppo, che prende spunto da quelli attualmente esistenti e cerca di superarne alcuni dei limiti. Questo metodo è stato implementato all'interno di Gempipe, uno strumento bioinformatico liberamente accessibile (github.com/lazzarigioele/gempipe) che si occupa anche delle fasi di correzione e analisi dei modelli, e che rappresenta il prodotto principale della tesi. Esso è stato confrontato con le principali alternative disponibili, dimostrandosi un passo avanti nell'ambito delle ricostruzioni multi-ceppo. Inoltre, è stato utilizzato in tre diversi casi studio riportati in questa tesi.

Nel primo, i modelli sono stati impiegati per predire un terreno di coltura adatto a sostenere la crescita di *Candidatus* Erwinia dacicola, un batterio endosimbionte noto per essere indispensabile nella riproduzione della mosca dell'olivo, insetto che danneggia i frutteti provocando perdite economiche nel settore agroalimentare. Il terreno di crescita predetto ha permesso di sostenere una coltura preliminare dell'endosimbionte, ma non di isolarlo, rivelando comunque fattori di crescita mai riportati prima.

Nel secondo caso studio, i modelli sono stati utilizzati per investigare la biodiversità metabolica intraspecifica di *Lactiplantibacillus plantarum*, un batterio lattico comunemente usato nelle produzioni alimentari, valutando una possibile correlazione tra potenziale metabolico e nicchia ecologica da cui i ceppi sono stati campionati. L'approccio implementato ha permesso un confronto simultaneo di oltre mille ceppi, che sono stati suddivisi in gruppi metabolicamente omogenei e apparentemente non correlati alla nicchia, in linea con una stile di vita definito nomade.

Nel terzo caso studio è stato modellato *Thauera* sp. *Sel9*, un ceppo batterico isolato da un impianto di trattamento di acque reflue e capace di produrre biopolimeri plastici a partire da acidi grassi a catena corta. L'informazione contenuta in genomi filogeneticamente vicini, unita ad uno screening fenotipico ad alto rendimento, ha permesso di correggere la topologia del modello, punto di partenza necessario per l'uso prospettico volto ad aumentare la resa in bioplastiche.

Parole chiave: modelli metabolici a scala genomica, ricostruzioni e analisi multi-ceppo, endosimbionti di insetti, requisiti nutrizionali, batteri lattici, screening fenotipici

Abstract

This PhD thesis is about the study, development, and application of computational methods for the reconstruction and analysis of genome-scale metabolic models. Specifically, it focuses on state-of-the-art methods for the simultaneous modeling of multiple bacterial strains of the same species or phylogenetically close species. Each model encapsulates the metabolic potential of a strain and can be used to simulate cellular growth under various conditions, enabling genome-based phenotype predictions.

The development of a novel multi-strain reconstruction method is here reported, inspired by existing approaches and designed to overcome some of their limitations. This method has been implemented into Gempipe, a freely accessible bioinformatics tool (github.com/lazzarigioele/gempipe) which also deals with model correction and analysis, and represents the main outcome of the thesis. It has been compared against the leading alternatives, proving to be a step forward in the field of multi-strain reconstructions. Furthermore, the tool has been applied in three case studies described in this thesis.

In the first one, models were used to predict a suitable growth medium for sustaining the growth of *Candidatus* Erwinia dacicola, an endosymbiotic bacterium known to be essential for the reproduction of the olive fly, insect that damages orchards and causes economic losses in the agrifood sector. The predicted growth medium enabled a preliminary culture of the endosymbiont, but not its isolation, nevertheless revealing previously unreported growth factors.

In the second case study, models were employed to investigate intraspecific metabolic biodiversity of *Lactiplantibacillus plantarum*, a lactic acid bacterium commonly used in food productions, assessing an eventual correlation between metabolic potential and the ecological niche from which the strains were sampled. The implemented approach enabled the simultaneous comparison of more than a thousand strains, which were divided into metabolically homogeneous groups seemingly unrelated to the niche, consistently with a lifestyle described as nomadic.

In the third case study *Thauera* sp. *Sel9* was modeled, a bacterial strain isolated from a wastewater treatment plant and capable of producing plastic biopolymers from short-chain fatty acids. The information contained in phylogenetically related genomes, combined with a high-throughput phenotypic screening, enabled the correction of the model's topology, a necessary starting point for prospective use aimed at increasing bioplastic yields.

Keywords: genome-scale metabolic models, multi-strain reconstructions and analysis, insect endosymbionts, nutritional requirements, lactic acid bacteria, phenotypic screenings

Contents

Aims and outline

Genome-scale metabolic models (GSMMs) have emerged as key tools in systems biology, providing a computational framework to study the metabolic capabilities of organisms. GSMMs can recapitulate physiological states of cells, and they also provide a simulation platform to test metabolic outcomes under different conditions [1]. As they are based on the annotation of a genome, GSMMs can be considered "children" of genomics. However, while traditional genomics essentially suggests gene functions and compares gene presence across organisms, GSMMs organize these genes into a data structure that links genotype to metabolic fluxes, enabling the next level of genomics where biological problems can be described in a more mechanistic way [2,3].

While working in a genomics and bioinformatics laboratory, I was tasked with developing expertise in the reconstruction and analysis of genome-scale metabolic models (GSMMs), an area that had not yet been explored within our lab. To achieve this, I relied primarily on self-directed study of peer-reviewed literature. This process allowed me to independently acquire knowledge in GSMMs, which I consider a significant personal and professional achievement that has contributed to expanding the research capabilities of the laboratory. Moreover, thanks to Prof. Bas Teusink, I had the opportunity to spend three months at the Systems Biology Lab of the Vrije Universiteit (NL), where I interacted with experts in the field. My understanding of genome-scale metabolic modeling is summarized in **Chapter 1**, which starts by describing the working principles and the inner components of GSMMs, and then it delves into their uses and limitations.

After learning the fundamentals of genome-scale metabolic modeling, the primary aim of my PhD program was to apply GSMMs to the exploration of metabolic diversities within a single bacterial species or closely related species. In this context, GSMM reconstruction methods capable of highlighting differences across strains are essential. These methods are summarized at the end of **Chapter 1**, with a focus on their strengths and limitations. The identified limitations were the drivers for the development of a new reconstruction method, which I included in a tool named Gempipe, whose implementation and validation are described in **Chapter 2**. Here, the main aims were to provide competitive features with respect to other state-of-art tools, and a performance comparison using publicly available datasets. Moreover, efforts were made to distribute Gempipe as a Conda package, ensuring it can be installed seamlessly. Not only multi-strain reconstructions are covered by Gempipe: some aspects of the manual curation and analysis are streamlined too. Given its versatility,

Gempipe has been employed throughout the rest of this thesis, which continues with three different case studies.

The first case study, reported in **Chapter 3**, focuses on the isolation of a yet-uncultured bacterium. *Bactrocera oleae* is a fly that lays its eggs inside green, unripe olive drupes. When growing, larvae produce holes that lead to secondary infections, causing economic losses in the olive markets [4]. Interestingly, larvae are known to develop on unripe olives only in presence of *Candidatus* Erwinia dacicola, the main endosymbiont of *B. oleae* [5]. The impairing of this symbiotic relationship could provide the basis for a pest control system [6], therefore the biology of *Ca.* E. dacicola should be characterized in more detail. In this perspective, the aim of the study was to obtain stable laboratory cultures of this endosymbiont. Here, a multi-strain GSMM reconstruction was employed to characterize *Ca.* E. dacicola and its closest free-living relative at the species level, comparing their growth on the same medium to deduce possible growth factors for the endosymbiont. This project is in collaboration with the microbiology laboratory led by Prof. Giovanna Felis, where Dr. Ilaria Checchia curated all the wet-lab experiments, including the search for antibiotic resistance genes, based on which selective growth conditions were established.

Dealing with the biodiversity of bacteria is both a challenge and an opportunity in microbial studies. While the definition of bacterial species describes its members as *"genomically and phenotypically coherent"* [7], it is well-known that strains of the same species may show differences on both the genomic and phenotypic level [8–10]. With advances in sequencing technologies, vast collections of bacterial genomes are now accessible, so these differences can be evaluated systematically. In **Chapter 4**, multi-strain reconstruction and analysis were performed to study the intraspecific diversity of *Lactplantibacillus plantarum*, a lactic acid bacterium with a broad application spectrum in health and food industries [11]. In this second case study, the primary objectives were to identify potential groupings of metabolically similar strains and to assess whether these data-driven clusters correlated with the environmental niches from which the strains were originally isolated.

Chapter 5 briefly reports the third case study, where a GSMM for *Thauera sel9* was drafted and curated in its reaction topology, using the information coming from both a multi-strain reconstruction of closely-related genomes and a high-throughput phenotypic screening (Biolog® Phenotype MicroArray plates). This strain, isolated from an activated sludge plant for wastewater treatment, is of biotechnological interest given its particular ability to produce polyhydroxyalkanoates from fatty acids [12]. Here, the main aim was to maximize the match between qualitative predictions and experimental data on

alternative substrate usage. This study is part of a larger project aiming to characterize the metabolism of *Thauera sel9* for a process optimization directed at improving the yield of biopolymers. The project is in collaboration with the microbiology lab led by Prof. Silvia Lampis, where Dr. Mehrdad Jaberi curated the Biolog® screenings.

Chapter 1 | Introduction to genome-scale metabolic models

In the last forty years, the sequencing of genomes has become exponentially cheap. Public databases like GenBank have been quickly populated, with the amount of stored nucleobases increasing 10-fold every 5 years. The early 2000's brought to light many technologies alternative to the Sanger sequencing, of which the Illumina short reads-based system prevailed. Its near monopoly was due not only to the reduced costs but also to the high sequencing accuracy. In the last few years, long-read sequencing gained popularity, with Nanopore and PacBio as leading technologies. These third generation approaches were born with issues in accuracy compared to Illumina, but today the gap is closing and they are becoming routinely used also for non-model organisms [13].

Thanks to the improvements in sequencing technologies, the research community is today flooded by genome sequences. In the past, the explosion of molecular biology, driven by a reductionist approach, was essential for determining the function of single genes, for example through knock-outs and complementation of *Escherichia* lab strains. This paved the way for the interpretation of new genomes, querying them for sequences similar to those previously characterized. In this sense, today it is possible to obtain an approximate "list of components" of an organism, just by reading its genome. However, to really grasp the functioning of the organism, it is needed to understand how these components work together, a task that is not trivial. Indeed, despite the huge amount of genome sequences publicly available today, the extraction of meaningful, mechanistic information out of them remains a bottleneck [14,15].

Components of an organism are so numerous that omics sciences, including genomics, intrinsically involve a "big data" analysis problem, which cannot be handled without computational techniques. To extrapolate mechanistic information out of these biological big data, the *structuring* and *contextualization* of single components are keys. These are the pillars of systems biology, a computational discipline that aims at the understanding of an organism as a system, making order in the chaos of data. As progressively complex biological questions arise, systems biology approaches are preferred over reductionist approaches, because they provide a means to look at the bigger picture, shifting the focus from single components to complex networks, ultimately leading to clues for the answer [16,17] .

In this context, metabolic models are system biology approaches to study the metabolism of an organism. They can be roughly divided into two categories: kinetic models and genome-scale models [18]. The latter are the main objective of this thesis.

Genome-scale metabolic models (GSMMs) were born in 1999, following the sequencing of the genome of *Haemophilus influenzae* strain Rd, the first complete genome of a free-living organism [19]. The assembled genome was subjected to genome-scale metabolic modeling, *structuring* and *contextualizing* the annotated genes into a coherent network of metabolic reactions [1]. This remarkable effort was made possible thanks to extensive manual curation, which today remains both a blessing and a curse for high-quality models [20,21]. Since then, thanks to the advances and decreasing cost of genome sequencing, the genome-scale modeling approach became established. In 2019, it was estimated that 6000 GSMMs had been produced, encompassing reconstructions for bacteria (the vast majority), archaea and eukaryotes. However, since manual curation is extremely time-consuming, most of these reconstructions were drafts, meaning that the organism has likely been represented inaccurately [20].

To understand the essence of GSMMs, a brief comparison with kinetic models is needed. Kinetic models use differential equations to describe the variation of metabolite concentrations over time. To obtain the concentration of a metabolite in a specific instant, the knowledge of enzymatic rates is required. As a metabolite can be produced and consumed by different enzymes at the same time, the contribution of different production and consumption rates must be accounted for. The rate of an enzyme is dependent not only on the concentrations of reactant and product metabolites, but also on the kinetic parameters of the enzyme, such as the Michaelis-Menten constants. While the description of metabolite concentrations over time is certainly valuable, in practice kinetic models present difficulties that prevent their usage [18].

The main difficulty is to obtain kinetic parameters for each modeled enzyme. Even in dedicated databases, the kinetic parameters of a certain enzyme could have been measured for a different (model) organism, or under in-vitro conditions that can be far from the in-vivo, physiological ones [18,22,23]. As a consequence, kinetic models usually describe small reaction networks, like isolated pathways, so they hardly describe the interdependency with the rest of the system [24].

In contrast, GSMMs do not require the knowledge of the kinetic parameters of enzymes. This is possible because GSMMs are based on the steady-state assumption, which defines the sum of production and consumption rates

constantly equal to 0 for each metabolite in the network. This can be mathematically described with the following equation [18]

$$\frac{dX}{dt} = \sum_{i=1}^{n} v_i = 0$$

where dX / dt is the variation of a metabolite concentration over time, and v_i are production and consumption rates of n enzymes that are producing or consuming the metabolite. This assumption can be justifiable as biochemical reactions express quickly (in the order of milliseconds) compared to other biochemical processes like the transcriptional regulation or the production of biomass; therefore, pathways observed for a sufficient amount of time tend to reach a (pseudo) steady-state [18,22]. Moreover, the steady-state can be seen as the average condition of a growing population of cells living different lifecycle stages [25]. In practice, the only growth phase that resembles a steady-state is the mid-exponential phase, where experimental data are collected to be integrated into GSMMs [14].

In steady-state, it is appropriate to refer to the enzyme rates simply as "fluxes". In this particular condition, the flux of a particular reaction is constrained by the fluxes of all the other reactions in the GSMM, by a set of linear equations. The constraints imposed by the steady-state, together with other constraints that will be later discussed, explain why the reconstruction and analysis of GSMMs is also known as "constraints-based modeling", and the associated methods are known as "COBRA methods" (COnstraints-Based Reconstruction and Analysis methods) [18].

Not being dependent on kinetic parameters means that GSMMs can be scaled up to reach a true "genome-scale", modeling all the metabolic enzymes in a genome. However, even if dynamic behaviors can be approximated, GSMMs essentially lose the time scale describing a single, optimal metabolic state reached in a constant environment. This intrinsic simplicity implies that GSMMs are lighter to run then kinetic models, and indeed they can be executed on personal laptops with no problems [14].

1.1 Internal components

Each GSMM, in its essence, is composed of at least five components: (1) reaction network; (2) gene associations; (3) biomass composition; (4) uptake / secretion rates; (5) energy parameters. They are collectively encoded in SBML files [26], a

commonly used file format to store and share GSMMs, and they will now be discussed in dedicated subsections.

1.1.1 Reaction network

The reaction network is made up of a stoichiometric description of all the reactions encoded by a genome. While such reactions should be in principle balanced in mass and charge, a GSMM is essentially described by the reaction stoichiometry alone. Each reaction is defined by the stoichiometric relation among its reactant and product metabolites, meaning that molar amounts are made explicit for each involved chemical species. In addition, to describe and guarantee mass and charge balance of reactions, each metabolite is associated with a chemical formula and a charge. Finally, each reaction and each metabolite has an unique ID, and the same metabolite can be associated to more then one ID only to differentiate its presence among different subcellular compartments, including the external compartment (outside the cell). This is particularly relevant for eukaryotic GSMM, as they typically possess many subcellular compartments [27].

Two are the principal ID systems (namespaces) used in genome-scale metabolic modeling: the BiGG-based [28] and the SEED-based [29]. Beside the slightly different set of reactions and metabolites they describe, the main difference is that BiGG IDs provide an easier reading, while SEED IDs are just numerical codes. For example, the ID for the metabolite glucose-1-phosphate is “g1p” in BiGG, while “cpd00089” in SEED (where “cpd” generically means “compound”); the ID for the reaction phosphofructokinase is “PFK” in BiGG, while “rxn00545” in SEED (“rxn” stands for “reaction”). To distinguish the same metabolite in different subcellular compartment, the symbol for the compartment is appended after an underscore in the ID, so that glucose-1-phosphate in cytosol is “g1p_c” or “cpd00089_c”, while in chloroplast it is “g1p_h” or “cpd00089_h”. The main standard symbols for subcellular compartments are “c” for cytosol, “p” for periplasm, “m” for mitochondria, “x” for peroxisome / glyoxysome, “r” for endoplasmic reticulum, “v” for vacuole, “n” for nucleus, “g” for golgi apparatus, “u” for thylakoid, “l” for lysosome, and “h” for chloroplast. The extracellular space is represented too, using “e” [28].

The conversion between BiGG and SEED IDs, as well as other biochemistry databases not specialized in metabolic modeling (such as KEGG [30] and MetaCyc [31]) is provided by the MetaNetX [32] conversion dictionaries; however, the translation between different namespaces is not a trivial nor a completely solved problem. This limits the reusability and integration of existing GSMMs, contributing to the need for manual curation [33,34].

The BiGG and SEED ID systems are named after two databases with similar names: the BiGG database [28] and the ModelSEED database [29]. The latter is defined as a biochemistry database that was built automatically integrating reactions and metabolites from many different, general purpose metabolic databases; of these, KEGG [30] and MetaCyc [31] are reported to be the only databases integrated in ModelSEED in their entirety. Thanks to the automated integration of different databases, the chemical space covered by ModelSEED is vast. However, it is known to contain unbalanced and redundant reactions [29].

While the BiGG [28] database covers a narrower chemical space compared to ModelSEED [29], the easier reading of its IDs makes the BiGG namespace a convenient choice. Indeed, the interpretation of metabolic maps drawn with Escher [35,36] (a drawing tool developed by the same research group that created BiGG) requires little or no effort if the loaded GSMM adheres to the BiGG namespace. However, BiGG [28] is technically not a biochemistry database, but a collection of manually-curated GSMMs, from which a collection of reactions and metabolites is derived. This has several consequences [33].

First of all, the same reaction in different models could have different reversibility. For example, the transaminase "ILETA2" appears both as reversible and irreversible, as models iCN718 [37] and iCN900 [38] define it differently. Second, the same metabolite can appear in different models with different IDs, leading to duplicate metabolites. For example, the cytosolic indole is encoded as "indole_c" in model iJO1366 [39], and "ind_c" in model iSynCJ816[40]. Since both models are included in BiGG [28], both indole IDs are available, causing an inconsistency. Some of these duplication cases can be detected and corrected thanks to the conversions provided by MetaNetX [32], some others cannot (like in the example of indole). Third, reactions can also have duplicates if they use duplicate metabolites. For example, the tryptophan synthase using indole is represented by both by "TRPS2" ("indole_c + ser__L_c $\rightarrow$ h2o_c + trp__L_c") and "TRPS2_1" ("ser__L_c + ind_c $\rightleftharpoons$ h2o_c + trp__L_c"), which also present inconsistency in their reversibility. Fourth, even if a metabolite appears in different models with the same ID, it could be defined with a different chemical formula or charge depending on the original model, leading to another layer of inconsistency. For example, tagatose-6-phosphate ("tag6p__D") was modeled with charge -1 or -2, while flavin-mononucleotide ("fmn") was modeled with formulas $C_{17}H_{18}N_4O_9P$ or $C_{17}H_{19}N_4O_9P$. This leaves also space for the coexistence of different modeling paradigms, like in the case of acyl-carrier-protein (ACP)-ligated fatty acids that requires a description of ACP, many times abstracted using "R" or "X" artificial atoms representing part of proteins. For example, cis-hexadec-9-enoyl-ACP ("hdeACP") appears with

one of the following formulas: $C_{27}H_{49}N_2O_8PRS$, $C_{16}H_{29}OSR$, $C_{16}H_{29}OX$, or $C_{400}H_{631}O_{143}N_{96}P_1S_3$. Of course, differences in mass and charge can lead to duplicate reactions with different stoichiometry, like for example in the case of the pantetheine-phosphate adenylyltransferase, present in BiGG [28] both as "APPAT" ("atp_c + 2.0 h_c + pan4p_c ⇌ dpcoa_c + ppi_c"), and as "PTPATi" ("atp_c + h_c + pan4p_c ⇌ dpcoa_c + ppi_c").

While the approach implement in BiGG leaves flexibility to define, for the same metabolite, different protonation states eventually occurring in different subcellular compartments or physiologically different organisms, this creates also inconsistency issues, again limiting the integration of existing GSMMs and the development of automated methods.

Inside the network of reactions encoded in a GSMM, reactions must never violate the stoichiometric consistency. This condition is not dependent on the reaction balances in mass and charge, so it is not dependent on the chemical formula and charge associated with metabolites. Instead, it depends solely on the stoichiometry of reactions, which is consistent only when it is possible to assign a positive mass to all the metabolites in a GSMM. In other words, stoichiometric inconsistency implies that at least one metabolite in the network is forced to have zero mass [41,42]. For example, the following network composed by two reaction is stoichiometrically inconsistent, as C is forced to have zero mass:

$R_1 = A \longrightarrow B$
$R_2 = A \longrightarrow B + C$

In every GSMM, each reaction is also associated with a pair of constraints, called lower bound (*lb*) and upper bound (*ub*). They represent the minimum and maximum flux allowed for that reaction, respectively. These constraints, also known as "capacity constraints" provide a rough link with kinetics: they are used to represent experimentally measured uptake or secretion rates and, more in general, to provide indications on the reversibility of reactions. Indeed, in absence of further experimental data, reactions known to be irreversible are described by (lb=0, ub=+∞), where the symbol ∞ represents a non-limited flux value. Instead, reversible reactions are associated with (lb=-∞, ub=+∞) [18]. To define the reversibility of a reaction, its thermodynamics must be studied [8]. In absence of thermodynamics data, rules of thumb can be applied: irreversible reactions are those transferring a phosphate group from ATP molecules, as well as those involving quinones; in the other cases, reactions should be left reversible [21]. By convention, an "infinite" flux value is represented computationally with an arbitrary, high constant like 1000 or 999999.

1.1.2 Gene associations

In GSMMs, each reaction is associated with the set of genes in the genomes that are needed for its expression. This association is expressed using boolean logic, so enzyme complexes and alternative isoforms can be described. Specifically, isoforms are linked by an "OR" statement, while different subunits are linked by an "AND". Using parentheses, it is possible to represent even complicated enzymatic complexes, like for example "gene_a OR (gene_b AND (gene_c OR gene_d))". This kind of expression is historically known as GPR, acronym for Gene-to-Protein-to-Reaction association [3,43]. When a reaction occurs spontaneously, the GPR often contains a dedicated label such as "spontaneous", to distinguish it from the so-called "orphan" reactions, where the underlying genes are unknown or not found in the genome [21,44,45]. When a gene is included in a GSMM, it should appear in at least one GPR [34].

The link between reactions and their underlying genes demonstrates how genome-scale modeling can be seen as a branch of genomics. Indeed, the reconstruction of GSMMs is strongly dependent on the annotation of a genome. However, the process of translating a set of unknown genes into a stoichiometrically coherent and balanced set of reactions is notoriously time-consuming. The workload for the reconstruction of a GSMM varies from many months to several years, depending on the complexity of the organism, its available body of knowledge, and the number of researchers involved. Many bioinformatic tools are available to partially automated specific steps of the reconstruction; however, as stated in the beginning, manual curation is still required to produce high-quality models [21].

GSMMs should be seen not only as tools to provide clues to biological questions. They should be seen also as organism-specific repositories of knowledge, like an advanced metabolic database specifically tailored for an organism. Like every database, GSMMs can and should be updated each time new knowledge is available for the organism. For this and other reasons, the reconstruction of a GSMM is defined as an "iterative" process. Modelers should update the models each time new experimental data is available. For example, the GSMM for *Escherihia coli* relies on more than twenty years of refinements [21].

As it depends on genome annotation, the reconstruction of a GSMM relies on homology-based gene function prediction, which transfers a function from a characterized gene of a model organism to an unknown gene of the organism under study [2]. Genomes are shaped by many different events such as duplication, chromosome rearrangements, gene loss, horizontal gene transfer,

and so on; consequently, functional assignment based on homology can be complicated [46].

Homology exists in two instances: orthology and paralogy. The first implies that the unknown gene and the reference gene originate from the same gene of a common ancestor. In the second, one of the two genes derives from a duplication event in the ancestor [2]. As orthologs tend to conserve their original function more than paralogs, the determination of orthologs is a critical step in the reconstruction of genome-based reaction networks [46,47].

Given the gene sequences of a reference organism and of an organism to model, various bioinformatic methods are available for the determination of orthologs. A simple, commonly accepted standard method is the BRH (best-reciprocal hits) alignment, where gene sequences of an organism are aligned over the sequences of the other, and vice versa: if two genes result associated as best hit homologs in both the alignments, then they are assumed as orthologs. These all-vs-all comparisons are usually computationally intensive [47–49], also because they are often performed with the accurate BLAST aligner [50].

To solve this issue, various methods have been developed to speed up the orthologs determination, trying to balance speed with accuracy. Usually, such methods group genes into "orthogroups" containing genes from two (or more) organisms originating from the same ancestor gene; this means that orthogroups describe a many-to-many association. However, some tools also take into account synteny, which is the order in which gene appear in genomes, to distinguish between orthologs and paralogs in the same orthogroup; this enable orthogroups to be split in smaller groups, describing associations closer to one-to-one [47,49,51–54].

The reconstruction of reaction networks is difficult not only because of the ortholog determination: once an enzyme has been identified, it could be involved in several different reactions. To classify molecular functions of genes, the EC codes are used. When a reference is not available, the function of a gene is hypothesized by automated tools aligning its sequence against a database of annotated sequences, typically inheriting the EC code of the best hit [2]. Subunits of the same enzymatic complex are usually assigned to the same EC code, helping in the definition of GPRs (for example, all the subunits of ATP synthase are given the EC number 3.6.3.14). However, the same enzymatic complex can express different molecular functions, so it can be associated with different EC codes; moreover, the same EC code can represent several reactions involving slightly different reactants or products. Finally, EC codes do not cover all reactions happening in a cell [55]. These facts make the EC codes

sub-optimal for reconstructing the stoichiometrically balanced network of reactions needed by GSMMs; indeed EC codes were not made for network reconstructions [3].

Several databases exist providing the conversion between EC codes and metabolic reactions, such as KEGG [30] and MetaCyc [31]. While the first remained freely accessible throughout the years, the second became available only through a paid subscription, limiting its usage. However, they were not created specifically for metabolic modeling: they contain unbalanced or generic reactions, as well as metabolites not defined in their physiological state. To give some examples for the KEGG database, the metabolite ATP (code C00002) is defined with chemical formula $C_{10}H_{16}N_5O_{13}P_3$: however, inside the cell it is usually deprotonated as $C_{10}H_{12}N_5O_{13}P_3^{4-}$; the reaction long-chain fatty acid:acyl carrier protein (code R01406) is defined with a generic "long chain fatty acid" (code C00638) among its reactants. Therefore, when such generic databases are used to convert automatically annotated EC codes into metabolic reactions, manual curation will be needed to ensure the network of reactions is balanced and stoichiometrically coherent [21].

Another approach to convert unknown gene sequences into a set of reactions suitable for metabolic modeling is to use dedicated databases such as BiGG [28], containing not only metabolites and reactions, but also their associated genes. Indeed, each reaction in BiGG is linked to the models using it, giving access to the model-specific GPR and gene sequences for that reaction. Therefore, one can align unknown gene sequences directly on the genes in the BiGG [28] database, making use of GPRs and reactions already defined. However, as said in subsection 1.1.1, BiGG is not a coherent biochemistry database, but rather a collection of manually curated GSMMs. This implies the presence of duplicated metabolites and reactions, and multiple options of charge and chemical formula for the same metabolite. Therefore, even this approach requires manual curation to ensure a balanced and stoichiometrically coherent final network. Additionally, the current version of BiGG (v1.6) still contains few GSMMs, and they are strongly biased on model species: 108 GSMMs are present, of which 88 modeling bacteria, of which just 22 modeling bacteria not part of the *Escherichia* / *Shigella* species complex [28,56]. Therefore, the set of genes contained in BiGG lack representativeness, so they are hardly usable with organisms phylogenetically distant from model species.

Independently of the strategy used to go from a set of gene sequences to a reaction network, another key issue is that automated annotations are usually incomplete, sometimes reaching a poor 52% of completeness [57]. Moreover, EC codes retrieved during genome annotation may be incomplete (for example

"3.6.1.-"). These issues mean that, when translating genes into reactions, many reactions could be missing from the GSMM reconstruction, representing metabolic gaps in the network [57] (see subsection 1.2.2).

1.1.3 Biomass composition

A special, unbalanced reaction known as the biomass equation is always included in the reaction network of a GSMM. Through millimolar coefficients, this reaction precisely defines the quantity of each biomass precursor contained in 1 gram of cellular dry weight. However, many GSMMs are known not to respect this normalization at 1 gram [58].

In general, the biomass equation can be formulated using different levels of detail, according to the experimental data available, and the level of accuracy desired for the GSMM. At the most basic level, the biomass equation captures the macromolecular content of a cell. This level includes, for example, the four nucleotides and ribo-nucleotides, forming DNA and RNA; the twenty protein-forming amino acids; different fatty acids and phospholipids, forming cell membranes. At a more detailed level, the biomass equation describes species- or strain-specific components always found in vital cells, less easy to measure. These precursors include, for example, teichoic and lipoteichoic acids, as well as peptidoglycan units, forming cell walls and other envelope structures; different essential cofactors and vitamins; glycogen, polyphosphates, polyhydroxyalkanoates, or other storage compounds forming granules; inorganic ions found in ashes [59–62].

The biomass composition is determined for cells growing in the mid-exponential phase, in accordance with the steady-state assumption. Experimental methods for the determination of the macromolecular composition have been reviewed and benchmarked, specifically for the purpose of GSMMs reconstruction [14,62]. It must be kept in mind that the biomass composition is species-dependent, or even strain-dependent, and it is usually a function of the growth conditions applied to the organism [61]. Therefore, prior to the experimental determination of biomass composition, the growth conditions must be decided with care: they should be representative of a "mean" cellular lifestyle, or perhaps representative of a particular industrial setting. However, a GSMM has no limit in the number of biomass equations it can harbor; therefore, multiple biomass equations can be included in the same model, and the optimal one can be selected for each particular application [63].

When available experimental data are not sufficient, and the missing information is not available in literature, the biomass equation (or part of it)

can be inherited from phylogenetically close organisms already modeled in a quality GSMM [57,61,64]. When working with bacteria, if there are no phylogenetically close GSMMs with a biomass equation to copy, it is possible to use a generic (mean) biomass composition either for gram positives, gram negatives or cyanobacteria [44,65,66].

1.1.4 Uptake / secretion rates

The identification of transporters has always been an issue in bioinformatics, because they share a high degree of sequence similarity, with just small differences determining their substrate specificity [55]. However, for every compound known to be imported in the cell or secreted to the medium, a dedicated transporter should be included in the GSMM. The same is true for each molecule known to pass from a subcellular compartment to another, which is particularly relevant for eukaryotic models.

The gold standard database for transporters is the TCDB [67], containing the reference genes and the information on how they combine to form the transport complex. Moreover, each transporter is classified using a TC code, which takes into account both the transport mechanism and the phylogenesis of the organism. In a GSMM, reactions describing transporters are respectful of the transport mechanism. For example, ATP-binding cassette importers always have water and ATP as reactants, and phosphate, ADP and H^+ as products, in addition to the imported substrate. Diffusions are modeled too: small, hydrophilic molecules should be left free to diffuse through the membranes, and the GPR associated to the corresponding transport reactions is marked as "spontaneous" (provided that diffusion is not facilitated) [21].

The reaction network encoded in a GSMM always contains a particular type of reactions known as exchange reactions. They are artificial, unbalanced reactions that allow a metabolite to enter or leave the system, and they must not be confused with the transport reactions, which are biological reactions. Like all reactions in a GSMM, exchange reactions have a lower bound (lb) and an upper bound (ub): when (lb=0, ub=+∞), the associated metabolite is free to be eventually secreted by the cell (provided the presence of a dedicated transporter); when (lb=-∞, ub=0), the metabolite is available to the cell with no limits [21]. As previously mentioned, an "infinite" (unconstrained) reaction bound is represented computationally with an arbitrary, high constant (usually 1000 or 999999).

In each GSMM, the growth conditions are described by means of the exchange reactions. Growth conditions are intended in terms of a chemically defined

nutritive environment, while other abiotic factors such as the temperature cannot be directly included into a variable. When experimental uptake and secretion rates are determined, they can be used as additional constraints for the GSMM. In this case, the exchange reaction for a particular compound is constrained to (lb=-(v + e), ub=-(v - e)) if the compound is uptaken, or to (lb=v - e, ub=v + e) if it is secreted, where v and e are the measured rate and its associated experimental error, respectively, both having unit of measure $mmol \cdot g\text{DW}^{-1} \cdot h^{-1}$ [27].

However, there is no standard practice for the definition of the growth conditions. Exchange reactions can also be used to represent, instead of uptake / secretion rates, medium concentrations in $mmol/L$ [68] (see subsection 1.2.1). As recipes for growth media often include complex ingredients like for example tryptone, yeast extract or meat extract, their dissociation into the solvent must be taken into account. For example, given a recipe having 10 g/L of peptone and 5 g/L of meat extract, millimolar concentrations must be determined for each forming amino acid, vitamin, nucleotide, etc [68].

Apart from exchange reactions, there might be other artificial, unbalanced reactions in a GSMM. Demand reactions are irreversible reactions that allow the accumulation of a metabolite, independently from the biomass equation. Sink reactions are reversible reactions that not only allow accumulation, but also depletion of metabolites. Demand reactions are usually temporarily introduced to verify the availability of all the precursors needed to synthetize a particular metabolite [21]. Sink reactions are sometimes used to describe the depletion of cellular carbon stocks like starch or glycogen [69]. However, demand and sink reactions should be introduced only if strictly necessary [21].

Given their artificial nature, the biomass equation and the exchange, demand and sink reactions are associated with an empty GPR, but they do not count as orphan reactions [70].

1.1.5 Energy parameters

Each GSMM contains two energy parameters: the growth-associated maintenance (GAM) energy, and the non-growth-associated maintenance energy (NGAM), both expressed in the form of ATP hydrolysis [21,71].

GAM is part of the biomass equation, and takes into account the energy required for cell replication, of which some contributions may be partially unknown [8,21,71]. Therefore, GAM accounts for energy spent during biosynthesis and assembly of biomass precursors, including the polymerization

energy. If experimental determination in chemostats is not possible, GAM can be underestimated approximating the polymerization energy, for example accounting 2 ATPs for each nucleic acid monomer, or 4 ATPs per protein-forming amino acid. Additional costs for proof-reading and double helix unwinding can also be taken into consideration [8,21,59,62].

The NGAM is a separated reaction, and accounts for the energy spent to ensure cellular homeostasis (turgor pressure, protein turnover, cell repair, motility, etc). In the case of NGAM, the constraint for the GSMM does not manifest as stoichiometric coefficients, but instead as a fixed lower bound [21,71].

Experimental determination of GAM and NGAM requires a gap-filled GSMM (see subsection 1.2.2) and a chemostat culture with the same growth settings used to determine the biomass composition [8,61]. The "full coupling" assumption should be verified, meaning that energy (ATP) produced from the catabolism of a limiting carbon source is fully consumed for growth and maintenance, without producing cellular carbon stocks [8,72,73]. In these conditions, different growth rates are kept constant, and the corresponding carbon source uptake rate and eventual secretion rates are measured. Then, setting all these rates as constraints for the GSMM, the theoretical maximum ATP production rate is computed using the flux balance analysis (see subsection 1.2.1). A scatter plot is created, with fixed growth rates on the X-axis, and the computed ATP production on the Y-axis. Points are finally subjected to linear regression: GAM and NGAM correspond to the slope and intercept, respectively. More points can be obtained repeating the chemostat measurements for several different carbon sources [8,61,62].

Like the biomass composition, GAM and NGAM can be functions of the growth conditions. Therefore, multiple GAM and NGAM could be incorporated in a GSMM and then selected according to its usage [71,74].

1.2 Uses and limitations

Not only are GSMMs an organism-specific repository of knowledge. Their underlying data structure enables GSMMs to be used in growth simulations, leading to predictions. The predictive power exerted by GSMMs can range from qualitative to quantitative, depending on the type and amount of experimental data used to constrain the model, and depending on the extensiveness of the manual curation applied.

1.2.1 Flux balance analysis

Among the many available techniques to analyze a GSMM, a simple, robust, and widely diffused technique for simulating growth is the flux balance analysis (FBA). It treats the model as a matrix ***S*** associated with two vectors ***LB*** and ***UB***. ***S*** has metabolites in rows and reactions in columns. Different rows are used for the same metabolite in different compartments, including the external environment (see subsection 1.1.1). Moreover, all reactions in the GSMM appear as columns, including the biomass equation and the other artificial reactions (exchange, demand and sink reactions). Cells of ***S*** contain the stoichiometric coefficients appearing in the reaction, with negative sign if the metabolite is consumed rather than produced. For this reason, ***S*** is commonly known as the "stoichiometric matrix". For those metabolites not involved in a particular reaction, the corresponding cells will be filled with zeroes; indeed, ***S*** is described as a sparse matrix, as it mainly contains zeroes. Vectors ***LB*** and ***UB*** describe the capacity constraints, so they contain the lower and upper bound, respectively, of all the reactions in ***S***. These vectors are fundamental because they contain a description of both the thermodynamics of reactions and the nutritional environment in which the cell is placed (see subsection 1.1.4) [27,43,75].

FBA typically requires a reaction to be indicated as "objective", chosen from the reactions included in the GSMM. Given the stoichiometric matrix ***S*** and the lower and upper bound vectors ***LB*** and ***UB*** (i.e., the constraints), FBA computes a flux distribution that maximizes (or minimizes) the flux through the objective reaction. In other words, the objective reaction is the driving force which distributes fluxes over the entire reaction network, in a way that is "optimal" for the objective reaction to carry the highest flux possible, in compliance with the stoichiometry ***S*** and the constraints ***LB*** and ***UB***. Therefore, the FBA solution not only includes the flux through the objective reaction or "objective value" (v_{obj}), but also a vector $\boldsymbol{v}$ containing a flux for each reaction in the GSMM [8,18,27,59].

Mathematically, performing an FBA means to solve the following linear optimization problem (system of linear equations) [27,76]:

$\max(v_{obj})$
subjected to:

$$\mathbf{S} \cdot \boldsymbol{v} = 0$$

$$\boldsymbol{LB} \leq \boldsymbol{v} \leq \boldsymbol{UB}$$

The flux distribution depends on the chosen objective reaction [25]. A widely used objective reaction is the biomass equation; other objectives include for example the ATP formation, or the production of a metabolite of interest [25,27,59].

When the objective reaction is the biomass equation, and exchange reactions represent uptake / secretion rates ($mmol \cdot g\mathrm{DW}^{-1} \cdot h^{-1}$), then the objective value represents the maximal theoretical growth rate ($1 / h$). Instead, if exchange reactions represent medium concentrations ($mmol / L$, see subsection 1.1.4), then the objective value represents the maximal theoretical growth yield ($g\mathrm{DW} / L$) [27,68,77].

The objective value will be zero if at least one of the biomass precursors cannot be synthesized because of gaps in the network [64] (see subsection 1.2.2). For the linear optimization problem, zero is however a valid solution; if the imposed constraints cannot be satisfied, the solver will inform that the solution is "infeasible".

Many COBRA methods evolved from the plain FBA. They all depend on the presence of a numerical optimization solver, for example the open-source GLPK or the commercial CPLEX [78–80]. Solvers are interfaced with libraries, where the actual COBRA methods are implemented. Two of the most adopted libraries are COBRA Toolbox [78] and COBRApy [81]. The first is implemented using MATLAB, a commercial language, and it is probably the most complete library due to historical reasons: it was one of the earliest implementations, first published in 2007 [75,78,82]. The second is a lighter Python implementation, which integrates smoothly with the free and open-source Python ecosystem [79,81].

As realistic GSMMs have more reactions than metabolites, the system has more unknown variables than linear equations. Therefore, there is usually not a unique solution but rather a "solution space", composed by many different flux distributions (many alternative $\boldsymbol{v}$) leading to the same objective value, if additional constraints are not applied [27,55]. The flux variability analysis (FVA) and the parsimonious FBA (pFBA) can be used to cope with this underdetermination issue [8,18]. For each reaction, FVA determines the flux range (from minimum to maximum) which leads to the same objective value determined by FBA (or a fraction of it), thus giving an indication on how flexible the metabolism is in reaching its optimum. Essential reactions can then be identified as those where zero is not included in the allowed range [8,18,43]. Instead, pFBA selects, among the alternative solutions, the one that minimizes the number of active reactions, under the assumption that faster growth is

favored. Indeed, the smaller the number of enzyme complexes to be synthesized, the faster should be the cellular growth [83].

To better assess the solution space, flux sampling techniques can be used. Sampling algorithms generate per-reaction probability distributions indicating the likelihood of alternative flux values, while the solution feasibility is granted considering all reactions at the same time. This is more informative than FVA, which only returns per-reaction maximum and minimum fluxes while evaluating each reaction independently (meaning that, for example, maximums obtained for two reactions may be infeasible to sustain contemporarily) [84–86]. Importantly, sampling algorithms work even in absence of a defined objective, making them useful also when the actual objective is not clear or not compatible with assumptions (see below) [84,87,88]. Moreover, different sets of experimental constraints can be used to generate and compare different per-reaction probability distributions, highlighting how the metabolism adapts to, for example, different growth conditions, environmental stresses, or genetic perturbations [84,86].

A limitation of FBA, and GSMMs in general, is that it is not possible to directly simulate transcriptional regulation (native or externally induced), nor protein activation (like the phosphorylation of protein kinases in signaling pathways). Examples of phenotypes not directly simulable include: the growth repression due to an high concentration of a toxic molecule; the growth inhibition due to high temperatures; the catabolite repression phenomena, preventing a substrate to be consumed until another one is fully depleted [9,27,32,89]. However, a physiological flux distribution in all these scenarios can in principle be obtained, provided that a sufficient number of experimental rates has been measured and imposed as constraints [76]; in these applications, GSMMs are useful for their descriptive potential, rather than their predictive potential.

While no stage of gene regulation can be directly simulated, its consequences can be taken into account as constraints. Indeed, omics-datasets can be integrated into GSMMs, shrinking the solution space toward a context-specific conformation, so the number of possible flux distributions is reduced [90]. The procedure of model restriction can be roughly divided into switch-based and valve-based: the first applies thresholds to omics data determining which reactions have to be kept or removed; the second scales reactions' bounds in accordance with the omic signal, without removing reactions [90,91]. In the context of transcriptomics, switch-based methods might be preferable, as they are more robust to noise and less affected by the erroneous assumption of correlation between transcripts and flux levels [92]. Among switch-based methods, a further distinction can be made in the requirement of objective

functions or other metabolic functions that have to be guaranteed, at the cost of reintroducing reactions previously excluded according to thresholds [92,93]. These are often methods of choice when modeling microorganisms, as an objective is more frequently identified with respect to cells belonging to multicellular organisms [93].

Speaking of cellular objectives within the FBA framework, it is important to realize that the maximization of the growth rate always implies a maximization of the growth yield per *mmol* of substrate, i.e. a maximization of the substrate usage efficiency [77]. However, this is just one of the many fitness strategies that an organism could have. Other strategies include, for example, the fast depletion of substrate or the release of toxic metabolites. In other words, simply setting the biomass equation as the objective reaction in FBA-based simulations may not faithfully model the physiology of many organisms. Simply put, the objective reaction (or more generally, the objective function) is never known a priori for a wild strain, and it could have evolved really far from the substrate usage efficiency [8].

Regarding growth strategies, microbial metabolism could be roughly divided into slow and efficient, or fast and inefficient. In the first case, the classic maximization of the flux through the biomass equation could be the appropriate choice for the FBA objective, since the maximum ATP amount is extracted from nutrients to fuel the assembly of biomass precursors. In the second case, energy is "wasted" through the secretion of energetic molecules, such as, for example, lactic acid in homolactic fermentation. This metabolic behavior is defined as an "overflow metabolism" and it is quite diffused among bacteria. To complicate things, the same organism can switch between different growth strategies depending on the environmental conditions, usually the substrate concentration. Indeed, when substrate is abundant, less efficient metabolism is favored, switching to a high efficiency metabolism at lower growth rates, when substrate becomes limiting [94–96].

In the case of overflow metabolism, as in that of transcriptional regulation discussed above, obtaining a realistic flux distribution with FBA is, in principle, always possible, provided that a sufficient amount of experimentally determined constraints is applied to the GSMM. However, doing so, the model progressively loses its predictive potential. In other words, the more constraints are added, the more a GSMM shifts from being predictive to being descriptive [8,14,24,77,94]. Researchers who rely on GSMMs just for their predictive power will agree with the words of Schuster and colleagues [94], stating that this practice "*diminishes the elegance of FBA and may lead to a circular reasoning, where the result to be computed has already been used as input information*" [94].

1.2.2 Gapped networks

Despite the efforts put into the generation of a balanced and stoichiometrically consistent network of reactions, the network itself could present holes or gaps due to missing reactions. They could originate from (i) incomplete genome annotation; (ii) inaccurate translation of genes into reactions; (iii) presence of yet-unknown pathways; (iv) promiscuous enzymes; (v) underground metabolic pathways, given by low-rate, weak side activities of enzymes. These gaps could break connectivity of pathways, preventing interdependent reactions to carry flux. For example, when pathways leading to synthesis of biomass precursors are blocked due to the presence of gaps, then the *in silico* growth fails. Gaps may generate the so-called "dead-end" metabolites, which can be a consequence of two different scenarios: (a) a reaction has a reactant that never appears as a product in other reactions; (b) a reaction has a product that never appears as reactant in other reactions. Dead-end metabolites originate from incomplete reconstructions, and sometimes they represent true "knowledge gaps": an enzyme that produces or consumes a known metabolite may have been at most assumed, while its coding sequence never been identified. From an FBA point of view, reactions associated with dead-end metabolites will be blocked (unable to carry flux) in any growth condition, due to the missing connections. The procedure of closing gaps, either manually using literature and dedicated wet-lab experiments, or automatically using specific algorithms, is known as gap-filling [8,21,27,55,97,98].

Manual gap-filling can be complex and time-consuming, taking months of effort [97]. Generic metabolic maps provided by KEGG [30] or MetaCyc [31] are of great help to identify the metabolic context of a dead end metabolite, providing a starting point for searches in literature or in the genome [21]. Moreover, methods based on linear optimization are available to suggest possible reactions to be included [99–102].

A widely diffuse method is SMILEY [99], implemented in COBRApy [81] within its *gapfill* function. With this method, two GSMMs are required: the first is the model to be gap-filled; the second defines the repository of possible reactions to be used in order to guarantee a specified minimum objective value during FBA. Of course, the search of candidate reactions is limited to the content of the second GSMM. With this method, different alternative solutions can be proposed, which are usually prioritized to contain the smallest possible set of reactions. Moreover, reaction sets can be prioritized by assigning a weight to each reaction in the repository GSMM [81,99].

When a set of reactions is selected to restore the specified phenotype, the genome should always be searched for the underlying genes, for example through BLAST alignments, which could ultimately lead to an update of the genome annotation. This is needed to create GPR associations, since the number of orphan reactions in a quality GSMM must be minimized [64,97,103,104]. Automated gap-filling methods like SMILEY [99] only propose known reactions to be inserted; however, some gaps could be resolved by changing the reversibility of a reaction, correcting the biomass composition, or defining whole new reactions (including transport and exchange reactions) [55,98].

Gap-filling has been defined as a "tricky business" [21]. Given the existence of multiple gap-filling solutions, and the possible introduction of gap-filled reactions as orphans (i.e., with no GPR), gap-filling usually introduces uncertainty in a GSMM [103,104]. Gaps that are mandatory to close are the ones strictly required for biomass formation on known media, otherwise the GSMM will not be able to perform growth simulations. Additional gap-fillings should be performed only if required to model specific, known metabolic phenotypes. When this parsimony criteria is not applied and gaps are closed only to maximize connectivity of the reaction network or to solve all dead end metabolites, artifact (false positive) reactions might be included [21,55,97,105]. However, many gap-filling approaches do not follow this criteria [44,97,106].

1.2.3 Infeasible thermodynamics

Solutions given by FBA do not consider thermodynamics. Therefore, thermodynamic feasibility of a metabolic network must be verified beforehand. For example, in a mammalian GSMM, aerobiotic catabolism of glucose should yield more or less 34 ATPs per glucose. However, in earlier human GSMMs, glucose could yield way more ATP, an error that was corrected in updated reconstructions. Such errors are due to stoichiometrically balanced cycles (SBCs) or energy-generating cycles (EGCs). They are caused by an insufficient number of constraints in the reaction network, typically regarding reversibility of reactions. As a consequence, a loop of reactions can be formed, which violates the second law of thermodynamics generating energy out of nothing in form of energy-containing molecules like for example ATP (and other nucleoside triphosphates), NADH, NADPH, FADH2, menaquinol, etc [21,43,107].

Such reconstruction artifacts are widely diffused in GSMMs that did not receive extensive manual curation. For example, it was calculated their presence in over 85% of models from ModelSEED [29], while they are rare among the curated

models contained in BiGG [56]. Consequences can be severe, as these artifact cycles typically inflate the predicted growth rate by 25% [43,107].

1.2.4 Quantitative predictions

The most classical quantitative feature among those predictable with a GSMM is probably the growth rate. As discussed above (subsection 1.2.1) FBA models the production of biomass in such a way that the substrate utilization efficiency is maximized; this is the case where GSMMs are suited for the prediction of the growth rate [77].

In batch, the growth rate to measure is the maximal and constant growth rate μ_{max} reached during the bacterial mid-exponential phase, in accordance with the fundamental assumption of steady-state. During this phase, cells double at constant rate:

$$N_1 = N_0 \cdot 2^n$$

where N_1 is the number of cells after n generations, and N_0 is the starting number of cells. Since n is equivalent to the elapsed time $t = t_1 - t_0$ divided by the generation time t_d, the above equation can be rewritten in the following logarithmic form:

$$ln\left(\frac{N_1}{N_0}\right) = \frac{t_1 - t_0}{t_d}\, ln2$$

As t_d and $ln2$ are both constant, μ_{max} is also constant, provided that t_0 and t_1 are both taken during the mid-exponential growth phase when substrates necessary for growth are not limiting. Moreover, since the number of cells N is directly related to their dry weight X, that is easier to measure or estimate by means of the optical density, the final expression can be written as [108]:

$$\mu_{max} = \frac{0.693}{t_d} = \frac{lnX_1 - lnX_0}{t_1 - t_0}$$

The above expression is valid for batch growth, and it is a function of both the growth conditions, for example the temperature, and the strain used. Therefore, μ_{max} should remain constant for a strain if growth conditions do not change, assuming that no impacting mutations occur during strain propagation in laboratories [109].

In chemostats (continuous cultures) where the steady-state is established, the growth rate equals the dilution rate D, i.e. the number of complete volume changes per hour. The chemostat is the preferred equipment to collect experimental data to constrain or validate GSMMs, as any number of steady-states can be obtained at different dilution rates, so the physiology of a strain can be studied at different growth rates on the same substrate [108,110].

For a quantitative growth prediction, GSMMs must be constrained with appropriate uptake / secretion rates: in this context, the approach of setting the exchange reactions as medium concentrations is not sufficient to obtain a reliable output and must be avoided [68]. Substrate-specific experimental uptake / secretion rates ($mmol \cdot g\mathrm{DW}^{-1} \cdot h^{-1}$) can be measured for a specific growth rate using a chemostat:

$$q_s = \frac{S_F - S}{X} D$$

where S_F and S are the substrate concentration ($mmol / L$) in the feeding medium and in the bioreactor, respectively, D is the dilution rate ($1 / h$), and X is the biomass concentration ($g\mathrm{DW} / L$) in the bioreactor [76,109–111].

Apart from setting realistic uptake / secretion rates as constraints for the GSMM, a mass- and charge-balanced network of reactions is a prerequisite for quantitative predictions, as unbalanced networks could lead to the generation of energy out of nothing. For example, extra ATP molecules could be produced due to an incorrect proton balancing of reactions [2,21,58]. Moreover, the stoichiometric consistency must be assured (subsection 1.1.1), the network must be appropriately gap-filled for biomass production (subsection 1.2.2), and thermodynamically infeasible cycles must be removed (subsection 1.2.3). Another prerequisite for quantitative predictions, often neglected, is the biomass equation to be set up to exactly 1 gram of dry cellular weight [58] (subsection 1.1.3). Respecting these requisites, it is possible to start comparing FBA results with the experimental data.

To obtain an experimental validation, the predicted value should fall between the error range of the experimental value. When the growth conditions to be quantitatively validated are too numerous, a researcher can validate just a subset, as long as it spans the entire growth rate range from minimum to maximum [96,112].

If the growth rate simulation is not in accordance with the real data, there could be several possible explanations, for example: (i) presence of missing or limiting

nutrients; (ii) incorrect estimation of the energy parameters (GAM and NGAM); (iii) imprecise setting of uptake or secretion constraints; (iv) presence of thermodynamically infeasible cycles; (v) regulatory mechanisms that cannot be modeled; (vi) biomass equation not normalized to 1 gram; (vii) different fitness strategy (not following the maximization of substrate usage efficiency) [21].

In case of missing or limiting nutrients, a reduced costs analysis (also known as sensitivity analysis) can be performed. Each reaction is associated with a reduced cost (r_i) as an integrating part of every FBA solution. The reduced cost is an indication of how much an infinitesimal flux alteration in a particular reaction (dv_i) would alter the computed objective value: $r_i = dv_{obj} / dv_i$. In other words, if the flux v_i of reaction i is increased by Δv_i, then the objective value will be updated as follows: $v_{obj}' = v_{obj} + r_i \cdot \Delta v_i$. Therefore, by inspecting the reduced cost of the exchange reactions, it is possible to understand which substrate is missing or limiting. When missing substrates are not of interest, to detect the most limiting substrates among those already being uptaken, scaled reduced costs are computed: $r_i^s = r_i \cdot v_i / v_{obj}$ [8,21,68,113].

1.2.5 Qualitative predictions

The output of a GSMM is always numeric. Anyway, there are cases where numeric accuracy is not needed or simply it cannot be reached because the model has not properly been parameterized with experimental data. For example, uptake rates could not have been measured, so an accurate growth prediction cannot be obtained. However, an investigator could be interested to understand "whether" something can potentially happen, rather than "how much". In these cases, exchange reactions may also be set as concentrations (see subsection 1.1.4), and the numeric output is interpreted as a binary response (yes or no), which is still useful when investigating a metabolism [68,77]. For example, when the predicted growth is greater than a specified threshold, it is possible to affirm that the organism *can potentially* grow on the provided set of nutrients.

One of the typical qualitative predictions that can be performed with GSMMs is that of substrate utilization. Given a C, N, P, or S source, it is possible to predict whether it can potentially be catabolized by the organism. Frequently, a GSMM is set to simulate growth on a minimal medium composed of glucose, ammonia, phosphate, and sulfate as the sole C, N, P and S sources, respectively. Then, for each alternative C substrate, the lower bound of the exchange reaction for glucose is set to 0, while the one for the alternative substrate is set to an arbitrary value (e.g., a plausible concentration). If the objective value is greater than a threshold, say 0.001, then it is possible to conclude that the alternative

substrate can potentially be catabolized by the organism. A similar procedure is repeated for N, P, and S substrates [114,115]. However, a strain could present particular nutritional requirements, like auxotrophies for aminoacids or vitamins. In this case, FBA is applied on a medium supplemented with the essential growth factors, and the objective value v_{obj} is saved; then the substrate to test is added to the medium, and an updated objective v_{obj}' is computed: if v_{obj}' > v_{obj} + 0.001, then the substrate can potentially be catabolized [9]. Moreover, this screening is often repeated both in aerobic and anaerobic conditions [114,116].

To validate substrate utilization predictions, a dedicated growth assay must be performed for each nutritive source [9]. Alternatively, a high-throughput phenotypic screening can be carried out, like the widely diffused Biolog® assays. This omics technology, also referred to as "phenomics", emerged around twenty years ago, proposing a 96-well plate-based screening. It was commercially named Phenotype MicroArray (PM), and it is today available as a set of 25 different PM plates [117], of which PM1, PM2, PM3 and PM4 are frequently used for GSMM validation. Specifically, PM1 and PM2 are used to test C sources, PM3 is for N sources, while the PM4 plate tests P and S sources.

In principle, each of the 96 wells of a Biolog® PM plate contains a different substrate. To verify its consumption, the color shift of tetrazolium dyes is evaluated. Indeed, the uptake and catabolism of a substrate is associated with the generation of redox potential such as the production of NADH, which transfers electrons to a tetrazolium dye causing its color shift, that can be measured at regular intervals of time by spectrophotometry. Interestingly, this signal is not dependent on cellular replication, nor on cell morphology, contrary to the optical density. When plotted over time, data points assume a sigmoid distribution resembling a growth curve, but without death phase, as tetrazolium dyes do not turn back in their original state [117].

Usually, when reading Biolog® plates, two wavelengths are recorded at the same time: 750 *nm*, to monitor turbidity, and 590 *nm*, to monitor color shifts of tetrazolium dyes. Since the tetrazolium read may be affected by cell turbidity, absorbance at 750 *nm* is subtracted from that at 590 *nm* for each well. Then, the signal of the blank (well A0) is subtracted from that of each substrate. Finally, the signal at time $t = 0$ is subtracted from those at $t > 0$, for each substrate, and eventual negative signals are set to zero [118–120]. However, not all authors implement this data normalization [121]. There is no standard method to convert the color shift kinetics into a binary response (can grow / cannot grow), a process known as "growth calling". Indeed, an arbitrary threshold is applied, either on the height reached by the sigmoid [122,123] or on the area under it

[124,125]. However, abiotic dye reductions have been described, and they depend on the chosen dye [126–128].

After the growth calling, to qualitatively validate a GSMM, simulated and experimental substrate utilization data are compared. Each substrate can be classified as true positive (TP) when both data agree on growth, true negative (TN) when both data agree on growth absence, false positive (FP) when growth is predicted but not experimentally observed, or false negative (FN) when no growth is predicted while it is experimentally observed. To evaluate how well a GSMM recapitulates substrates utilization, some metrics are usually computed [44,129]:

$$\text{precision} = \frac{TP}{TP+FP}$$
$$\text{recall} = \frac{TP}{TP+FN}$$
$$\text{specificity} = \frac{TN}{TN+FP}$$
$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Importantly, before being used to validate a GSMM, substrate utilization assays are used to curate its network topology, guiding gap-fillings and eventual removals of erroneous reactions included during the reconstruction process. In case of FN, missing reactions must be added, together with their underlying genes. The case of FP is trickier, as there could be different explanations: (i) some genes have been associated with the wrong reactions, therefore some reaction will end up to be removed; (ii) a transcriptional repression have occurred in vitro, but this cannot be predicted by GSMMs; (iii) the inoculation time was not sufficient for the organism to reprogram its metabolism and express the enzymes needed to consume the substrate [64,130]. Another important point to keep in mind is that Biolog® assays from different laboratories could differ when the same plate and the same strain are used: inconsistencies may arise from the use of different thresholds for growth calling, different inoculation volumes, or different incubation times [114]. Moreover, results may be confounded by abiotic dye reductions [126–128].

Another kind of prediction achievable with GSMMs is that of auxotrophies for amino acids, vitamins, and other growth factors in general. For example, to predict if the organism is auxotroph for a particular amino acid, it is sufficient to simulate growth in a minimal medium containing all the other 19 proteinogenic amino acids. Provided a curated network topology, if no growth is observed *in silico*, then the organism is a potential auxotroph for that amino acid [9,114,131].

The experimental validation is usually performed with single-omission growth experiments [132].

GSMMs can also be used to predict essential genes and synthetically lethal gene pairs. In this case, GPRs are fundamental as they determine if a reaction stays or will be removed when a particular gene is deleted from the model. An accurate biomass equation, at least in terms of involved precursors rather than their quantity, is fundamental in gene-deletion simulations. This is why vitamins and cofactors are usually included in the biomass equation, even if they are present in traces. Moreover, essential genes are also a function of the growth medium composition [8,27]. Simulation results are usually compared with the experimental data obtained using techniques like TraDIS (transposon-directed insertion site sequencing) [133,134], using the same metrics presented above (precision, recall, specificity, accuracy).

1.2.6 Multi-strain qualitative predictions

Qualitative methods have been applied also to multi-strain reconstructions. In this particular application context, hundreds of GSMMs are created, one for each strain of a species or group of phylogenetically close species. This of particular interest in biodiversity studies (see subsection 1.3.2), where contents and predictions of hundreds of strain-specific GSMMs enable (i) to screen strains for desired metabolic traits; (ii) to classify strains according to their metabolic capabilities; (iii) to define species in terms of their core metabolic potential [114].

The first, pioneering study of this kind was published in 2013 by Monk and colleagues [114], where 55 strains of *Escherichia coli* and *Shigella* were modeled. Here, for the first time, the concepts of core- and pan-reactome were introduced, in analogy with those of core- and pan-genome used in pangenomics: the first describes the set of metabolic reactions that is always present in all the strains of a particular clade (for example, a species); the second describes the set reactions encoded in at least one of the genomes of the clade. In addition, the concept of accessory reactome was introduced, defined by the reactions in the pan-reactome that are not part of the core-reactome.

In [114], it was found that the accessory reactome of the *Escherichia* / *Shigella* group contained mostly reactions for the catabolism of alternative C, N, P and S sources and, most importantly, that it could be used to cluster strains according to their environment niche. The clustering was performed starting from the binary matrix of C, N, P and S substrates utilization, applying the Ward's method on the Jaccard dissimilarity matrix, this way creating what could be

defined a “phylometabolic” tree. From this tree, for example, it was possible to see that *Shigella* strains have lost the ability to catabolize D-allantoin, D-malate, and xanthine, which is instead retained by *Escherichia* strains. Moreover, *Escherichia* strains could be further divided into intestinal pathogen, commensal and extra-intestinal pathogen strains, so the phylometabolic tree reflected the strain-specific environmental niches. Monk and colleagues [114] also used the deck of GSMMs to predict auxotrophies, detecting strain-specific auxotrophies for several amino acids and vitamins.

Other similar GSMMs-based biodiversity studies followed [37,115,116,134–136]. In [116], for example, 64 strains of *Staphylococcus aureus* were modeled and analyzed, again creating a phylometabolic tree based on binary features such as the presence of auxotrophies and the growth on alternative substrates. Such trees, however, are suited to handle any layer of binary features, including those derived from classic genomics, which can be supplemented to those derived from metabolic modeling. Indeed, an additional binary layer describing the presence / absence of various virulence factors was included. Together, these binary features lead to the creation of a classification tree, which allowed a precise strain-specific host prediction: human-associated or livestock-associated.

Another example is [115], where 410 strains of *Salmonella* spanning 64 different serovars were reconstructed and analyzed. In this case, the accessory reactome contained reactions involved not only in the catabolism of substrates, but also in the cell wall production, in line with the diversity of wall antigens on which the traditional *Salmonella* strains classification is based. Again, a phylometabolic tree was built, enabling the prediction of serovars and host preference.

Similar qualitative analysis of GSMMs was performed in [137], where 24 different species of *Penicillium* (a representative strain for each species) were reconstructed and analyzed. The phylometabolic tree, this time, was based on the presence / absence of modeled reactions, and was then compared with a traditional phylogenomic tree based on shared single-copy orthologs [138]. The comparison showed that the two trees were not superimposable, highlighting the inadequacy of traditional phylogenomic trees to represent metabolic differences.

One of the limitations set by these multi-strain studies is the difficulty to validate each GSMM created, since reconstructions are hundreds and it is clearly hard to individually validate them all. Therefore, once a phylometabolic tree has been divided into clusters, one or two representative strains for each

cluster can be selected for validation, for example using the Biolog® assays discussed above. As clusters are in principle metabolically coherent, this strategy should improve the validation coverage [114,115].

1.3 Automated reconstructions

Unfortunately, as described in previous sections, the production of high quality models is today still dependent on manual curation and, for certain applications, on parametrization with experimental data. However, many aspects of the reconstruction of GSMMs can today be automated. Several bioinformatic tools exist for this purpose, which implements methods that can be roughly divided into bottom-up or top-down. The latter became widely adopted, thanks to their ability to produce stoichiometrically consistent, extensively gap-filled, annotated, simulation-ready GSMMs; this is possible because reconstructions start from a metabolically comprehensive, pre-curated model used as template, from which reactions are subtracted rather than added (also known as a "universal model" or "universe") [34,139].

Quality of reconstructed GSMMs can be evaluated using Memote [70], a community standard tool made for this purpose. It provides a total score per model, taking into account 5 different metrics: stoichiometric consistency, metabolite annotations, reaction annotations, gene annotations, and SBO (System Biology Ontology) annotations. Stoichiometric consistency is weighted the most in the total score, while minor weights are given to the annotations [70]. The latter have to follow the MIRIAM namespace (Minimum Information Required In the Annotation of Models), which is a collection of perennial identifiers provided by the European Bioinformatics Institute [140]. Unfortunately, Memote reports include information on, for example, dead end metabolites, unrealistic growth rates, duplicated reactions, and thermodynamically infeasible cycles, but these important aspects do not contribute to the total score [141].

1.3.1 Bottom-up methods

Bottom-up methods roughly follow the workflow described in subsection 1.1.2, where coding sequences of a genome are functionally annotated gaining EC codes, and then manually translated into balanced reactions. Such reactions are then added one by one into a coherent network, which is later curated creating a dedicated biomass equation, setting metabolite compartments, filling gaps, removing thermodynamically infeasible cycles, and so on. This is clearly the worst case scenario in terms of manual effort, and it was the obligatory workflow for earlier reconstructions, which has been detailed in a 96-steps

protocol published in Nature Protocols [21]. After more then ten years, this resource still remains the gold standard for the reconstruction of high quality, manually curated GSMMs [142,143].

In the rush of automatizing the workflow, earlier reconstruction tools enabled the user to automatically convert unknown gene sequences into a set of reactions, without guaranteeing their stoichiometric consistency or mass and charge balance. Therefore, while this was a first step towards the speeding up of reconstructions, much of the manual work remained. Moreover, such tools were sometimes based on tabular-like graphical user interfaces, designed to facilitate the collection and curation of genes, reactions, and metabolites, an approach far from being adopted in automated pipelines [34,144,145]. In addition, when multiple resources were queried to retrieve reactions, such as both KEGG [30] and MetaCyc [31], the work of making an integrated, interconnected network of reactions was even more demanding [34,146,147].

When the earlier GSMMs were published, they represented reconstruction tools *per se*. Indeed, methods called "template-based" were soon developed, requiring the presence of a curated GSMM from which genes are extracted and subjected to orthologs determination (see subsection 1.1.2) with respect to an input set of genes. The identification of ortholog genes greatly improves the reaction network reconstruction speed, as the associated reactions can simply be copy-pasted updating their GPR [148]. Moreover, the biomass equation is inherited too, ready to be used or adapted according to the organism under study. The resulting draft is then extended with new reactions, based on functional annotation as described above [147].

Interestingly, the template based methods can be generalized to a prioritized list of quality GSMMs, usually sorted from the phylogenetically closest to the most curated, comprehensive model (typically belonging to a model species such as *Escherichia coli* or *Saccharomyces cerevisiae*). In this case, input genes can be associated with orthologs from different organisms, linked by GPRs to slightly different reactions (for example in terms of cofactor usage): the only transferred reaction is the one from the closer template in the sorted list. Such prioritized template-based methods are today widely used for the reconstruction of high-quality GSMMs [149,150].

1.3.2 Multi-strain methods

As described in previous sections, each genome-scale model is traditionally based on a particular genome. In the case of bacteria, this means that each GSMM can be described as "strain-specific".

Even if the definition of bacterial species describes its members as "*genomically and phenotypically coherent*" [7], it is known that strains of the same species may display genomic and eventually phenotypic differences. These differences are due to the adaptation to different environmental niches, where possible mutations, gene losses, and/or horizontal gene transfer events can determine different evolutionary patterns [9,10]. The differences between strains and closely related species are the focus of biodiversity and bioprospecting studies, where several phylogenetically close genomes are processed to select strains for desirable traits (see subsection 1.2.6).

In this kind of studies, given the predictive power of GSMMs, it would be very helpful to have a strain-specific GSMM for each evaluated strain. As the reconstruction of quality models takes considerable effort (as perfectly depicted in the gold standard protocol), there is the need to recycle the information present on an already available model, adapting it to different strains, similarly to template-based methods. After some pioneering works [114–116], a general method was developed for the fast, scalable generation of strain-specific GSMMs, one for each input genome, employing the information present in a high quality, manually curated, reference GSMM. This method was published in 2020 as an extension of the gold standard protocol [48], and it can be described as a multi-strain reconstruction.

The first phase in this protocol is to filter input genome assemblies for quality. Two basic metrics used to evaluate an assembly are the N50 and the number of contigs. The first is defined as the length of the shortest contig at 50% of the total assembly length, sorting contigs by descending length. The second is usually higher when genomes are fragmented, as the assembler is unable to produce more contiguous sequences; ideally, a finished bacterial genome has a number of contigs equal to $c + p$, where c is the number of chromosomes (usually 1), and p is the number of eventual plasmids. To filter genomes, authors suggest a sequencing coverage > 70x, assuming it will lead to satisfying N50 and number of contigs [48].

Following the filtering, remaining genomes are annotated using always the same tool for consistency (e.g., Prodigal [151]), and then a blastp [50] best-reciprocal hits (BRH) alignment (see subsection 1.1.2) is performed against the reference genes detecting strain-specific orthologs. After applying thresholds for e-value, identity and coverage, a binary homology matrix is created, having reference genes in rows, genomes in columns, and ortholog presence (1) or absence (0) in cells. For each missing orthology (cells with 0), the corresponding reference gene is searched with blastn [50], and the eventual

recovered sequence is checked for the presence of premature stop codons [48,111].

When the binary homology matrix is updated accounting for gene recovery, a copy of the reference model is made for each quality-filtered genome. Then, reference genes are subtracted, leading to an eventual loss of reactions according to the GPR. Later, for each retained reaction, the GPR is translated using strain-specific genes [48].

At this point, a strain-specific GSMM could not grow, despite the removal of strain-specific biomass precursors from the reference biomass equation and the provision of supplementary amino acids or vitamins for which the strain could present auxotrophy. For this reason, a gap-filling for biomass is applied to each GSMM using the reference as the source of gap-filling reactions. This must be based on a growth medium, better if minimal, known to support the growth of all the strains in input (provided dedicated auxotrophs handling) [48].

Finally, the protocol reaches a critical point: the addition of new reactions, not present in the reference [48]. Until now, it would have been possible to automate all the steps, but the addition of new, strain-specific reactions impose the adoption of procedures typical of the bottom-up methods, as the reaction network has to be expanded respecting balances and stoichiometric consistency.

Without considering this last expansion phase, as well as the initial genome filtering phase, the remaining steps have been implemented in a reconstruction tool named Bactabolize, published in 2023 [129]. One observation arises naturally, pointed out by the same authors of the tool: in order to really cover the strain-specificity of the genomes in input, the reference model must be representative of diverse strains in terms of modeled genes and reactions. Otherwise, the strain-specific GSMMs in output would represent just a subset of the reactions of a single strain (the one of the reference model).

This reasoning introduces the concept of pan-GSMM. These kinds of models have to represent the metabolic features not of a single genome (a single strain), but several genomes (several strains) at the same time. Indeed, they do not encode genes of a genome, like in traditional GSMMs, but representative sequences for one or more phylogenetically close species. If the goal is avoiding manual curation using automatic multi-strains reconstruction tools like Bactabolize [129], a pan-GSMM should be the only type of reference model to use (see **Chapter 2**). However, the concept of pan-GSMM has been developed also in metagenome-based modeling, where it tends to have a different

application scope; to better clarify this topic, a broader overview of pan-modeling will be given in subsection 1.3.4.

However, it must be noted that some of the published biodiversity studies based on GSMMs did not consider the step of adding new reactions, so the strain-specific models in output were just a subset of the reference strain reactions [37,134,152], which is a clearly limiting approach.

1.3.3 Top-down methods

As outlined in the previous section, it is crucial that the reference GSMM captures as much metabolic diversity as possible. Top-down methods are based on a universe, that is a GSMM covering in principle all the metabolic pathways that can be found in a particular domain of life. Since the universe is semi-curated and annotated beforehand, the GSMMs generated from it will inherit these same characteristics. Moreover, another crucial aspect of top-down methods is their ability to produce simulation-ready models, meaning that a sufficient degree of gap-filling has been applied to ensure *in silico* growth under several nutritive conditions. In such models, growth follows universal biomass equations, containing all the biomass precursors commonly found in a domain of life; for example, the same biomass equation will be used for all gram positive bacterial reconstructions, at the cost of limiting the achievable specificity. However, tools capable of generating simulation-ready models have gained significant popularity in the last five years, making genome-scale modeling more accessible to those without experience in the lengthy reconstruction process [34,44,65,66,153,154]. Even if models created by these methods are still useful to infer metabolic patterns, their accuracy is significantly lower than manually curated GSMMs, and they should be considered “drafts” [34].

Among the tools implementing top-down methods, CarveMe [44] plays a lead role. When it came out in 2018, it brought much novelty in the field of automated reconstructions. It was the first command-line program able to generate simulation-ready models in seconds; moreover, it proposed a new gap-filling algorithm relying on genetic evidence (sequence alignment scores), designed to *“enforce network connectivity”* [44]. The strengths of this particular implementation were evident when processing a high number of fragmented genomes, leading CarveMe [44] to pave the way for metagenomics-based microbial community studies [139,155–157].

The three universes implemented in CarveMe [44] (gram-negative bacteria, gram-positive bacteria, and cyanobacteria) are derived from an automated

merge of the high-quality bacterial models available in the BiGG database [56], followed by a semi-automatic curation (for example to improve stoichiometric consistency). While this confers wide metabolic coverage, the resulting universes suffer from issues inherited from BiGG, as described in subsection 1.1.1. This means that CarveMe-generated models [44] may contain duplicated metabolites and reactions, as well as unbalanced reactions both in mass and in charge.

Moreover, not only the universal reaction networks are derived from BiGG, but also the reference genes (BiGG genes) which are used to connect the input coding sequences to the reactions of such networks. This implies that the gene database internally used by CarveMe [44] is small, containing genes from a maximum of 22 bacterial organisms not included in the *Escherichia* / *Shigella* species complex, as described in subsection 1.1.2. Therefore, the level of genetic representativeness is low, and when CarveMe [44] is used to model distant-from-model organisms, several genes could not be included in the GSMM, compensating for their absence with the extensive gap-filling, contributing to the high degree of orphan reactions (see **Chapter 2**).

In practical terms, Diamond [158,159] is used to perform a one-way, fast amino acid alignment of the input sequences against the BiGG genes (**Table 1.1**). This alignment is then filtered to retain HSPs (high-scoring segment pairs) with a bit score of at least 100. To convert the input sequences into GPRs for the universal reactions, a GPR table is used (**Table 1.2**), which represents a central element of the reconstruction process. This table contains, for each BiGG gene, all the possible model-specific protein complexes in which the BiGG gene could be involved, together with the corresponding universal reaction. After filtering the alignment, a top-scoring sequence is selected for each BiGG gene based on the highest bit score, discarding the others. The IDs of the top-scoring sequences are then joined by "AND" operators to form preliminary GPRs based on the protein complex variants listed in the GPR table. To give some examples based on **Table 1.1** and **Table 1.2**, the protein complex variant "M2_G2 + M2_G3" is translated into "CDS_4 and CDS_5", while "M3_G1 + M3_G2 + M3_G3" is translated just into "CDS_6". Each preliminary GPR is also associated with a "protein complex score", computed as the mean of the bit scores of the sequences involved. Next, preliminary GPRs for the same reactions are joined by OR operators to form the final GPRs. For example, reaction R1 will receive GPR equal to "CDS_1 or CDS_2", while the GPR of reaction R3 will remain "CDS_6". Similarly to the previous step, each final GPR is associated with a "reaction score" given by the maximum protein complex score among the joined preliminary GPRs. These final scores are normalized by their median and then

used to feed the gap-filling algorithm, which will prioritize the incorporation of reactions with higher genetic evidence (i.e., higher final scores).

Table 1.1. Schematic example of Diamond alignment table produced by CarveMe [44].

	CDS	BiGG gene	bit score
0	CDS_1	M1_G1	240
1	CDS_1	M1_G2	55
2	CDS_2	M1_G2	56
3	CDS_2	M2_G1	200
4	CDS_3	M2_G2	190
5	CDS_4	M2_G2	191
6	CDS_5	M2_G3	210
7	CDS_6	M3_G1	215

Table 1.2. Schematic example of GPR table internally used by CarveMe [44].

	BiGG gene	protein complex variants	BiGG reaction	BiGG source model
0	M1_G1	M1_G1	R1	M1
1	M1_G2	M1_G2 + M1_G3 + M1_G4	R2	M1
2	M1_G2	M1_G2 + M1_G3 + M1_G5	R2	M1
3	M1_G3	M1_G2 + M1_G3 + M1_G4	R2	M1
4	M1_G3	M1_G2 + M1_G3 + M1_G5	R2	M1
5	M1_G4	M1_G2 + M1_G3 + M1_G4	R2	M1
6	M1_G5	M1_G2 + M1_G3 + M1_G5	R2	M1
7	M2_G1	M2_G1	R1	M2
8	M2_G2	M2_G2 + M2_G3	R2	M2
9	M2_G2	M2_G2 + M2_G4	R2	M2
10	M2_G3	M2_G2 + M2_G3	R2	M2
11	M2_G4	M2_G2 + M2_G4	R2	M2
12	M3_G1	M3_G1 + M3_G2 + M3_G3	R3	M3
13	M3_G2	M3_G1 + M3_G2 + M3_G3	R3	M3
14	M3_G3	M3_G1 + M3_G2 + M3_G3	R3	M3

The above examples highlight two additional issues associated with the CarveMe [44] reconstructions. A first issue is that highly similar, equally valid input coding sequences are ignored during the building of GPRs. Indeed, for example, the final GPR of reaction R2 should be “(CDS_3 or CDS_4) and CDS_5”, while it will instead be reported as “CDS_4 and CDS_5”, with a missing gene. This behavior can lead to an inaccurate description of the organism, for example in case of naturally occurring gene duplications.

The second issue is that the completeness of original protein complexes is never checked and it can be violated. Indeed, for instance, reaction R3 can be imported in the reconstruction with the incomplete GPR “CDS_6”, corresponding to “M3_G1”, while the reaction should probably be not included at all, due to the absence of a second and third term associated to M3_G2 and M3_G3, respectively. Again, this behavior likely leads to an inaccurate description of the organism being modeled. Both the issues were publicly reported in the CarveMe [44] GitHub repository under the IDs #180 and #182.

Despite the embodied amount of uncertainty, draft GSMMs generated by CarveMe v1.5.2 [44] (see also **Chapter 2**) are today widely adopted and they are featured in many relevant publications which use them as they are [160–165].

Another widely adopted top-down tool is gapseq [65], which was published three years after CarveMe [44]. It is based on the SEED namespace and its universal network of reaction is derived from the ModelSEED biochemistry database [29], manually curating for thermodynamically infeasible cycles producing free ATP (see subsection 1.2.3). The resulting network is much wider compared to CarveMe [44], so its metabolic coverage should be wider [20,60]. The internal gene database is also much larger, being based on all the metabolic and transport genes contained in UniProt [166] and TCDB [67] databases, respectively. Since the gene database is so comprehensive, the execution times are much slower compared to CarveMe [44]. Another difference with respect to CarveMe [44] is that universal reactions are annotated in pathways, and the presence/absence of metabolic pathways is evaluated in conjunction with GSMM reconstruction. The default alignment-based gap-filling algorithm is also different, but tries to reach the same goals of CarveMe [44], taking into account many potential growth conditions at the same time.

During the reconstruction process, transport reactions are selected from the universe giving priority to a transport mechanism similar to the one detected by aligning on the TCDB genes. This represents an issue: for example, let's suppose that genes encoding an active (ATP-consuming) transporter for a specific substrate are detected; if the only transport mechanism present in the universe for that substrate is a facilitated diffusion, the generated model will contain an erroneous description of the transport reaction, with potential consequences during simulations [65].

Moreover, GPRs generated by gapseq [65] derive from text mining of the sequence FASTA headers of the UniProt genes. When they were compared against those from reference models (assuming their GPR as correct), a 19% of equivalence was reached by gapseq [65], which was better than the 6% obtained by CarveMe [44], but still very low, highlighting a great need of improvements in this area.

1.3.4 Pan-modeling methods

In subsection 1.3.2, the concept of a pan-GSMM was introduced in the context of biodiversity or bioprospecting studies, where closely related strains are subjected to screening or clustering based on their strain-specific metabolic features. In this panorama, the focal point is to shed light on strain-specificity, meaning that the first goal is to create faithful, accurate, high quality models for each input strain. Indeed, in such studies, input genomes are always complete or at least filtered for quality, and the pan-GSMM is curated to be as comprehensive as possible, representing also rare metabolic features, since all the strain-specific models will be derived from it by subtraction [37,114–116,134,152,167,168].

However, the concept of a pan-GSMM was recently developed also in the context of metagenome-based modeling. Here, pan-GSMMs fulfill a different need: they are instrumental to cope with the incompleteness and contamination of metagenome-assembled genomes (MAGs). Indeed, when a low-quality MAG is used to reconstruct a metabolic network, the resulting GSMM will be largely incomplete and/or will present spurious reactions due to exogenous sequences; spurious reactions, together with the high number of orphan reactions resulting from the extensive gap-filling phase, will likely result in an unreliable model [169].

To solve the challenge of metagenome-based modeling, the tool pan-Draft [169] has just been released by De Bernardini and colleagues. Its implemented method requires a collection of pre-made GSMMs generated with gapseq [65], using MAGs belonging to the same SGB (species-level genome bin). The SGB is built by grouping MAGs sharing an ANI (average nucleotide identity) of at least 95%, therefore it ideally clusters genomes of the same species. At this point, pre-made GSMMs are processed to extract the frequency of occurrence of each reaction. Then, based on the hypothesis that contaminating sequences are non redundant among the MAGs of the SGB, a frequency threshold (e.g. 6%) is applied to discard spurious reactions as well as those reactions that are correct but rare. Frequency of occurrence is finally converted into weights to be used during a gap-filling step which brings to the creation of a so-called “species-level” GSMM [169].

Such species-level GSMMs are designed to be the objects to work with in metagenomics-based microbial community modeling. They provide a more robust description of the organisms acting in a community, as incompleteness and contamination of MAGs is reduced by combining them together, retaining their prevalent metabolic features. However, the concept of pan-GSMMs

leading to such species-level models goes in the opposite direction compared to the concept of pan-GSMMs used in multi-strain reconstruction. Indeed, in the first case strain-specificity is completely lost [169] in favor of a consensus/mean reconstruction, while in the second case strain-specificity is retained and emphasized as it is needed for the subsequent generation of strain-specific GSMMs.

Chapter 2 | Gempipe: a tool for drafting, curating and analyzing pan and multi-strain reconstructions of genome-scale metabolic models

2.1 Introduction

Different strains of the same bacterial species can exhibit marked differences at the phenotypic level, such as the ability to catabolize different substrates, the presence of specific auxotrophies, or the acquisition of biosynthetic pathways through lateral gene transfer [10,170]. From an ecological point of view, such differences can explain, for example, why specific ecological niches are preferred over others [171], or which metabolic features distinguish one species from another [172]. From a biotechnological perspective, strain-specificity is key during the screening of new strains to be used for a bioprocess. For instance, in the food industry, the presence of certain amino acid degradation pathways can lead to specific flavors [173], and the presence of complete vitamin biosynthetic pathways is crucial for fortified food production [174].

Genome-scale metabolic models (GSMMs) are established systems-biology tools able to recapitulate the metabolic potential encoded by a genome. Assuming a steady state and given a biomass composition, their constraint-based simulations enable predictions of cellular growth under specific nutritive inputs [1]. Therefore, the availability of a GSMM for each strain of interest enables *in silico* screening of phenotypic characteristics, offering a faster and more cost-effective alternative to traditional experimental methods. This approach can provide valuable insights that can subsequently be validated through experimental verification.

Unfortunately, the creation of high-quality GSMMs is time expensive because manual curation is required, which involves tasks such as the gap-filling of missing reactions and the removal of thermodynamically infeasible cycles [21]. The bottleneck imposed by manual curation was the main driver for the development of automated pipelines [44,65,66,146,147], some of which quickly became established thanks to their ability to produce simulation-ready GSMMs, using a universe-based top-down approach [34].

Since the first pioneering work by Monk and colleagues on *E. coli* [114], the general steps for the multi-strain reconstruction of GSMMs remained quite the same [37,115,116,134,152,167,168]. Briefly, genomes are collected and filtered for quality. Then, genes are predicted and clustered together, creating orthologous gene families sometimes referred to as the pangenome. One of the strains includes a high-quality, manually curated GSMM that is used as reference. For each strain, a copy of the reference is created and all the genes that do not have an orthologous with the reference are subtracted along with their metabolic reactions. Finally, gap-filling is usually limited to a minimal medium as the starting genomes are quality-filtered and an extensive gap-filling could hide true strain specificities.

The above reference-based method was formalized in 2019 by Norsigian and colleagues [48], where metabolic functions are inherited from the reference GSMM after orthologous genes are detected via a blastp [50] best reciprocal hits (BRH) alignment. This method was then implemented in Bactabolize [129], a recent tool published in 2023.

This approach has proven effective in exploring metabolic diversity at the strain level, provided the availability of a curated and phylogenetically close GSMM as a reference – something that is often lacking, particularly for non-model organisms. However, this method has a key limitation: the strain-specific GSMMs in output represent only a subset of the reactions contained in the reference, excluding unmodeled strain-specific reactions. The protocol [48] requires a manual curation of the strain-specific GSMMs, adding new reactions that were not modeled in the reference, a step that was not automatized in Bactabolize [129]. For this reason, to fully capture strain-specific metabolic features, a pan-GSMM that encompasses the metabolic diversity of the entire species (or genus) should be provided as reference instead of a GSMM modeling a specific strain. However, while a curated strain-specific GSMM is already rare, obtaining a comprehensive pan-GSMM for the species under study is even more challenging.

The production of a pan-GSMM is more complex and time-consuming than a single genome GSMM. While such pan-models are still rare in literature, the most prominent example of a pan-GSMM is probably the KpSC-pan developed for the *Klebsiella pneumoniae* species complex, that required years of development, from the first core of 37 manually-curated strain-specific GSMMs in 2022 [134] to the latest version of the pan-GSMM KpSC-pan-v2 in 2024 [175], not to mention the creation of the first *Klebsiella* reference back in 2011 [176]. Few other curated pan-models have been published [129], including those for *Escherichia coli* [177], *Salmonella enterica* [115] and *Bacillus subtilis* [152,168].

Given the vast number of strain-specific genome sequences now available and the scarcity of comprehensive pan-GSMMs, there is a growing need for tools that perform multi-strain genome-scale metabolic reconstructions efficiently, even in the absence of a pan-GSMM. These tools should not only be capable of using a reference strain as a starting point, but also of integrating new, strain-specific reactions to fully capture the metabolic diversity across strains. This would allow for a more accurate representation of the complete biodiversity at the strain level.

In this work Gempipe is introduced, a novel package that, to the best of our knowledge, is the first to offer both expanded reference-based and reference-free pan and multi-strain reconstructions of GSMMs. Within Gempipe, genomes are automatically downloaded and subjected to rigorous quality filtering based on both technical and biological criteria. Genes are subsequently annotated and grouped into clusters, followed by a comprehensive gene-recovery to mitigate errors arising from genome assembly or gene calling processes. These gene clusters are then used to build a reference-free reconstruction, leveraging the semi-curated metabolic reaction database from CarveMe [44]. This process employs enhanced gene-to-reaction (GPR) association rules, accounting for different gene isoforms while preserving the original enzyme complexes as defined in BiGG [56].

The reference-free reconstruction serves as a source for introducing new reactions to an optional user-provided reference GSMM, incorporating strain-specific features beyond the scope of the reference. This expansion respects the design decisions of the reference, including how metabolite formulas, charges, and reaction balances were defined. The resulting draft pan-GSMM is then annotated *de novo* with accessions from various databases, with the option to remove duplicated metabolites and reactions. To facilitate the manual curation of the draft pan-GSMM, a dedicated application programming interface (API) is provided. While manual curation is actively encouraged, Gempipe offers also a fully automated reconstruction mode ("gempipe autopilot").

Once the pan-GSMM is finalized, it is utilized in the multi-strain reconstruction, generating strain-specific GSMMs for each input genome or proteome by leveraging gene clustering data, ensuring biomass production across a range of user-defined growth media.

Gempipe was evaluated using 3 different datasets and results were compared with the current state-of-the-art reference-free and reference-based pipelines, namely CarveMe [44], gapseq [65], and Bactabolize [129]. To validate the

multi-strain reconstruction, substrate usage predictions were compared with strain-specific Biolog® phenotypic screenings taken from literature.

Gempipe can be easily installed using the dedicated conda package (https://anaconda.org/bioconda/gempipe) and its source code is freely available on GitHub (https://github.com/lazzarigioele/gempipe). Moreover, a comprehensive documentation is available on ReadTheDocs (https://gempipe.readthedocs.io/en/latest/).

2.2 Results

2.2.1 Pipeline

Gempipe is a Python pipeline divided into three main components to be used in the following order (**Figure 2.1**): 1) the command-line program "gempipe recon", which creates a draft pan-GSMM and an associated gene presence / absence matrix (PAM), starting either from genome sequences or proteomes, and using an optional reference GSMM; 2) the Gempipe API, useful to perform the manual curation of GSMMs, including the draft pan-GSMM previously generated; 3) the command-line program "gempipe derive", which derives strain-specific GSMMs, starting from the PAM and the curated pan-GSMM. Moreover, even if the manual curation of the draft pan-GSMM is strongly encouraged, Gempipe offers an additional automated mode through the "gempipe autopilot" command-line program, which derives strain-specific GSMMs directly from the input genomes or proteomes, performing an automatic gap-filling on the draft pan-GSMM (**Figure 1**). Finally, the Gempipe API can be used once again to analyze the strain-specific GSMMs in output.

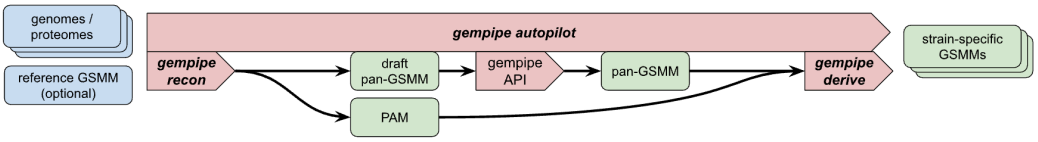

Figure 2.1. General overview of the Gempipe pipeline.

Accepted inputs

Gempipe can start from either genome assemblies or proteomes (**Figure 2A**), which are here defined as all the genome coding sequences in amino-acidic format. Therefore, the choice depends on whether or not the strains under study already possess a reliable gene annotation. If proteomes are available, they can be provided either as multi-fasta files or as GenBank files. Otherwise, local genome assemblies in FASTA format can be provided. Alternatively, users

can specify one or more NCBI species-level taxonomy IDs, and Gempipe will automatically download all the available assemblies from NCBI [178] for the selected species.

Genome filtering and gene prediction

When starting from genome assemblies, a filtering process based on technical and biological metrics is applied to retain only high-quality assemblies (**Figure 2.2A**). First, genes are predicted using Prodigal [151] and then assessed with BUSCO [179] to evaluate genome completeness. Next, the N50 value and the and number of contigs are calculated. This process enables users to retain high-quality genomes by setting thresholds for the maximum number of missing or fragmented BUSCO's single-copy orthologs, minimum N50, and maximum number of contigs.

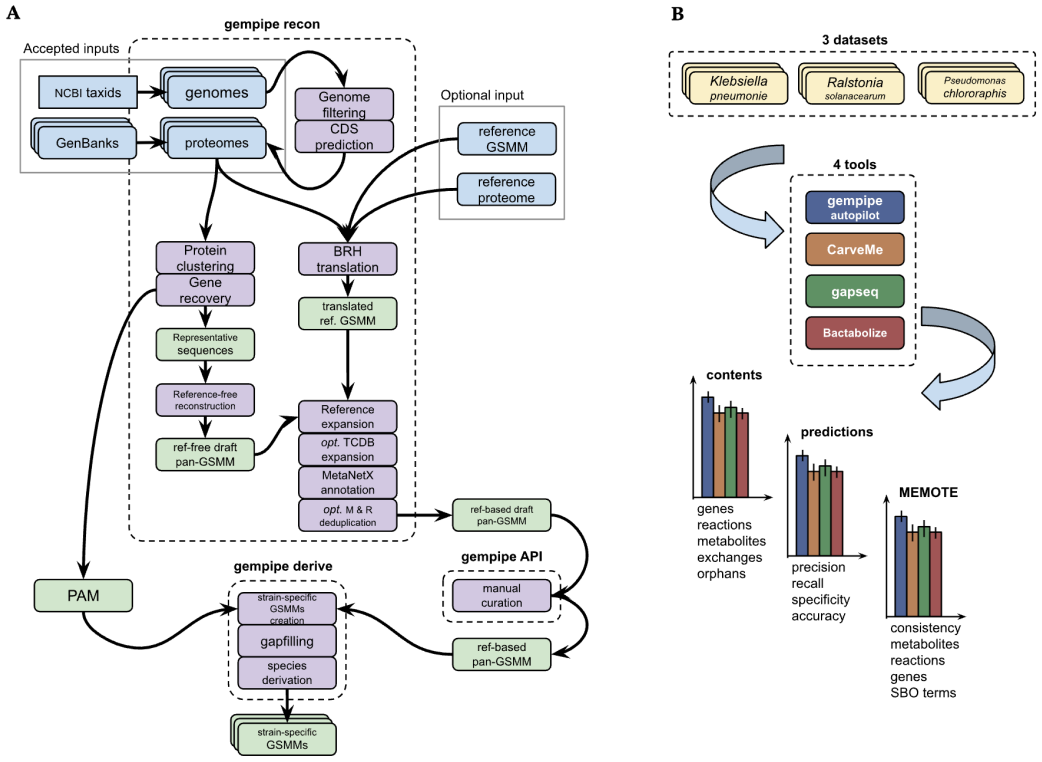

Figure 2.2. Overview of Gempipe and its validation. **A)** Simplified schema of the Gempipe pipeline. **B)** Simplified schema of the validation procedure.

Protein clustering and gene recovery

Once proteomes are obtained either from direct input or from internal *de novo* annotation, coding sequences are grouped into clusters based on aminoacidic sequence identity (**Figure 2.2A**). A representative sequence is taken from each cluster and labeled with the cluster ID. After the clustering, a presence-absence

matrix is generated, with clusters as rows and strain genomes as columns. The matrix cells contain the gene IDs from a specific genome that are associated with a given cluster.

When genome assemblies are provided, Gempipe uses the representative sequences of clusters to execute a gene recovery in order to identify potentially unannotated genes in genomes (**Figure 2.2A**). For each genome, Gempipe focuses on the gene clusters still without any match in that specific genome. The implemented algorithm considers the following two scenarios. Scenario 1: presence of premature stop codons that break the protein. To address this, when two consecutive coding sequences together match, with high identity, the representative sequence of a missing cluster, the algorithm assumes this is caused by sequencing or assembling errors, and restores the entire sequence into the correct cluster. Scenario 2: failure of the prediction algorithm to identify coding sequences, leading to unpredicted genes. To address this, the algorithm first aligns the representative sequences of missing clusters to the genome excluding all the previously predicted genes, then it evaluates potential overlapping coding sequences [180]; only high-identity and high-coverage alignments are considered to populate the PAM (see 2.5.2 Supplementary Methods).

Reference-free reconstruction

Representative sequences of clusters are aligned on the same gene database used by CarveMe [44], derived from the bacterial genes collected in the BiGG database [28], each of them linked to a different BiGG model. This alignment is needed to create a reference-free reconstruction that will be used either as a repository of reactions for the expansion of the reference GSMM (if provided), or to constitute the draft pan-GSMM when no reference is provided (**Figure 2.2A**). This reference-free reconstruction is created by copying reactions from the gram-positive or gram-negative universe used by CarveMe, after filtering the alignment with user-selectable thresholds like the minimum percentage coverage and the minimum percentage aminoacidic sequence identity.

Each reaction in the BiGG database can appear in several different BiGG models, and each time it can be described with a different protein complex through its gene-to-reaction association (GPR). For example, the mannitol transport reaction "MNLpts" is encoded with 2, 3 or 4 AND-linked subunits in model iCN900 [38], iYO844 [122] or iNF517 [181], respectively. In Gempipe, a reaction is copied from the selected universe only if at least one of the BiGG's original protein complex definitions is fully satisfied. For example, in the case of "MNLpts", if just 1 subunit is found, then the reaction will not be included.

Moreover, if two or more slightly different proteins align equally well on the same BiGG gene, all the involved GPRs will take into account the alternative isoforms. It is believed that these two design decisions lead to more accurate GPRs, solving two known public CarveMe issues, #180 and #182.

Bacterial models in BiGG v1.6 [56], which are not part of the *Escherichia* / *Shigella* species complex, are just 22. This means that BiGG is biased towards model organisms, and the few genes it contains can hardly represent the known bacterial diversity. To alleviate this issue, not only representative sequences of clusters are aligned on the CarveMe/BiGG gene database, but they are also functionally annotated with eggNOG-mapper [182]. Then, for each gene cluster involved in the reference-free reconstruction, alternative equivalent clusters are retrieved by searching for those having the exact same eggNOG-mapper functional annotation. For each extra gene cluster retrieved, the involved GPRs in the reference-free reconstruction are expanded accordingly.

Reference expansion

Users can specify an optional reference GSMM and its associated proteome. In this case, the reference is the cornerstone of the reconstruction process, providing key features such as the non-growth associated maintenance energy (NGAM) [21] and the biomass equation [59], which would be otherwise inherited from the CarveMe universe [44]. Moreover, the reference is expanded with new reactions taken from the reference-free reconstruction, to include new strain-specific reactions (**Figure 2.2A**). In this expanded reference-based reconstruction mode, orthologs are determined for each strain with a BRH alignment with the reference proteome, similarly to the approach suggested in the Nature Protocol Extension [48]. This enables the mapping of the reference genes to the respective gene clusters.

The transfer of reactions from the reference-free reconstruction to the reference respects the following principles. (i) If the reaction ID is already included in the reference, the corresponding GPR is updated including eventual new gene clusters not yet included. (ii) If the reaction ID is not included in the reference, all possible synonym reactions are searched, having the same reactant and product IDs, ignoring protons. If a synonym reaction is found in the reference, the expansion is limited to new gene clusters that may be included in its GPR. (iii) If a reaction is new, it is added to the reference model. New metabolites and gene clusters are also transferred if needed. If a metabolite is encoded both in the reference and in the reference-free reconstruction with the same ID but a different formula and/or charge, the metabolite definition of the reference is

maintained. During the reference expansion, however, users are allowed to superimpose specific reaction balances and metabolite formulas and charges.

Transport reactions expansion

The accuracy of any simulation involving cellular import or export of molecules heavily relies on the quality of transporter annotations [183]. Few published tools are available for this purpose, like TranSyT [184] and TransAAP [185], but they could not be integrated within Gempipe. Indeed, a tool was needed that would (i) run locally (no web-servers), (ii) build BiGG-compatible reactions, and (iii) provide GPR for the whole transport complex. Gempipe implements its own annotation system, that can be activated on users request, and that integrates the transport reactions coming from the optional reference and the CarveMe universe [44] with others derived from the TCDB [67] (see 2.5.2 Supplementary Methods).

Draft pan-GSMM annotation

Once the draft pan-GSMM has been built, its metabolites and reactions are annotated *de novo* by means of MetaNetX v4.4 [186] (**Figure 2.2A**). This process adds links to several databases, including HMDB [187], KEGG [30], MetaCyc [31], Rhea [188], Chebi [189], and many others, including MetaNetX itself.

This annotation can be optionally used to detect and remove duplicate metabolites in the draft pan-GSMM, such as for example “glc_D_B” and “glc__bD”, both representing beta-D-glucose. Duplicates are replaced by a consensus metabolite, giving priority to metabolites in the reference GSMM, if available. If duplicates exist in the reference, they are retained.

MetaNetX annotation [186] is also used to remove duplicate reactions, with the same logic applied: the consensus reaction is selected giving priority to those in the reference GSMM. If duplicates are present in the reference, they are kept. Otherwise, the GPR of the consensus reaction is updated to incorporate all the gene clusters appearing among the duplicates. The final draft pan-GSMM represents the main output of the "gempipe recon" command-line program, along with the PAM (**Figure 2.1** and **Figure 2.2A**).

Facilitated manual curation

The manual curation of the draft GSMMs encompasses several different tasks such as for example adding missing reactions and gene clusters, and removing erroneous energy-generating cycles (EGCs) [21]. Gempipe includes a Python application programming interface (API), which was developed to speed up the

curation using Jupyter Notebooks [190], not to replace the community effort MEMOTE [70] (**Figure 2.2A**). With this API, for example, users can quickly check which biomass precursors cannot be synthesized and perform a dedicated gap-filling; moreover, new reactions and metabolites can quickly be added, EGCs can be detected, and limiting input fluxes can be shown with a sensitivity analysis; in addition, the Gempipe API can be used to simulate the Biolog® Phenotype MicroArray™ (PM) screening system.

The API works with any GSMM loaded in COBRAPy [81], but becomes especially useful for curation of draft pan-GSMMs produced by Gempipe, providing specific functions. In this context, for example, the PAM can be functionally queried to facilitate the identification of new gene clusters to model. Therefore, it provides a quick and convenient way to check the main sanity standards and perform gap-fillings before going on deriving strain-specific GSMMs with "gempipe derive". Regarding the implemented functions, further information is provided in 2.5.2 Supplementary Methods, and tutorials are available on the Gempipe documentation.

Strain-specific GSMMs generation

Once the pan-GSMM has been sufficiently curated, "gempipe derive" can be used to derive strain-specific GSMMs. Briefly, a copy of the pan-GSMM is made for each quality-filtered genome or proteome. Then, gene clusters are removed according to the PAM, leading to an eventual loss of reactions. Next, each GPR is translated from gene clusters to the actual strain-specific gene IDs.

Finally, each strain-specific GSMM is gap-filled to ensure biomass production on one or more user-defined growth media, usually minimal, known or assumed to enable growth of all the strains in input. During the gap-filling, a minimal flux through the objective reaction can be specified. Moreover, the gap-filling can optionally be skipped, which is useful for example when auxotrophies need to be studied on a minimal medium [177]. At this point, strain-specific GSMMs have the minimum requirements to be used in simulations.

Multi-strain analysis

Once strain-specific GSMMs have been generated with "gempipe derive", users can request simulations of the Biolog® PM screening. Moreover, to better evaluate potential metabolic patterns among the input strains, binary metabolic features can be predicted and used to divide strains into metabolically coherent clusters. These binary metabolic features include, for example, presence of reactions in GSMMs, capability to catabolize alternative C,

N, P or S substrates, and presence of auxotrophies for amino acids and vitamins. These features can be combined together to create dendrograms which could be called “phylometabolic” trees, as strains sharing similar metabolic potential are shown in close proximity, conceptually similar to phylogenomic trees.

Once metabolic clusters are localized on a phylometabolic tree, the cluster attribute can be compared to other strain-specific attributes such as the environmental niche or the species of origin. Moreover, discriminant metabolic features, whose presence or absence characterizes each cluster, can be extracted. These are examples of multi-strain analysis implemented within specific functions of the Gempipe API.

2.2.2 Validation

To validate Gempipe for the reconstruction of strain-specific GSMMs (**Figure 2.2B**), publicly available datasets were searched with the following characteristics: (a) organisms belonging to the same species or species complex; (b) publicly available genome assemblies; (c) strain-specific Biolog® PM screenings at least for carbon sources; (d) a published manually curated, phylogenetically close GSMM, belonging at least to the same genus. Three datasets were collected: 37 strains belonging to the *Klebsiella pneumonie* species complex [134]; 11 strains belonging to the *Ralstonia solanacearum* species complex [149]; 36 strains of *Pseudomonas chlororaphis* [191].

Gempipe was compared with the current state-of-the-art reference-free reconstruction pipelines, namely CarveMe [44] and gapseq [65], as well as a recent reference-based pipeline specifically developed for strain-specificity studies, Bactabolize [129]. The *--reference* option was used in CarveMe, even though in reality the reconstruction is reference-free, as highlighted in 2.3 Methods. Using each of these tools, a GSMM was reconstructed for each strain. In addition to analyzing the content of the GSMMs, their ability to predict substrate utilization prior manual curation was assessed, assuming literature Biolog® PM data as ground truth.

Content comparison and similarity between tools

After reconstructing the strain-specific GSMMs with the four tools, their content was compared against their relative manually-curated reference in terms of number of genes, reactions, metabolites, and exchange reactions (**Figure 2.3**). Moreover, every tool was compared against each other using the mean Jaccard index of the reaction content (**Figure 2.S1**).

By leveraging its hybrid reconstruction mode, Gempipe produced strain-specific GSMMs that were based on the manually curated reference, but also expanded with new contents. The reference coverage in terms of reactions and metabolites was maximized, and was in line with purely reference-based reconstructions (Bactabolize), if not better. On the other hand, Gempipe reconstructions went beyond the reference with respect to the number of reactions and metabolites, aligning with the reference-free methods.

Gempipe models were more similar to Bactabolize ones, with CarveMe models in second position probably due to the BiGG-based database in common. However, it is evident that Gempipe had better reference coverage compared to Bactabolize, with a lower Jaccard index only for the *Klebsiella* dataset due to the addition of new reactions during the reference expansion.

gapseq models, despite having a consistently higher number of metabolites compared to the other tools, were the most divergent from the reference. Anyway, the conversion between SEED IDs and BiGG IDs provided by MetaNetX [186] is not perfect, so the metrics reported for gapseq are likely underestimated.

Regarding genes, not only a consistently higher number of modeled genes was observed in Gempipe models, but also a consistently higher reference coverage, probably due to the cluster-based reconstruction system and the custom-tailored gene recovery.

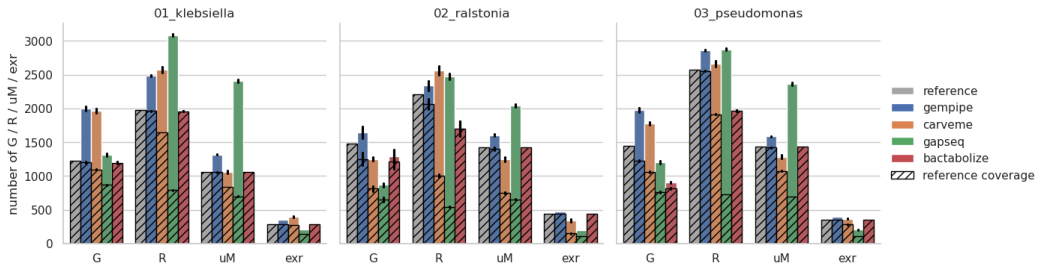

Figure 2.3. Comparison of the general reconstruction metrics. 'G': number of genes; 'R': number of reactions excluding exchange reactions; 'uM': number of unique metabolites, i.e. not considering their compartment; 'exr': number of exchange reactions. Bar height corresponds to the mean between strains, while error bars represent standard deviations. Hatched area represents contents in common with the reference (reference coverage).

Comparison with the Biolog phenotypic screenings

To evaluate the ability to recapitulate phenotypic traits, publicly available binarized Biolog® PM data were used as a benchmark (**Figure 2.4**), meaning that the kinetic signal was converted into a binary response "can grow" / "cannot grow" in the source publication. The *in silico* simulations were performed with the Gempipe API, which are based on the flux-balance analysis (FBA) [27]. The cases of no growth and infeasible solution were distinguished: the first case is when no growth is observed *in silico* while the solver status is "optimal"; the second case is when the linear programming problem defined by the FBA has no valid solution, not even 0, corresponding to the solver status "infeasible" (**Table 2.S4**). Moreover, to make the comparison fair among the tools, it was not performed any manual curation on the draft pan-GSMMs created by Gempipe (pan-GSMMs were generated with "gempipe autopilot"), and every substrate requiring a manual curation of the draft pan-GSMMs (in terms of the additions of genes or reactions) was excluded from the comparison (see 2.3 Methods).

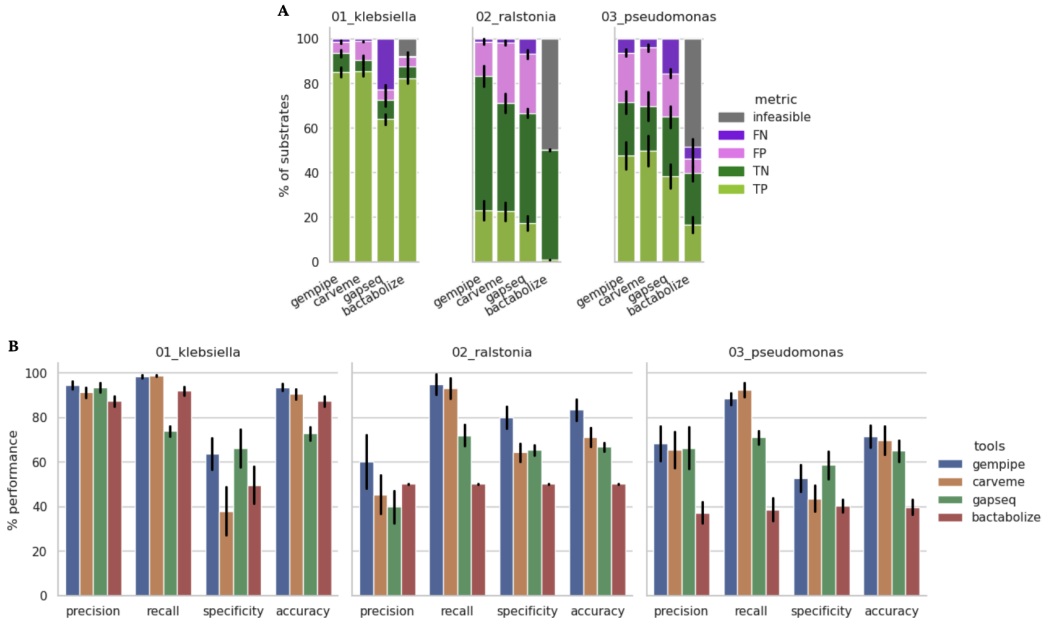

Figure 2.4. Comparison between experimental and simulated Biolog® PM growth assays. Bar height corresponds to the mean between strains, while error bars represent standard deviations. **A)** Outcome of the comparison considering single substrates. TP: true positive; TN: true negative; FP: false positive; FN: false negative; infeasible: FBA without solution. **B)** Overall performance of the substrate utilization prediction. Infeasible simulations have been penalized (see 2.3 Methods).

In general, Gempipe performed better or in line with the other tools (**Figure 2.4A**). Its novel hybrid reconstruction method, based both on a reference and on a reference-free draft, resulted in an improved accuracy in every dataset (**Figure 2.4B**).

While CarveMe ranked second in terms of mean accuracy, it was observed that its internal gap-filling algorithm, designed to *"enforce network connectivity"* [44], tends to maximize the number of substrates for which growth is predicted; this leads not only to a reduced number of FN, resulting in better recall, but also to a greater number of FP, resulting in a detrimental specificity score. Despite using the same gene databases and reaction universes of CarveMe, the different and more conservative reconstruction method of Gempipe resulted in a greater representation of the no-growth phenotype. In addition, the difference with CarveMe is more evident when modeling a genus that still does not have a reference GSMM deposited in BiGG [56], and therefore cannot be part of the CarveMe gene dataset and universes, like in the case of *Ralstonia*.

However, the presence of some FP could not only be due to the erroneous introduction of reactions but also to an intrinsic limitation of this (classical) approach of genome-scale metabolic modeling, which does not account for any type of regulatory mechanisms such as, for example, the transcriptional repression of genes [192].

In this analysis the absence of exchange reaction is evaluated as inability to utilize the correspondent substrate. Indeed, while exchange reactions are artificial (not biological) reactions and thus not associated with a GPR (see **Chapter 1**), their inclusion by automated reconstruction tools is usually bonded to the inclusion of some associated transport reaction; when genes underlying transport reactions are not detected, the corresponding exchange reactions are not introduced (unless forced during the gap-filling phase). gapseq, in general, had a lower number of TP compared to CarveMe and Gempipe, as well as a higher number FN. While the Biolog® simulations provided by the Gempipe API are compatible with the SEED namespace [29], the number of Biolog®-compatible exchange reactions in gapseq models could be lower with respect to other tools, as already showed elsewhere [129], which could explain this behaviour.

The purely reference-based method implemented in Bactabolize failed to recover orthologs for many reference genes, resulting in a highly gapped reaction network, despite the medium-specific gap-filling step in common with the other tools. With many important reactions missing, a considerable number of FBAs performed by Biolog® simulations resulted as "infeasible", preventing

the prospective usage of the Bactabolize-generated GSMMs, in particular for the *Ralstonia* and *Pseudomonas* datasets.

MEMOTE metrics

As an additional comparison between tools, the scores obtained by MEMOTE [70], a widely used tool for assessing the quality of GSMMs (**Figure 2.5**), were evaluated. Five metrics are merged into a total score: stoichiometric consistency and metabolite, reaction, gene, and Systems Biology Ontology [193] (SBO) annotations. Of these, stoichiometric consistency, the condition for which a positive mass can be associated to all metabolites in a GSMM [41], is weighted the most in the total score [70].

In general, gapseq reconstructions invariably achieved high scores in terms of stoichiometric consistency, whereas CarveMe reconstructions were more variable. Conversely, gapseq annotations resulted as scarce or totally absent for reactions, genes, and SBO terms, leading to lower total scores. The stoichiometric consistency and annotations provided by Bactabolize are completely inherited from the reference, meaning that strain-specific GSMMs cannot score better than the reference itself. In contrast, Gempipe *de novo* annotation based on MetaNetX [186] produced consistently higher scoring annotations, particularly for metabolites and reactions. This not only resulted in higher total scores, but also allowed Gempipe to surpass the metrics of the reference GSMMs.

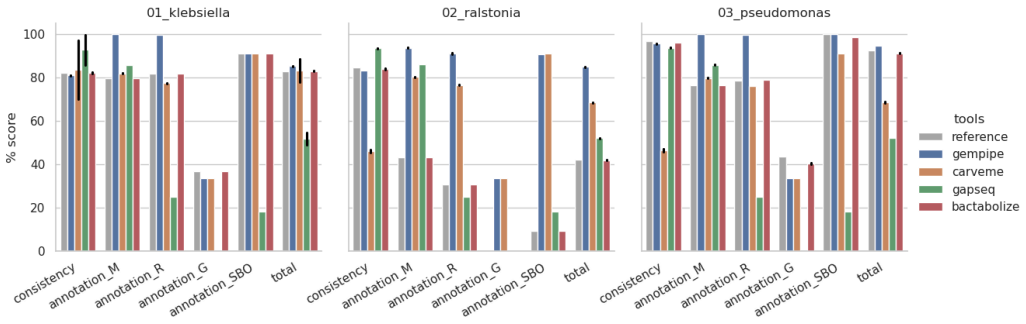

Figure 2.5. Comparison of the MEMOTE metrics and the MEMOTE total score. Bar height corresponds to the mean between strains, while error bars represent standard deviations.

Effects of gene recovery

As reported above, Gempipe provides a gene recovery feature to cope with possible errors due to sequencing, assembling, or gene prediction. This recovery is divided in three steps executed in order: the first step tries to reconstitute proteins broken in two pieces, the second searches for missing genes along

unconsidered genomic regions, while the third deals with overlapping open reading frames (see 2.5.2 Supplementary Methods).

In general, as expected, the entire gene recovery procedure was conservative for all the tested datasets, as the overall number of recovered genes (**Table 2.S5**) and reactions (**Table 2.S6**) contained in strain-specific GSMMs (**Figure 2.6**) resulted low compared to the total (**Figure 2.3**), with a median of 5 recovered genes and 2 recovered reactions, regardless the dataset. However, two evident outliers were found, namely strains K27 and ChPhzS23 from the dataset *Pseudomonas*, both sequenced with PacBio technology [194]. Their corresponding accessions (see **Table 2.S3**) were both flagged on NCBI as *"removed from RefSeq due to many frameshift proteins"*, therefore these genomes should normally be excluded from an analysis. Furthermore, the 100% of the recovered reactions for these two strains were always present in the other strains of the dataset, demonstrating that the gene recovery procedure is effective in improving the metabolism reconstruction even for genomes with low sequence quality.

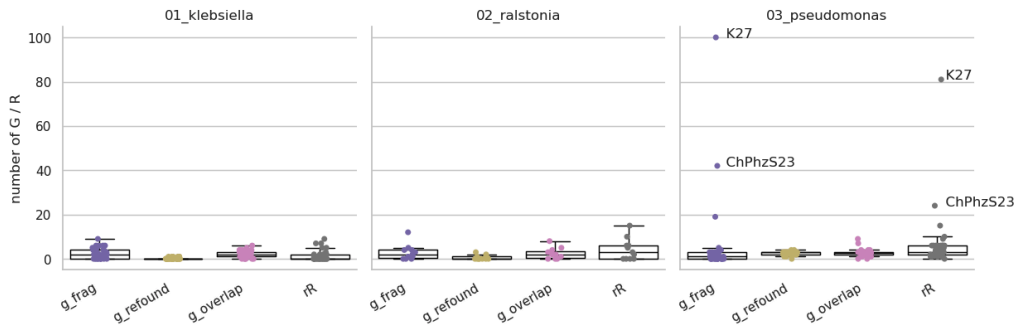

Figure 2.6. Effects of gene recovery. "g_frag", "g_refound" and "g_overlap" are the genes included in strain-specific GSMMs which have been recovered during the first, second, and third step of gene recovery, respectively. "rR" is the number of reactions included in the strain-specific GSMMs consequently to the entire gene recovery procedure.

Reactions with empty or wrong GPR

The accuracy of reconstructions could also be evaluated by the number of modeled metabolic reactions (not exchanges, sinks, nor demands) which have not been associated with genes and, at the same time, have not been labeled as spontaneous. These reactions without GPR, also known as "orphan" reactions [27], could be added by internal gapfillers of specific reconstruction tools to improve network connectivity [44] and should be manually checked, possibly linking them to genes.

The presence of orphan reactions was compared (**Figure 2.7**), and Gempipe reconstructions contained the lowest fraction of such reactions in all the three datasets. In Bactabolize reconstructions, orphans were copied for the reference, and fraction was even increased in *Ralstonia* and *Pseudommoas* datasets as a consequence of the gap-filled reactions. In Gempipe, orphans are also copied from the reference but, during the reference expansion process, their GPRs are supplemented with missing gene clusters taken from the reference-free reconstruction. This resulted in a fraction of orphans lower than the reference in all the three dataset, and even more low with respect to reference-free reconstruction tools. Respect to CarveMe, in particular, the difference seemed to be more accentuated when the reference is not (yet) part of the BiGG collection [56], as in the *Ralstonia* dataset.

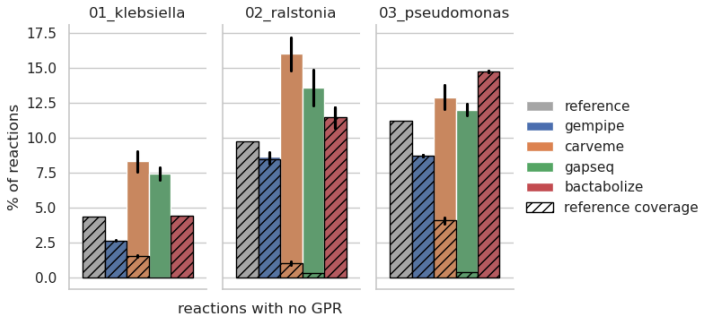

Figure 2.7. Relative number of modeled reactions with no GPR (orphans). Exchange, sink, and demand reactions are excluded, as well as reactions containing the substring “diffusion” in their name. Bar height corresponds to the mean between strains, while error bars represent standard deviations. Hatched area represents orphans in common with the reference.

Another aspect concerning accuracy, difficult to be evaluated systematically, is that of GPR correctness. To highlight this aspect, it was decided to report results for an exemplificative metabolic feature, the usage of mannitol, for the strain TAMOak81 of the *Pseudomonas* dataset (2.5.1 Supplementary Results). Briefly, TP matches with experimental substrate usage data were obtained by models built with CarveMe [44] or gapseq [65], even if the biology of the organism was not correctly represented in terms of reaction set and associated GPRs.

2.3 Methods

This work is referred to Gempipe v1.37.3. Pipeline implementation methods are detailed in 2.5.2 Supplementary Methods.

2.3.1 Generation of strain-specific GSMMs

Strain-specific GSMMs were reconstructed for three different genome datasets: "*Klebsiella*", "*Ralstonia*" and "*Pseudomonas*" datasets. Reconstructions were performed with four different pipelines: Gempipe ("autopilot" mode), CarveMe v1.6.2 [44], gapseq v1.3.1 [65], and Bactabolize v1.0.3 [129] (*draft_model* command), all installed via conda. As a backend linear programming solver, CPLEX v22.1.1 [195] was used.

For each dataset, a strain-specific reference GSMM was available (see below for details). The reference GSMM and associated coding sequences were specified using the options *-rm/-rp* in Gempipe, *--reference* in CarveMe, and *--ref_model_fp/--ref_proteins_fp/--ref_genes_fp* in Bactabolize. In gapseq, it is not possible to specify a reference GSMM, therefore reconstructions are fully based on the gapseq internal universe. Reconstructions are purely universe-based also in CarveMe, but the *--reference* option increases the reaction scores of those reactions in common with the reference, based on reaction ID exact matching. Moreover, the use of the reference GSMM as an alternative CarveMe universe already showed poor performances [129], so this reconstruction mode was ignored here. The gram-negative universe was specified with *-s* in Gempipe and *-u* in CarveMe, while gapseq relies on a staining autodetection feature.

The growth media used during gap-filling was specified using the options *--media* in Gempipe, *-g/--mediadb* in CarveMe, *-n* in gapseq, *--media_type/--atmosphere_type* in Bactabolize (after copying the medium definition in the installation directory). Unconstrained exchange of protons, water, oxygen, and CO_2 was allowed. In Bactabolize, oxygen and CO_2 were not included in the medium definition as the *--atmosphere_type aerobic* option was always used. Trace elements were approximated as an unconstrained uptake of Ca^{2+}, Cl^{-}, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, K^{+}, Mg^{2+}, Mn^{2+}, Zn^{2+} and MoO_4^{2-}. As Bactabolize does not include an automatic gap-filling feature, models generated with Bactabolize were automatically gap-filled with COBRApy v0.29.0 [81] using the specified growth medium and the reference GSMM as source of reactions, requiring a minimal growth rate of at least the 70% with respect to the reference.

In order to maximize the comparability among pipelines, given that gene calling is outside the scope of CarveMe, all the tools were run starting from the raw aminoacidic sequences generated by Gempipe, which in turn were derived from Prodigal v2.6.3 [151] run through Prokka v1.14.6 [196] with options *--noanno --norrna --notrna* (see 2.5.2 Supplementary Methods). The aminoacidic sequences were specified using the *-M 'prot'* option in gapseq, while they are the

default input type in CarveMe. In Bactabolize, multifasta proteomes are not allowed in input, therefore the aminoacidic sequences were formatted with Biopython v1.80 [197] as Bactabolize-compatible genbank files, which were later given in input specifying the *--no_reannotation* option.

In Gempipe, the genome filtering feature was not used, to prevent the exclusion of strains during the comparison between tools. Therefore, Gempipe parameters *--buscoM*, *--ncontigs* and *--N50* were set to 100%, 10000 and 0, respectively.

Klebsiella dataset

Genomes for 37 strains belonging to the *Klebsiella pneumonie* species complex were downloaded from NCBI [178]. The strain names and accessions are listed in **Table 2.S1**. More information on these strains such as the isolation source or the sequencing platform, are available at [134]. Binarized Biolog® data for 91 carbon sources in aerobiose were taken for each strain from the supplementary tables of [134] and assumed as ground truth.

The curated reference model iYL1228 [176], specific for *Klebsiella pneumonie* MGH 78578, was downloaded from BiGG [56] and slightly edited as follows. The biomass equation was edited by removing the strain-specific precursors dTDP-rhamnose ("dtdprmn_c"), UDP-galacturonate ("udpgalur_c") and UDP-galactose ("udpgal_c"), as indicated in [134]. Diffusion reactions involved in the transport of Biolog® substrates were systematically set as reversible [134].

With Biopython [197] a nucleotide and an aminoacidic fasta files, containing the gene sequences modeled in iYL1228, were created starting from the entire set of coding sequences retrieved from NCBI (CP000647.1). The resulting GSMM and associated coding sequences were considered as reference for the dataset, and were used in all the reconstruction tools where a reference can be specified.

Regarding the reconstruction process in Gempipe, the reference artificial gene indicating spontaneous reactions was specified as *-rs KPN_SPONT*, and the minimum flux through the objective was arbitrary set with *--minflux 30*; moreover, to avoid stoichiometric inconsistencies in the produced GSMMs, a blacklist of reactions not to be included during the expansion phase ("ACPpds", "ENTERH", and "NFORGLUAH2") was specified with *--mancor*.

For all reconstruction tools, the aerobic growth medium on which to guarantee strains' biomass production had glucose, ammonium, phosphate and sulfate as C, N, P, and S sources, respectively (unbounded uptake).

Ralstonia dataset

Genomes for 11 strains belonging to the *Ralstonia solanacearum* species complex were downloaded from NCBI [178]. The strain names and accessions are listed in **Table 2.S2**. More information on these strains such as the isolation source are available at [149]. Binarized Biolog® data for PM1 and PM2A plates in aerobiose were retrieved for each strain from the supplementary tables of [149] (time-normalized, maximal intensity) and assumed as ground truth.

The curated reference model iRP1476 [198], specific for *Ralstonia solanacearum* GMI1000, was downloaded in a COBRApy-compatible version from supplementary materials of [44], and slightly edited as follows. All unconserved metabolites were detected with MEMOTE v0.17.0 [70] and manually corrected. The gene RSp042 was changed to RSp0421 in the "ARBTNLSYN" reaction GPR. To uniform the reconstruction paradigm among models, prefix "G_" was removed from gene ids; genes whose ID started with "e" (indicating an exchange reaction) were removed; genes whose ID started with "d" or "s" (indicating diffusions and spontaneous reactions, respectively) were all converted to an artificial "spontaneous" gene; the gene "NoAssignment" was removed from GPRs. As indicated in [149], cobolamin ("adocbl_c") and sperimidine ("spmd_c") are non-essential for growth, and were thus removed from the biomass assembly. Finally, an uptake for Ca^{2+} was introduced as a proton antiport.

With Biopython [197] a nucleotide and an aminoacidic fasta file, containing the gene sequences modeled in iRP1476, were created starting from the entire set of coding sequences retrieved from NCBI (NC_003295.1, NC_003296.1, AL646052.1, AL646053.1). Genes not recovered were removed from the model without removing reactions. The resulting GSMM and associated coding sequences were considered as reference for the dataset, and were used in all the reconstruction tools where a reference can be specified.

Regarding the reconstruction process in Gempipe, to avoid stoichiometric inconsistencies in the produced GSMMs, a blacklist composed by the only reaction "ACPpds" was specified using *--mancor*.

For all reconstruction tools, the aerobic growth medium on which to guarantee strains' biomass production had L-glutamine, ammonium, phosphate and sulfate as C, N, P, and S sources, respectively (unbounded uptake).

Pseudomonas dataset

Genomes for 36 strains of *Pseudomonas chlororaphis* were downloaded from NCBI [178]. The strain names and accessions are listed in **Table 2.S3**. More information on these strains such as the isolation source are available at [191]. Binarized Biolog® carbon source utilization assays (PM1 and PM2A) in aerobiose were retrieved for each strain from supplementary material of [191] and assumed as ground truth.

The curated reference model iJN1463 [167], specific for *Pseudomonas putida* KT2440, was downloaded from BiGG [28] and slightly edited as follows. Its unconstrained bound constants were changed from +/- 999999 to +/- 1000, and the objective reaction was set to the preloaded "core" biomass equation described in [167].

With Biopython [197] a nucleotide and an aminoacidic fasta file, containing the gene sequences modeled in iJN1463, were created starting from the entire set of coding sequences retrieved from NCBI (NC_002947.4). Genes not recovered were removed from the model without removing reactions. The resulting GSMM and associated coding sequences were considered as reference for the dataset, and were used in all the reconstruction tools where a reference can be specified.

Regarding the reconstruction process in Gempipe, the reference artificial gene indicating spontaneous reactions was specified as *-rs PP_s0001*, and the minimum flux through the objective was arbitrary set with *--minflux 15*; moreover, to avoid stoichiometric inconsistencies in the produced GSMMs,, a blacklist of reactions not to be included during the expansion phase ("AALDH", "ACPpds", "ASR2" and "PROR") was specified with *--mancor*.

For all reconstruction tools, the aerobic growth medium on which to guarantee strains' biomass production had glucose, ammonium, phosphate and sulfate as C, N, P, and S sources, respectively (unbounded uptake). For Gempipe and Bactabolize reconstructions, with respect to other datasets, the unconstrained uptake of Na^{+} and Ni^{2+} was added to satisfy the growth requirements of iJN1463.

2.3.2 Comparison between tools

Content of reconstructed GSMMs

Each strain-specific GSMM was loaded into COBRApy [81], and four sets of IDs were extracted: IDs for genes ("G"), reactions ("R"), metabolites ("uM") and exchange reactions ("exr"). Reaction IDs fell into the "exr" set if the corresponding reaction had just one involved metabolite belonging to the external compartment, otherwise it fell into the "R" set. Metabolite IDs were extracted without considering their compartment. Moreover, the number of reactions without GPR (orphans) was computed ignoring reactions with either just one metabolite involved (exchanges, sinks, and demands) or with "diffusion" appearing as substring in their name (reactions marked as spontaneous were not considered). For GSMMs created by gapseq, the IDs of metabolites and reactions were translated from SEED to BiGG namespace using the MetaNetX v4.4 [186] mappings.

Each GSMMs was then subjected to the computation of the MEMOTE metrics [70] which are five (consistency, metabolite annotations, reaction annotations, gene annotations, SBO annotations), plus a total score. The MEMOTE command line tool was not used in favor of the MEMOTE API for Python, precisely *memote.suite.api.test_model*, which was implemented in a parallelized process to speed up the computation time. From the partial results obtained with the API, the original MEMOTE metrics were replicated applying the formulas described in the original paper [70].

To assess the reference coverage, i.e. the degree of superimposition between the output models and the reference, the intersection between each strain-specific GSMM and its reference was computed in terms of ID sets. This intersection was directly possible for metabolites and reactions, even for gapseq reconstructions thanks to the namespace translation provided by MetaNetX [186]. Instead, the intersection was not directly possible for genes, as the IDs are strain-specific. To connect the gene IDs in the output GSMMs to the gene IDs in the reference, the following approach was used. First, each reference gene ID was converted to the corresponding set of cluster IDs computed during the BRH step of Gempipe; second, for each cluster ID, all the correspondent strain-specific gene IDs were collected by reading the PAM (see 2.5.2 Supplementary Methods). A reference gene was considered as "covered" if at least one of these equivalent strain-specific genes was present in the strain-specific GSMM.

To compare the similarity among the reconstruction tools in terms of reactions, a correlation matrix based was computed. Each cell of the matrix reported the similarity among two tools, represented by the mean Jaccard index

$$\sum_{i=1}^{n}\left(\frac{R_i^A \cap R_i^B}{R_i^A \cup R_i^B}\right)\frac{1}{n}$$

where R_i^A and R_i^B are the BiGG-based set of reaction IDs of the tool *A* and *B*, respectively, and *n* is the number of strains in the dataset; when a tool is compared against a reference GSMM, then each strain-specific GSMM is compared against a copy of the reference.

To identify the reactions introduced by Gempipe as a consequence of the gene recovery, recovered genes were subtracted. Specifically, for each GSMM in output, all the modeled genes having the *'_frag'*, *'_refound'*, or *'_overlap'* suffixes in their ID were removed, eventually leading to a reaction loss. Removed reactions were those gained by the gene recovery.

Simulations using reconstructed GSMMs

For all dataset and pipelines, Biolog® substrate utilization was simulated using the *biolog_preview* function from the Gempipe API, indicating glucose, ammonium, phosphate and sulfate as starting C, N, P, and S sources, respectively (see 2.5.2 Supplementary Methods). The argument *seed=True* was utilized for gapseq reconstructions. If the exchange reaction for a particular substrate was missing from a GSMM, this was interpreted as inability to catabolize that substrate.

To compare substrate utilization predictions among tools, the following metrics were considered [44,129]:

$$\text{precision} = \frac{TP}{TP+FP}$$
$$\text{recall} = \frac{TP}{TP+FN}$$
$$\text{specificity} = \frac{TN}{TN+FP}$$
$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP is the number of true positive substrates (predicted growth on experimental growth), TN is the number of true negative substrates (predicated non-growth on experimental non-growth), FP is the number of false positive

substrates (predicted growth on experimental non-growth), and FN is the number of false negative substrates (predicted non-growth on experimental growth). To penalize reconstructions with severe gaps, each metric was multiplied by the fraction of "optimal" (i.e., non-"infeasible") FBAs performed, where each FBA is associated to a specific substrate.

When evaluating the ability to recapitulate Biolog® data, substrates were excluded from the comparison if they were FN in at least one Gempipe strain-specific GSMM and, at the same time, if they could not sustain the growth of the draft pan-GSMM. This decision was taken because such mismatches clearly indicate a lack of manual curation in the draft pan-GSMM (in this case, reactions that should be introduced), and the comparison was aimed to be focused solely on the level of automation.

2.4 Discussion

In the present work Gempipe was introduced, a package for pan and multi-strain reconstructions of GSMMs. It can be easily installed via conda (*conda install gempipe -c bioconda*), its source code is publicly hosted on GitHub (https://github.com/lazzarigioele/gempipe) and its documentation is available on ReadTheDocs (https://gempipe.readthedocs.io/en/latest/).

Gempipe adopts an hybrid reconstruction approach that lies between reference-free and reference-based methods, as an optional reference is expanded with new reactions coming from a universal GSMM. Together with an internal clustering of strain-specific genes, extensive gene recovery, and conservative generation of GPRs, the implemented approach proved to be effective for multi-strain reconstructions, with better or similar performances compared to current established reconstruction tools when focusing on metabolic features and manual curation is not considered.

Gempipe represents a third option in the panorama of reconstruction tools: reference-based methods use a manually curated model (or sorted list of models, from the phylogenetically closest to the most curated) to be used as template, from which reactions are copied after orthologs genes are determined [148–150]; instead, reference-free methods use a semi-curated universal model with a generic biomass equation, from which reactions are copied and gap-filled based on sequence homology (one-way alignment) [44,65]. From the first, Gempipe inherits the ortholog determination to translate reference genes into equivalent gene clusters; from the second, Gempipe inherits the homology-based insertion of new, strain-specific reactions, with GPRs already based on gene clusters.

In the case of multi-strain reconstruction, reference-based methods are usually applied [37,114–116,134,152,167,168]. The main concern, in this context, is that the template must be representative of the metabolic diversity of the entire species or genera [129], otherwise strain-specific reactions must be added afterwards to each generated model [48]. This is why a comprehensive (pan) GSMM is usually curated prior to running reference-based reconstruction tools [175]. However, such pan-GSMM, representing a species or genera, are complex to build and still rare in literature [115,168,175,177], with as little as 4 species reported [129]. Indeed, in some GSMM-based biodiversity studies, strain-specific models have been used as reference [37,134,152]. This is a limiting approach, as generated models are just a subset of a single strain. In this context, Gempipe was able to grasp strain-specificities better than purely reference based methods like Bactabolize [129] when a single-strain GSMM is used as reference instead of a pan-GSMM, which is a common case. Moreover, inheriting contents from a manually curated reference, models built with Gempipe provided more accurate predictions with respect to those built with reference-free methods like CarveMe [44] and gapseq [65]. All this was achieved while maintaining high MEMOTE metrics [70], thanks to the internal *de novo* annotation phase.

Considering their paucity, Gempipe emerges also as a valuable tool to quickly build and curate pan-GSMMs to be used in biodiversity or bioprospecting studies. It must be noted, however, that the concept of pan-GSMM was introduced also in the context of metagenomics-derived models [169], which is remarkably different. In this context, indeed, pan-GSMMs are instrumental to cope with the incompleteness and contamination of metagenome-assembled genomes (MAGs), and strain-specificity in terms of genes and reactions is lost in favor of a consensus/mean reconstruction representing a species-level genome bin (SGB) [169]. Instead, in the GSMM-guided exploration of biodiversity, the context where Gempipe operates, strain-specificity must be retained and emphasized as it is needed for the subsequent generation of strain-specific GSMMs [48].

In the development of Gempipe, particular attention was put into how metabolic features are encoded in the model. In this regard, it was shown that metrics commonly used to evaluate and compare reconstructions (accuracy, precision, recall, specificity) [44,65,129] do not take into consideration the faithfulness of the reaction network in representing the organism. Indeed, behind a TP match with the experimental data (e.g., Biolog® screenings) there could be cases of (i) wrong reaction mechanisms (e.g., erroneous transporter types); (ii) presence of metabolic reactions not supported by genes (orphans);

(iii) reactions with seriously impaired GPRs, lacking many components of a protein complex. Provided that manual curation still remains essential for obtaining faithful GSMMs, users should be aware of the reconstruction principles that each tool follow to draft a model. Gempipe includes reactions only when a protein complex is fully supported by genes, otherwise mismatches (FN) will lead the manual curation in closing the gaps. On the other side, tools providing (a) internal gap-filling procedures aimed to improve the network connectivity and (b) too permissive mechanisms of GPR generation and reaction inclusion, could lead not only to predict a higher number of growth-supporting substrates (higher recall and lower specificity) but, most importantly, to match the experimental data with a biologically inaccurate representation of the metabolism, which can be easily overlooked during the manual curation.

Finally, Gempipe is not meant to be just a reconstruction tool, but also an analysis tool in the context of biodiversity exploration. Indeed, while the Gempipe API includes functions to curate models, it also contains functions to analyze the deck of strain-specific GSMMs in output. For example, functions are included to cluster strains according to their metabolism, and to compare metabolic clusters with respect to other attributes, for example niche metadata or the formal species classification. This aids users to achieve goals including (i) the screening of strains for desired metabolic traits, (ii) the classification of strains according to their metabolic capabilities, and (iii) the definition of species in terms of their core metabolic potential.

Limitations of Gempipe are mostly due to the resources it relies on. The BiGG namespace [28] was adopted as it was convenient for two main reasons: (i) the availability of high-quality, manually curated reference models based on the same namespace [56]; (ii) the human readable IDs, particularly useful during manual curation, when GSMM-based metabolic maps have to be hand-drawn or interpreted [35,36] (for example, D-glucose is "glc__D", immediately recognizable with respect to its ModelSEED [29] equivalent "cpd00027"). While convenient, the BiGG database has imperfections due to its structure: it is not a coherent biochemical database, but rather a collection of models from which content a biochemical database is derived. Depending on the model of origin, the same metabolite or reaction can be defined differently. Consequences are many: (i) the same reaction can be represented with different reversibility (e.g. "ILETA2"); (ii) the same metabolite can be represented with different IDs, leading to duplicate metabolites (e.g. "ind_c" / "indole_c"); (iii) the same reaction can use duplicate metabolites, leading to duplicate reactions (e.g. "TRPS2" / "TRPS2_1"); (iv) metabolites with the same ID can be represented with different chemical formula or charge (e.g. "fmn_c").

Therefore, the integration of a BiGG-based model with reactions coming from another BiGG-based model can lead to the introduction of unbalanced reactions or, even worse, stoichiometric inconsistencies [41]. Gempipe tries to circumvent this issue by enabling users to superimpose particular metabolite charges and formulas or reaction balances during the reference expansion phase.

Another limitation depending on BiGG [56] is that of the bacterial diversity coverage, mainly in terms of genes. Indeed, in its current version (v1.6), BiGG contains as little as 108 GSMMs, of which 88 are prokaryotic, of which only 22 do not belong to the *Escherichia* or *Shigella* genera. The set of BiGG genes, used by Gempipe and CarveMe [44], is therefore biased on model species and does not cover much bacterial diversity, potentially leading to missed genes (and thus reactions) during the reference-free reconstruction phase.
It is anticipated that further development on Gempipe will focus on the introduction of new universes, such as one for yeasts [199], and on new API functions for the analysis of the deck of strain-specific GSMMs created.

In conclusion, Gempipe will facilitate metabolic biodiversity studies for a wide range of bacterial species, including those not having a dedicated pan-GSMM, which are currently the large majority.

2.5 Supplementary Material

2.5.1 Supplementary Results

GPR correctness

The correctness of GPRs generated by reference-free reconstruction tools can be evaluated when a manually curated reference GSMM is available for the same strain being reconstructed [65]. In case of multi-strain reconstructions, as reference models are not available for each strain, the evaluation of GPRs correctness is more difficult to be evaluated systematically. To highlight the issue, it was decided to report results for an exemplificative metabolic feature: the usage of mannitol in *Pseudomonas chlororaphis* TAMOak81 from the *Pseudomonas* dataset.

According to the Biolog® PM screening, TAMOak81 and all the other 35 strains of the dataset are able to catabolize mannitol. Anyway, none of the respective strain-specific GSMMs generated with Gempipe predicted growth on mannitol; among the CarveMe [44] reconstructions, 20 out of 36 were able to grow on

mannitol, including TAMOak81; instead, gapseq [65] reconstruction were all concordant with the observed phenotype, meaning that a TP match was reported for all the 36 strains.

The usage of mannitol is encoded in the CarveMe gram-negative universe – in common with Gempipe – with 4 reactions (**Figure 2.S2A**): MNLtex, MNLptspp and MNLpts, representing a phosphotransferase system (PTS)-type transporter; M1PD, converting mannitol-1-phosphate to fructose-6-phosphate, providing the only connection to the central carbon metabolism.

The GSMM created by CarveMe for TAMOak81 correctly predicted growth on mannitol, but two issues were noted: (i) the M1PD reaction resulted without GPR (orphan), so the catabolic pathway was not fully supported by genes; (ii) the PTS-type transporter (MNLpts) had a 1-component, incomplete GPR ("DNLIPFME_05586"), while the manually curated GSMMs stored in BiGG [56] encode the same reaction with a 2-to-5-component GPR (as indicated in the file "bigg_gprs.csv", see 2.5.2 Supplementary Methods). By design, Gempipe introduces reactions only if at least one of the original protein complex definitions is fully satisfied, therefore neither M1PD nor MNLpts were introduced in the TAMOak81 reconstruction.

Besides these two technical issues, an additional biological issue summed up: (iii) it is known from literature that mannitol is imported in *Pseudomonas* with a 4-components ATP-binding cassette (ABC)-type transporter MtlEFGK [200,201], not a PTS-type transporter. Therefore, not only the BiGG-derived CarveMe gram negative universe [44] is incomplete, but the resulting GSMM was also biologically inaccurate in representing this metabolic feature.

For comparison, "gempipe autopilot" was run again, exactly as described in 2.5.2 Supplementary Methods, but with the addition of the *--tcdb* option, therefore trying to expand the transport reactions using the information derived from TCDB [67]. The resulting GSMM for TAMOak81 correctly included an additional reaction encoding the ABC-type mannitol transporter, supported by an accurate 4-component GPR, with AND-linked subunits.

As stated before, the GSMM created by gapseq for TAMOak81 (**Figure 2.S2B**) correctly predicted growth on mannitol. Anyway, two issues were noted: (i) the ABC-type transporter was present (rxn13853), but other two transport systems were included as well, comprising a symporter and a PTS-type transporter, the latter universally blocked; (ii) the GPR of rxn13853 is not compatible with a true ABC-type transporter for mannitol, being composed by a single gene ("DNLIPFME_02905").

None of the strain-specific GSMMs produced by Bactabolize [129] was able to grow on mannitol. Indeed, the purely reference-based reconstruction method implemented in Bactabolize imposes that models in output can only be a subset of the reference, which, in this case, do not encode a catabolic pathway for mannitol.

2.5.2 Supplementary Methods

Gempipe is a Linux AMD64 pipeline written in Python[1] v3.9+ and composed of three command-line programs, "gempipe recon", "gempipe derive", and "gempipe autopilot", plus a Python API intended to be used from Jupyter Notebooks [190]. It has the following list of dependencies: Biopython v1.80+ [197], Prodigal v2.6.3+ [151], Prokka v1.14.6+ [196], BUSCO v5.4.0+ [179], SeqKit v2.2.0+ [202], cd-hit v4.8.1+ [203], blast v2.12.0+ [50], eggNOG-mapper v2.1.7+ [182], DIAMOND v2.0.15+ [158], ncbi-genome-download[2] v0.3.3+, pigz[3] v2.5+, scipy[4] v1.10.0+, scikit-learn[5] v1.3.0+. Tabular files are handled with pandas[6] v2.0.0+ and plots are generated with matplotlib[7] v3.7.0+ and seaborn[8] v0.13.0+. GSMMs are handled with COBRApy v0.29+ [81], using CPLEX v22.1.1 [195] as backend linear programming solver. Many steps of the reconstruction are parallelized using the "multiprocess" Python standard library. Below is described the implementation of each part of Gempipe.

From genomes to protein clusters

"gempipe recon" reconstructs draft pan-GSMMs and supports four different and mutually exclusive types of inputs, in order of priority: proteomes in genbank format, proteomes in fasta format, raw local genome assemblies, and NCBI Species Taxonomy IDs (taxids) to download genomes on-the-fly. Each genome/proteome is treated as a separate strain.

When proteomes in genbank format are given in input, they are parsed with Biopython to extract all the coding sequences and their annotations. The formatting of the first sequence ID is checked to determine whether the genbank file derives from the GenBank or RefSeq database. According to the database of origin, a single query is built for each genbank file using the UniProt

[1] https://www.python.org
[2] https://github.com/kblin/ncbi-genome-download
[3] https://github.com/madler/pigz
[4] https://github.com/scipy/scipy
[5] https://github.com/scikit-learn/scikit-learn
[6] https://github.com/pandas-dev/pandas
[7] https://github.com/matplotlib/matplotlib
[8] https://github.com/mwaskom/seaborn

REST API [166] to retrieve the attributes “accession”, “gene_oln”, “xref_refseq”, “xref_embl”, “xref_kegg”, and “xref_geneid”, for all the coding sequences. These attributes are supplemented with the qualifiers “gene”, “locus_tag”, “old_locus_tag”, and “protein_id”, obtained parsing the genbank file itself. Taken together, these metadata are formatted to obtain five MIRIAM-compliant [140] gene annotations: “refseq”, “ncbiprotein”, “ncbigene”, “kegg”, “uniprot”. Such annotations are stored in a table that can be provided to “gempipe derive” to improve the MEMOTE [70] gene metrics of strain-specific GSMMs. With Biopython [197], a fasta proteome file is finally created extracting all the coding sequences in aminoacidic format from the genbank file. IDs in the fasta file are given after the “old_locus_tag” qualifier, if available, otherwise the “locus_tag” qualifier is used. With this input type, gene annotation is assumed as accurate or manually reviewed, so no gene recovery will be performed later.

When proteomes in fasta format are given in input, they are directly used in subsequent analysis. With this type of input, again, gene annotation is assumed as definitive and gene recovery will be skipped.

When raw local genome assemblies are given in input, they are first subjected to gene annotation running Prodigal [151] via Prokka [196] with options *--noanno --norrna --notrna*, this way obtaining a more convenient gene naming system. With this gene annotation, each assembly is associated with a proteome. Next, BUSCO [204] is run with options *--mode proteins*, indicating the user-provided database with *--lineage_dataset*. Users have to indicate just the name of the desired database, then it is automatically downloaded. If no database has been indicated, the generic “bacteria_odb10” is used. Subsequently, *seqkit stats* [202] is run with options *--tabular --basename --all*, in order to compute the number of contigs and the N50 for each assembly. At this point, all the assemblies that do not respect at least one of the following user-specified thresholds are discarded from subsequent analysis, together with their associated proteomes: maximum number of missing (default 2%) or fragmented (default 100%) BUSCO orthologs, minimum N50 (default 50000 [205]), and maximum number of contigs (default 200 [205]).

When taxids are given in input, the ncbi-genome-download tool is run with options *--no-cache --retries 100 --parallel 10 --section genbank --formats assembly-stats,fasta*, providing the taxids list with *--species-taxids* and asking the metadata table with *--metadata-table*. All the available genome assemblies for the indicated species are automatically downloaded from Genbank and then decompressed with unpigz. With the obtained genomes, Gempipe reprises from the genome filtering like if users started from raw genome assemblies.

Once that, regardless of the selected type of input, a proteome is obtained for each strain, all the coding sequences are combined together into a single aminoacidic multifasta file. This is subsequently inputted to CD-HIT [203] to group the sequences into clusters based on sequence homology. CD-HIT is run with parameters *-M -g 1 -aL 0.70 -aS 0.70 -c 0.90 -d 0*, meaning that sequences are grouped using high global sequence identity (90%). This command creates two main output files: (i) a clustering file (.clstr) indicating, for each cluster, the IDs of the representative and member sequences; (ii) a multifasta file (.faa), containing the aminoacidic sequences of the representatives. With Biopython [197], this latter file is edited to replace the original sequence header (strain-specific ID generated by Prokka [196]) with the cluster ID generated by CD-HIT (formatted as "Cluster_*"). Using the clustering file, an initial gene presence / absence matrix (PAM) is created, that is a plain text table (.csv) having the cluster IDs in rows, and the strains in columns. Each cell of the PAM contains the IDs of all the genes, separated by a semicolon, that are present in a specific strain and are members of a specific cluster.

The gene recovery steps

If Gempipe starts from genomes (either user-provided or automatically downloaded), then a three-steps gene recovery is applied. The first step recovers proteins broken into pieces due to eventual issues in genome sequencing or assembling. In this step, for each strain, the proteome is aligned with blastp [50] on the representative sequences that originated from the clustering. The alignment is filtered for high-scoring segment pairs (HSPs) having identity at least 90%, query coverage at least 70%, and e-value at most 1e-5. HSPs are then sorted by ascending e-value and grouped by cluster ID. Given a cluster-specific group, if two or more HSPs are associated with the same query ID, only the first HSP (best scoring) is retained. This results in a list of strain-specific coding sequences (queries) that are aligned with high identity to a particular cluster (subject): some of these sequences could potentially be just fragments. Given that Prokka [196] names the annotated genes using a progressive number according to the gene order, couples of sequences named with "n" and "n+1" are searched among the queries of a cluster-specific group. At this point, for each couple, it is computed: (A) the overall coverage, which include the eventual gap between the two sequences; (B) the superimposition between the two sequences, accounting for the eventual overlap; (C) the length of each couple member; (D) the relative frequency of the starting cluster of each couple member, indicating the number of strains in the PAM having sequences belonging to that cluster; (E) the relative frequency of new the cluster for which the couple members have been grouped in this gene recovery step. (A), (B), and (C) are all expressed in % relatively to the length of the subject (representative

sequence). For each detected couple, if the overall coverage is at least 70%, the superimposition is at most 30%, the relative length of the queries is below 90%, and the relative frequency of the new cluster is higher than that of both the starting clusters, then the couple members – taken together – are assumed as a recovered gene. The ID of the recovered gene will be "{prefix}_frag_{n}_{n+1}", where {prefix} is the constant genome-specific prefix generated by Prokka, and {n} and {n+1} are the progressive numbers associated with the couple members during the gene calling. Finally, the PAM is updated removing the couple members and adding the recovered gene ID to the corresponding cluster. After this procedure, eventual rows that remained empty (i.e., clusters remained without sequences) are removed from the PAM.

The second gene recovery step deals with unperfect gene calling, in particular with genomic regions that may be overlooked by the caller. In this step, all the coding sequences included in the updated PAM are masked from the original genome, this way accounting for the previous gene recovery. For each strain,a multifasta file (query) is created containing the aminoacidic sequences of the representatives of all the clusters that have no strain-specific members yet. These queries are then aligned with tblastn [50] over the masked genome and the resulting HSPs are grouped by cluster. For each cluster, HSPs are filtered for those having identity at least 90% and query coverage at least 70%. Each recovered sequence is named as "{prefix}_refound_{x}", where {prefix} is the prokka-specific prefix and {x} is a discriminative integer number. Then, for each recovered sequence, premature stop codons are searched. Briefly, the recovered sequence is extracted with blastdbcmd [50] and converted using Biopython [197]. If the length of the sequence translated until the first stop codon is below 95% of the translation until the end, then the ID of the recovered gene is marked with the suffix "_stop", indicating the presence of a premature stop codon. Finally, the PAM is updated introducing the new genes.

The third and last gene recovery step is again focused on issues in gene calling, in particular it deals with overlapping open reading frames. As in the second step, a query file is created for each strain, composed by the representative sequences of the clusters that are still without members from the strain. The queries are then aligned with tblastn [50] over the original (non-masked) genome assembly, and the resulting HSPs are sorted by ascending e-value and filtered for identity at least 90%, query coverage at least 70%, and e-value at most 1e-5. Next, the raw alignment coming from the previous step (alignment on masked genomes) is reloaded, its relative HSPs are sorted by ascending e-value and filtered to retain the queries targeted in the last alignment. Then, these HSPs are filtered for identity at least 90% and query coverage below 70%, this way isolating those high-identity segments that were discarded in the

previous recovery step due to an unsatisfying query coverage. These segments are grouped by cluster and, for each group, the first (best scoring) HSP is kept, annotating the contig ID and starting query coverage. Then, among the HSPs coming from the last alignment, the first (best scoring) HSP associated with the same contig ID but having a greater query coverage is searched. With this strategy, genes that (i) start in a genomic region overlooked by the caller, and (ii) end inside another previously annotated gene, are recovered. These recovered sequences are introduced in the PAM as "{prefix}_overlap_{x}", where {prefix} is the Prokka-specific prefix and {x} is a discriminative integer number. As in the previous recovery step, for each recovery sequence, the presence of premature stop codons is evaluated and the prefix "_stop" is eventually added to the sequence ID.

Reference-free reactions generation

Next, the representative sequences are processed to create a repository of reactions, used either to expand a reference GSMM with new strain-specific reactions or to constitute a reference-free reconstruction when no reference GSMM is provided. This repository could be seen as a pan-reactome, as it represents all the metabolic reactions theoretically present in at least one of the input strains. To accomplish this task, three files are copied from the CarveMe [44] assets, namely "bigg_proteins.faa", "bigg_gprs.csv" and the bacterial universal GSMM. The first is a multifasta file containing all the BiGG-derived [56] gene sequences that constitute the CarveMe gene database (hereafter referred to as the BiGG genes). The second is a table listing, for each BiGG gene, all its GSMM-specific protein complexes in which it is involved, together with the relative reaction ID. The third is a semi-curated, BiGG-based, COBRApy-compatible [81] universal GSMM, either for gram positive or negative bacteria, according to the staining specified by users.

The sequences of the BiGG genes are used to create a DIAMOND [158] database, on which the representative sequences are aligned with *diamond blastp* using parameters *--ultra-sensitive --top 10*. This alignment is translated to a set of reactions to be copied from the CarveMe [44] universe, this way composing the reference-free reconstruction. The translation procedure follows the CarveMe algorithm, which converts the bit score of the HSPs into reaction scores using the information contained in "bigg_gprs.csv". Anyway, it includes some important modifications to better manage the multi-strain reconstructions. First, the obtained HSPs are filtered using the user-specified parameters percentage identity (default 30%) and percentage query and subject coverage (default 70%), enabling users to discard low-confidence reactions based on sequence homology.

Then each BiGG gene (subject) is associated with a score and a partial GPR that is called "isoform GPR". This partial GPR shows all the CD-HIT [203] clusters passing the identity and coverage thresholds for the relative BiGG gene, linking these "isoforms" with the boolean operator "OR". For example, if "Cluster_7296", "Cluster_7339" and "Cluster_7891" are all included in quality-filtered HSPs having the BiGG gene PP_0225 [167] as subject, then the isoform GPR for PP_0225 will be "(Cluster_7296 or Cluster_7339 or Cluster_7891)". This is a fundamental difference compared to CarveMe [44], which by design does not consider protein isoforms, but retains just the protein (or cluster, in the case of Gempipe) with the highest bit score, ignoring all the others (see the GitHub public issue #180[9]). After this step, each BiGG gene is linked to 0 or more CD-HIT clusters, with an empty isoform GPR in case of 0. To associate a score to each BiGG gene, the maximum bit score among the isoforms is used. Like in CarveMe, the BiGG genes describing spontaneous reactions (artificial genes) are associated with a score of 0.

Next, each protein complex in "bigg_gprs.csv" is associated with a dedicated score and a higher-level partial GPR ("protein complex GPR"). In essence, the isoform GPRs of the BiGG genes involved in the complex are joined together using the boolean operator "AND". The protein complex score is computed as the mean of the gene scores, using 0 for those BiGG genes associated with no clusters. Importantly, the protein complex GPR is produced only if all the involved BiGG genes have been associated to an isoform GPR (i.e., at least 1 CD-HIT cluster), this way satisfying the protein complex definitions as they were encoded in the original GSMMs deposited in BiGG [56]. In contrast, CarveMe [44] produces a protein complex GPR with as little as 1 aligned gene, even if the complex was originally defined to be composed by 2 or more members (see the GitHub public issue #182[10]).

Lastly, protein complex GPRs and scores are converted into reaction GPRs and scores. As in CarveMe [44], every protein complex GPR linked to the same reaction is joined together using the boolean operator "OR". Doing so, when 2 or more protein complexes – coming from the same or different GSMMs deposited in BiGG – are linked to the same reaction ID, then their partial GPR becomes a distinct alternative in the GPR for that reaction. By design, when a reaction has no alternative protein complex fully satisfied, then that reaction is not included in the reference-free reconstruction. Finally, the reaction score is computed as the maximum score among the alternative protein complexes, and it is normalized by the median computed among reactions having scores > 0.

[9] https://github.com/cdanielmachado/carveme/issues/180
[10] https://github.com/cdanielmachado/carveme/issues/182

At this point, the reference-free reconstruction is created by making a copy of the selected CarveMe universe [44] and then subtracting it from all the previously identified reactions with empty GPR, except for the biomass equation ("Growth"). Then, each retained reaction is updated with the GPR previously built, and an exchange reaction with bounds (0, 1000) is created for each metabolite associated with the external compartment ("_e").

As reported in 2.2.1 Pipeline and 2.4 Discussion, genes included in the BiGG database (and thus in "bigg_proteins.faa") are scarce and biased towards model organisms. To cope with this issue, additional genes are recovered exploiting a detailed functional annotation. The representative sequences of the clusters are annotated by eggNOG-mapper [182] using options *-m diamond --itype proteins --trans_table 11*, automatically downloading its database if not already available. Once loaded the annotation table, for each cluster 5 discriminative attributes ("Preferred_name", "KEGG_ko", "KEGG_Reaction", "EC", "KEGG_TC") are picked up and concatenated to form a single string or "dense" annotation. Then, annotated clusters are divided in groups based on their dense annotation, ignoring those with no annotation in any of the 5 attributes (corresponding to the dense annotation "-----"). Next, for each annotated cluster in the reference-free reconstruction, the GPR of all of its linked reactions is updated, replacing that cluster with an "OR"-block enclosed by parenthesis, where all the members of the same group are connected by the boolean operator "OR".

Handling of the reference

Next, the pipeline continues with the handling of the optional reference GSMM, eventually provided by users together with its associated proteome. The first step is to translate the reference gene IDs into cluster IDs, making the reference GSMM compatible with the PAM and the reference-free reconstruction. For each strain, a blastp [50] best-reciprocal hits (BRHs) alignment is performed between the reference proteome and the strain-specific proteome, each time filtering both the sets of HSPs with query coverage at least 70% and e-value at most 1e-5, then sorting the remaining HSPs by ascending e-value. Alignments are parsed to obtain, for each strain, a set of 1-to-1 and a set of 1-to-0 associations, the latter representing best hits without a reciprocal. Next, for each reference gene, all the strain-specific 1-to-1 associations are collected, and the list of strain-specific orthologs is extracted. For each ortholog, the correspondent CD-HIT cluster [203] is retrieved by reading the PAM, obtaining a list of clusters for each reference gene. Then, a copy of the reference GSMM is made and, in each of its GPRs, every gene is replaced with an "OR" block enclosed by parenthesis, containing all the equivalent cluster IDs connected by

the boolean operator “OR”. Remaining without associated reactions, original genes are removed obtaining the reference GSMM fully translated.

The translated reference model is then supplemented with the reactions contained in the reference-free draft pan-GSMM, used as a reaction repository. These reactions are not imported blindly, but prioritizing the reference as much as possible, in terms of reactions mass and charge balance. Briefly, metabolic reactions contained in the repository are iterated, with exclusion of the biomass assembly reaction (“Growth”) and all the spontaneous reactions. For each repository reaction, an exact string match of reaction IDs is performed to check if it is already included in the reference. If the reaction is not found in the reference, it is likely that it is actually present but with a different ID. In this case, the first available “synonym” reaction is searched, defined as a reaction having the exact same reactants and metabolites (in terms of IDs), protons excluded, regardless of its reversibility or its upper and lower bounds. If an equivalent reference reaction is found (having the same ID or being a synonym), a check is made to verify whether it is linked to the same set of gene clusters. If it differs, then the reference GPR is updated by concatenating the GPR coming from the reference-free reconstruction, using the boolean operator “OR”.

If an equivalent reaction is not found, then the repository reaction is assumed as new to reference and it is imported together with its GPR and bounds. For each repository reaction to import, it is first checked which involved metabolites are still missing from the reference, adding them when needed. If the reference provided is purely stoichiometric, meaning that none of its metabolites is annotated with a chemical formula or a charge, then new metabolites created during the reaction transfer will be consistent with this design decision. Otherwise, by default, new metabolites inherit the same chemical formula and charge defined in the reference-free reconstruction. However, the same metabolite (same ID) could be present in different BiGG-based models with different chemical formula and / or charge, and this can lead to the import of unbalanced reactions or, in the worst case, to the violation of the stoichiometric consistency [41]. To avoid such issues, users can provide an optional text file in input, containing a set of rules to be applied during the import of new reactions from the reference-free reconstruction.

Rules are described using a simple syntax, with one rule per line. The chemical formula of a metabolite can be forced using the syntax “formula.{mid}:{formula}”, where {mid} is the metabolite ID without compartment (e.g., “isocapcoa”, not “isocapcoa_c”), and {formula} is the chemical formula to assign to the metabolite once included in the reference. Similarly, the rule “charge.{mid}:{charge}” can be used to force the desired

charge on a particular metabolite (regardless of the compartment in which it appears). To impose a particular reaction balancing, users can write the rule "reaction.{rid}:{rstring}", where {rid} is the reaction ID and {rstring} is the reaction string, e.g. "reaction.NTD12:dimp_c + h2o_c --> din_c + pi_c + h_c". To not include a particular repository reaction, the rule "blacklist.{rid}" can be used. With these four different kinds of rules, users are enabled to create totally balanced and stoichiometrically consistent pan-GSMMs (provided an input reference with these same characteristics).

Following the procedure described above, the reference GSMM is expanded in terms of metabolites, reactions, and gene clusters, exploiting the information contained in the reference-free reconstruction. Since the expanded reference accounts for new strain-specific features, it can be referred to as a draft reference-based pan-GSMM.

The optional expansion with TCDB

At this point, a draft pan-GSMM is available, whether it is reference-based or reference-free. Users have the option to expand its transport reactions, using information derived from TCDB [67].

All the TCDB sequences were downloaded as aminoacidic multifasta on 8th april 2024 from http://www.tcdb.org/public/tcdb. Their IDs were formatted as "TCDB.{code}_comp{n}", where {code} is the TC code of a specific protein complex and {n} is an incremental integer number assigned to the members of that complex. A DIAMOND [158] database was generated from this multifasta file and then stored as a Gempipe asset, together with the table "tcdb_gprs.csv". The latter listed all the protein complexes needed to sustain each transport system, using the same formatting of the "bigg_gprs.csv" file taken from CarveMe [44]. Next, a table listing a set of Chebi ontology [189] codes for each transport system was downloaded from https://www.tcdb.org/cgi-bin/substrates/getSubstrates.py. To convert each Chebi ID into BiGG metabolite IDs, the MetaNetX v4.4 [186] mappings were used. In particular, the notion of "child" metabolite offered by MetaNetX was used, where for example "alpha-D-glucose" and "beta-D-glucose" are both childs of "D-glucose". Therefore, each Chebi ID was first convert into a MetaNetX ID, and then in turn converted to BiGG ID, recursively looking among the MetaNetX childs when no direct conversion to BiGG existed (for example, the Chebi code 16362 described as "D-aldose" has no direct conversion to BiGG, so its MetaNetX childs were used, resulting in a final conversion equal to {"man", "gal", "gal_bD", "glc__D"}). However, MetaNetX mappings are not complete, so a set of manually defined conversions was hard-coded. Next, to extract important TC code-specific

annotations such as the life domain and the textual description, all the TCDB superfamily codes were collected, linking each of them to a representative family code, and then a single html page for each superfamily was downloaded using the lynx[11] v2.8.9.1 terminal-based web browser with *-source http://www.tcdb.org/search/result.php?tc={family}*, where {family} is the representative family code. Annotations were extracted from each html file with the BeautifulSoup[12] v4.12.3 library, and the list of working TC code was restricted to "Bacteria" filtering on the life domain annotation.

With all the gathered data, an attempt was made to build a BiGG-compatible reaction string for each of the remaining TC codes. As the number of TCDB superfamilies [67] is huge, it was decided to handle only some of them (2.A.1, 2.A.2, 2.A.3, 2.A.6, 2.A.8, 2.A.10, 2.A.14, 2.A.18, 2.A.21, 2.A.25, 2.A.26, 2.A.27, 2.A.46, 2.A.50, 3.A.1, 4.A.*, 4.C.1) and to divide the respective TC codes into 6 groups, characterized by a different template reaction string ("abc", "uni", "sym", "anti", "pts", "coa"). Since the TCDB-provided Chebi annotations [189] are not complete and do not distinguish between primary substrates and co-transported ions, the first sentence of the textual description was processed to extract eventual co-transported ions via an exact substring matching. With the same method, the transport direction ("uptake" or "secretion") was determined, while the reversibility ("-->" or "<=>") was assigned depending on the superfamily code. Then, for each TC code, a number of reaction strings was generated, being 0 if no BiGG metabolite ID was assigned, and > 1 in case of promiscuous transporter or alternative cotransported ions. Regarding the template reaction strings, group "abc" (ATP-binding cassette-like) was defined by "atp_c + h2o_c + {mid}_e --> {mid}_c + pi_c + h_c + adp_c" in case of uptake, otherwise "atp_c + h2o_c + {mid}_c --> {mid}_e + pi_c + h_c + adp_c" in case of secretion, where {mid} is the primary substrate; group "uni" (uniport-like) was defined by "{mid}_e <=> {mid}_c"; group "sym" (symport-like) was defined by "{mid}_e + {cot}_e {arrow} {mid}_c + {cot}_c" where {arrow} describes the reversibility or the transport direction, and {cot} is a cotransported ion; group "anti" (antiport-like) was defined by "{mid}_e + {cot}_c {arrow} {mid}_c + {cot}_e"; group "pts" (phosphotransferase system-like) was defined by "{mid}_e + pep_c --> {mid_pi}_c + pyr_c", where {mid_pi} is the BiGG ID for the substrate bonded to a phosphate group; group "coa" (fatty acid translocation-like) was defined by "{mid}_e + h_e + coa_c --> {mid_coa}_c + h_c + ppi_c", where {mid_coa} is the BiGG ID for the substrate bonded to a coenzyme-A (CoA). The BiGG-compatible reaction strings generated per TC code were tabled into the "tcdb_rs.csv" file, which was saved as Gempipe asset.

[11] https://anaconda.org/conda-forge/lynx
[12] https://anaconda.org/conda-forge/beautifulsoup4

If requested by users, the draft pan-GSMM is inflated with TCDB-derived transport reactions. First, the representative sequences of the clusters are aligned against the previously generated TCDB database using *diamond blastp --ultra-sensitive*. Then, the alignment is filtered to retain HSPs having identity at least 45%, positivity at least 60%, and query and subject coverage at least 80%. Using the remaining HSPs, the presence of each TCDB-derived reaction contained in "tcdb_rs.csv" is evaluated by using the exact same method described for the reference-free reconstruction, but replacing the "bigg_gprs.csv" file with "tcdb_gprs.csv". Therefore, a TCDB-derived reaction is added to the draft pan-GSMM only if each of the members of the transport complex is detected, this way respecting the protein complex definitions provided by TCDB [67]. Anyway, since the TCDB-derived transport reactions are created simply by following a reaction string template which is independent from any existing model, their introduction into the draft pan-GSMM could result in mass or charge unbalances or stoichiometric inconsistencies [41]. Given this downside, the expansion of transport reactions using information derived from TCDB is not active by default.

Reannotation and optional deduplication

After the optional expansion with TCDB [67], the draft pan-GSMM is subjected to a *de novo* annotation of metabolites, reactions, genes, and SBO terms. This facilitates prospective uses of the output GSMMs, and can lead to better scores in the community standard test-suite MEMOTE [70]. To annotate metabolites and reactions, a set of pre-made mappings is used, mainly derived from MetaNetX v4.4 [186] .

Briefly, MetaNetX [186] was downloaded and parsed to obtain mappings between BiGG IDs and IDs of several other databases. Regarding metabolites, MetaNetX mappings comprised KEGG (Compound, Drug, and Glycan) [30], MetaCyc [31], HMDB [187], ModelSEED [29], Chebi [189], SABIO-RK [206], LIPID MAPS [207], enviPath [208], Reactome [209], Rhea [188], SwissLipids [210] and BiGG [28] itself. Structural annotations like InChI [211], InChIKey [211], and SMILES [212] were also included. To include the mappings for PubChem [213], the official translations between InChIKey and PubChem IDs were downloaded from the PubChem FTP server[13] (release "2023-12-08"), then the InChIKey was used to map the BiGG IDs against PubChem. Regarding reactions, MetaNetX mappings comprised KEGG, MetaCyc, ModelSEED, Rhea, SABIO-RK, and BiGG itself, plus the relative EC-codes. To include the mappings for Reactome, the official translations between Rhea and Reactome IDs were

[13] https://ftp.ncbi.nlm.nih.gov/pubchem/RDF/inchikey/pc_inchikey2compound_*.ttl.gz

downloaded from the Expasy [214] FTP server[14] on 26th January 2024, then the Rhea IDs were used to map the BiGG IDs against Reactome. Lastly, to include the mappings for Brenda [215], translations were downloaded from the official Brenda website[15] (release "2023-07-11"), including those for KEGG, MetaCyc, SABIO-RK, and EC-codes, and they were all used to map the BiGG IDs against Brenda. All the collected mappings between BiGG and other databases were stored as unique Gempipe asset. This asset is used for the one-step *de novo* annotation of metabolites and reactions in the draft pan-GSMM.

For the annotation of genes, a different approach is used. Gene annotations are provided only when genbank proteome files are given in input to Gempipe, retrieving the needed attributes as described above. However, as genes are strain-specific, gene annotations are actually included only during the generation of strain-specific GSMMs by "gempipe derive".

For the annotation of SBO terms [193], the contents of the draft pan-GSMM are iterated and categorized. SBO terms are written for genes (SBO:0000243), exchange reactions (SBO:0000627), sink reactions (SBO:0000632), demand reactions (SBO:0000628), biomass reactions (SBO:0000629), transport reactions (SBO:0000655), purely metabolic reactions (SBO:0000176), and metabolites (SBO:0000247).

Once the draft pan-GSMM has been reannotated, it can be optionally subjected to the removal of duplicated metabolites and reactions. This process can potentially lead to mass or charge unbalances or stoichiometric inconsistencies [41], so it is disabled by default.

Briefly, metabolites are grouped based on their MetaNetX [186] annotation, and groups with 2 or more elements are retained, thus isolating duplicated metabolites. For each group, one of the members is selected as the one to keep (M^K) in the draft pan-GSMM, giving priority to metabolites included in the reference, if provided. All the other members (M^D) are replaced with M^K, unless they are also included in the reference. Reactions in the GSMM are iterated, searching for duplicated metabolites to replace. In case M^K is already present in the draft pan-GSMM but in a different compartment, then M^K is created inheriting chemical formula, charge and annotations. At the end of this process, M^D metabolites are involved in 0 reactions, and thus they are removed from the draft pan-GSMM.

[14] https://ftp.expasy.org/databases/rhea/tsv/rhea2reactome.tsv
[15] https://bkms.brenda-enzymes.org/download/Reactions_BKMS.tar.gz

After duplicated metabolites are removed, duplicated reactions can be detected. Reactions are grouped based on their MetaNetX [32] annotation, retaining groups with more than 1 member (exchange reactions excluded). As for duplicated metabolites, a reaction is selected for each group as the one to keep (R^K) in the draft pa-GSMM. The other members (R^D) are replaced with R^K, unless they are also included in the reference, or unless they have a different set of involved metabolites (excluding protons). When removing R^D reactions, their GPR is transferred inside R^K, concatenating them with the boolean operator "OR".

API for manual curation

In addition to the command-line programs, Gempipe provides an API which comprises functions based on COBRApy [81] for speeding up certain aspects of the manual curation of a GSMM, including the draft pan-GSMM produced by "gempipe recon". Some examples are reported below, grouped by topic.

(i) Sanity and formatting: isolation of energy-generating cycles (EGCs) [107]; overview of "artificial" atoms in metabolites chemical formula (i.e. "X" groups, "R" groups, etc); definition of the unconstrained flux constant (e.g. 1000 or 999999); identification of constrained metabolic reactions. (ii) Network topology: gap-filling for the production of a specified metabolite; biosynthesis verification for metabolites through demand reactions; biosynthesis verification of reactants of a given reaction (e.g. useful to detect blocked biomass precursors); quick import of reactions from a repository GSMM; quick definition of new metabolites and reactions. (iii) Nutritive inputs: facilitated setting of the exchange reactions to represent growth media; sensitivity analysis (eventually scaled) [113] to reveal missing or limiting nutrients.

Some functions were specifically developed to work with the outputs of "gempipe recon". For example, the PAM can be quickly queried to retrieve modeled and unmodeled gene clusters based on the eggNOG-mapper functional annotation [182]. Specifically, the search is based on the KEGG Orthology (KO) codes [216], the KEGG Reaction codes [30], the PFAM domains [217], the EC- and TC-codes [67], and the textual description provided by the eggNOG database [218]. Moreover, during the curation of the pan-GSMM, strain-specific GSMMs can be quickly previewed and used to simulate the Biolog® Phenotype MicroArray™ (PM) screening system (see below for the implementation). This is particularly useful to verify the efficacy of the manual curation before starting to derive strain-specific GSMMs with "gempipe derive". For further information on the curation functions included in the API, the

reader is referred to the Gempipe documentation available on ReadTheDocs, which includes *ad hoc* tutorials (https://gempipe.readthedocs.io/en/latest/).

Derivation of strain-specific GSMMs

"gempipe derive" takes the PAM and the curated pan-GSMM as main inputs, and produces a strain-specific GSMM for each of the strains listed in the PAM.

For each strain, a copy of the pan-GSMM is made. When the strain has no genes (including recovered genes) in a cluster, then the cluster is removed, leading to an eventual loss of reactions; if all the genes in a cluster are characterized by premature stop codons, then the cluster is removed anyway. Next, reactions are iterated updating their GPR: each remaining cluster is replaced by the corresponding strain-specific genes; when genes are more than 1, the cluster is replaced by an "OR"-block enclosed by parenthesis, where alternative genes are connected by the boolean operator "OR". Modeled genes are annotated if the MIRIAM-compliant [140] gene annotations ("refseq", "ncbiprotein", "ncbigene", "kegg", and "uniprot") produced by "gempipe recon" were given in input.

At this point, each model is gap-filled using a user-provided recipe for a medium, best if minimal, known or assumed to support growth of all the input strains. The file provided is encoded in Json, and lists lower bounds to apply to the respective exchange reactions; moreover, it allows the definition of several recipes at the same time, leading to multiple rounds of gap-filling. If no media file is provided, a generic minimal aerobic medium recipe is loaded, defined by the unconstrained availability of H_2O, H^+, O_2, glucose, phosphate, sulfate, ammonia, Ca^{2+}, Cl^-, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, K^+, Mg^{2+}, Mn^{2+}, Zn^{2+} and MoO_4^{2-}. The gap-filling algorithm applied is the one provided by COBRApy [81], using the pan-GSMM as source of reactions and a user-selectable minimum flux through the objective reaction. Eventual disconnected metabolites (i.e., involved in 0 reactions), consequence of the removal of gene clusters, are deleted from each strain-specific GSMM before it is saved to disk. Moreover, the strain-specific gap-filling phase can optionally be skipped, which is useful for example when auxotrophies has to be studied on a minimal medium [177].

Once strain-specific GSMMs are created, simulations inspired by the Biolog® Phenotype MicroArray™ (PM) screening system are performed if requested. Briefly, for each Biolog® PM plate, the substrate present in each well was collected and mapped to the corresponding BiGG [28] exchange reaction, creating a resource saved as "biolog_mappings.csv" among the Gempipe assets. Once the main sources for C, N, P, and S are defined – by default glucose,

ammonia, phosphate, and sulfate, respectively – wells of each plate are iterated. The exchange reaction for the main C, N, P, or S source is closed (lower bound = 0), depending on how the substrate is categorized in each well (for example, the L-glutamic acid is tested as a C source in PM01 well B12, but also as a N source in PM03 well A12). Growth is then simulated using flux-balance analysis (FBA) and the objective value (O_0) is recorded. Next, the exchange reaction corresponding to the well is opened (unconstrained input) and FBA is performed again, recording the new objective value (O_1). If $O_1 \geq O_0 + 0.001$, and the solver status is "optimal", then the strain is considered as able to utilize the substrate of that well.

An optional mode allows the simulation of alternative C, N, P and S sources on a minimal medium. In this case, a minimal medium is computed and applied to the GSMM using the *minimal_medium* function from COBRApy [81], specifying the minimal objective value to guarantee and activating the option *minimize_components*. Medium components are sorted by descending lower bound (considered with no sign), and the main sources of C, N, P and S are searched in the sorted list, taking the first containing a C, N, P or S atom, respectively, and giving priority to glucose, ammonia, phosphate, and sulfate, respectively, if present in the minimal medium. This may be useful when strains grow on complex media, as the contribution of some alternative substrates could be masked by other medium components when performing FBAs.

Strain-specific GSMMs are also used to create several binary feature tables, adhering to the same structure: columns contain strains (as reported in the PAM), rows contain binary features, cells contain 1 if the feature is present in the strain, otherwise 0. These tables can be useful during the multi-strain analysis, and are directly compatible with dedicated functions of the Gempipe API. The first binary feature table created is a reaction presence / absence matrix ("rpam.csv"), reporting which reactions are modeled for each strain. Other binary tables are then optionally created, such as the one reporting the presence of auxotrophies ("aux.csv"), or the one reporting the growth-enabling substrates ("cnps.csv"). The respective implementation methods are reported below.

To detect auxotrophies, a list of *n* compounds to test is defined. The default list include all the twenty protein-forming amino acids, plus the following vitamins / cofactors: biotine, folate, lipoate, pantothenate, pyridoxine, pyridoxamine, pyridoxal, riboflavin, thiamine, nicotinate, para-aminobenzoate, cobalamin, ascorbate. For each strain-specific GSMMs, the list of compounds is iterated: the exchange reaction for a compound is closed (lower bound = 0), while those for the other *n*-1 compounds are opened (unconstrained uptake). Then, FBA is

performed: if the solver status is “optimal” and the objective value is > 0.001, then the strain is not considered as auxotroph for the compound [9] and the respective cell in the binary feature table (“aux.csv”) is set to 0, otherwise 1.

To detect growth-enabling substrates, the exchange reactions of a strain-specific GSMM are first classified as C, N, P or S sources, depending on the presence of at least one atom of C, N, P or S, respectively (multiple classifications are possible for the same exchange reaction). Similarly to Biolog® simulations, once the main sources for C, N, P, and S are defined – by default glucose, ammonia, phosphate, and sulfate, respectively – all the exchange reactions for C sources are iterated. The exchange reaction for the starting C source is closed (lower bound = 0), growth is simulated using FBA and the objective value (O_0) is recorded. Then, the exchange reaction for the alternative C source is opened (unconstrained uptake) and FBA is performed again, recording the new objective value (O_1). If $O_1 \geq O_0 + 0.001$, and the solver status is “optimal”, then the strain is considered as able to utilize the alternative C source [9] and the respective cell in the binary feature table (“cnps.csv”) is set to 1, otherwise 0. The procedure is then repeated for N, P, and S sources.

Gempipe supports the processing of more than 1 species at a time. Using the reactions presence / absence information contained in “rpam.csv”, a species-specific GSMM is derived for each input species, defined by the set of reactions that are always present in all the strain-specific GSMMs belonging to the same species. This enables possible comparative studies at the species level, for example to determine which metabolic features distinguish a species from another. Once created, species-specific models are gap-filled on the same set of growth media used to gap-fill strain-specific models, providing the pan-GSMM as source of reactions; however, this species-specific gap-filling phase can optionally be skipped.

The “autopilot” mode

“gempipe autopilot” is the third and last command-line program provided, which creates strain-specific GSMMs directly from genomes or proteomes, without going through manual curation. This is done by internally calling “gempipe recon” and “gempipe derive”, automatically gap-filling the draft pan-GSMM for biomass production. Since “gempipe autopilot” calls “gempipe recon” at the startup, then all the four types of inputs previously discussed are supported: genbank proteomes, multifasta proteomes, assembled genomes, and species taxids.

Once the draft pan-GSMM is built, an expanded universal GSMM is created to be used as the source of reactions during its gap-filling. The expanded universal GSMM is created from a copy of the draft pan-GSMM, which is supplemented with additional reactions. In particular, all the reactions present in the selected CarveMe universe [44] or in the reference (if provided), if not yet modeled, are copied into the expanded universal GSMM. During this expansion, if a reaction appears with the same ID in both the reference and the CarveMe universe, then the reaction and the associated metabolites are copied as defined in the reference. Moreover, the same optional file of manual corrections adopted in "gempipe recon" is used to superimpose particular metabolite formulas and charges or reaction balances during the copy of reactions.

A prioritized gap-filling is performed using the COBRApy [81] built-in function to ensure biomass production, providing a penalty for each reaction in the expanded universal GSMM. With this approach, missing reactions are added to the draft pan-GSMM favoring those with better genetic support (better alignment scores). Penalty is equal to 0 for each reaction already present in the draft pan-GSMM; for the other reactions, penalty is defined as $1 / (1 + s)$, where s is the normalized reaction score computed using the same method described above for the reference free reconstruction, but with the following exceptions. (i) The DIAMOND database is created by expanding "bigg_proteins.faa" with the protein sequences coming from the reference, if provided. (ii) The normalized reaction scores are computed starting from an expanded version of "bigg_gprs.csv", containing also the genes, protein complexes, and reactions of the reference, if provided. (iii) The HSPs produced during the alignment are filtered using relaxed thresholds (identity at least 10%, coverage at least 40%), this way obtaining normalized reaction scores for a broader set of reactions compared to those already included in the draft pan-GSMM. (iv) reactions with no alignment-based genetic support are associated with a normalized reaction score of 0.

The same user-provided growth media used in "gempipe derive" are iteratively applied to the draft pan-GSMM, and a round of gap-filling is applied for each medium. If the medium requires the presence of exchange reactions not available in the draft pan-GSMM, then they are imported from the expanded universal GSMM beforehand. Once the draft pan-GSMM has been gap-filled to produce biomass on all the user-provided media, then it is used together with the PAM to derive the strain-specific GSMMs, internally calling "gempipe derive".

API for multi-strain analysis

The Gempipe API contains dedicated functions for multi-strain analysis. For example, any number of binary feature tables (such as those produced by "gempipe derive" for reaction content, auxotrophies and growth-enabling substrates) can be combined and used to create "phylometabolic" trees. Then, clusters of metabolically coherent strains can be defined, and features that characterize each cluster can be extracted.

Briefly, input binary feature tables are combined to form a single feature table, based on which the pairwise similarity between strains is computed using the Jaccard index. The produced distance matrix is processed to create a dendrogram using the Ward's agglomerative clustering [114]. The distance matrix and the dendrogram are created using the *linkage* and *pdist* functions from the scipy library, respectively.

Based on the dendrogram, clusters of metabolically coherent strains are extracted. To find the optimal number of clusters to extract, a silhouette analysis is performed: for each number of clusters, the average Silhouette score [219] is computed using the *silhouette_score* function from the sklearn library; the number of clusters that maximizes the Silhouette score is used, unless a user-specified number of clusters is given. Once clusters are extracted, it is visually possible to compare the cluster attribute to other strain-specific attributes, for example the species of origin or the environmental niche.

To extract features that characterize each cluster, varying features (i.e., features whose presence or absence is not constant in all the strains) are selected. Then, the frequency of presence of each feature is computed relatively to each cluster. Features having a relative frequency of ≥0.9 and ≤0.1 in two different clusters are selected. Based on the relative frequencies, pairwise similarity between features is computed using the Jaccard index. Features are represented in a heatmap, where they are sorted according to the leaves order of a dendrogram obtained by using Ward's agglomerative clustering on the Jaccard distance matrix.

2.5.3 Supplementary Figures

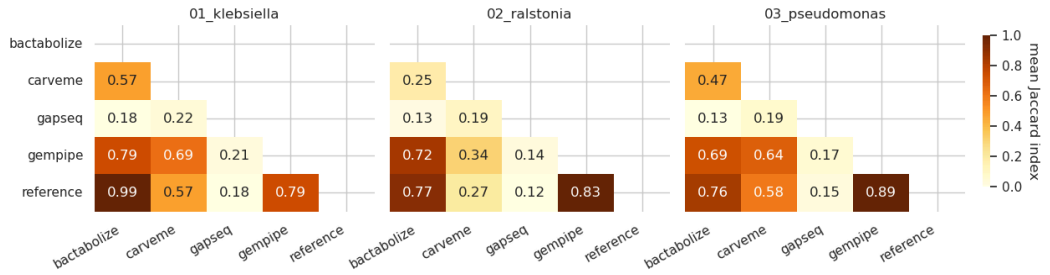

Figure 2.S1. Similarity between tools based on reaction content. Cells report the mean Jaccard index along the strains, computed for the reaction IDs.

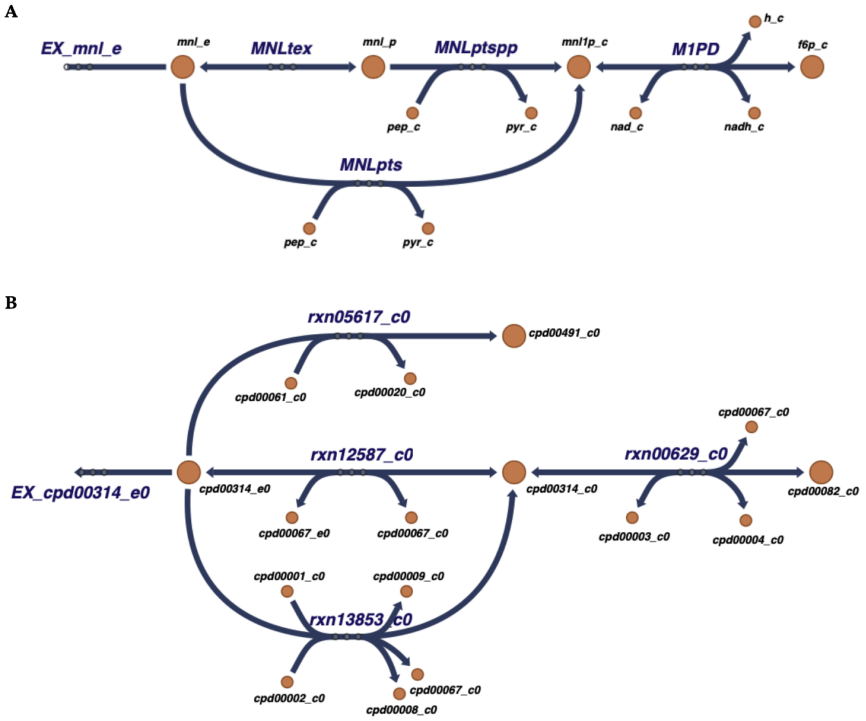

Figure 2.S2. Uptake of mannitol and relative connection to the central carbon metabolism. **A)** Representation in the CarveMe gram-negative universe [44]. M1PD: mannitol-1-phosphate 5-dehydrogenase; MNLpts: mannitol transport via PEP:Pyr PTS; MNLptspp: mannitol transport via PEP:Pyr PTS (periplasm); MNLtex: mannitol transport via diffusion (extracellular to periplasm); EX_mnl_e: exchange reaction for mannitol. **B)** Representation in the GSMM produced by gapseq [65] for *Pseudomonas chlororaphis* TAMOak81. EX_cpd00314: exchange reaction for mannitol; rxn05617: mannitol transport via PEP:Pyr PTS; rxn12587: mannitol transport via proton symport; rxn13853: mannitol transport via ABC; rxn00629: D-mannitol:NAD+ 2-oxidoreductase; cpd00314: D-mannitol; cpd00491: D-glucitol-6-phosphate; cpd00061:

phosphoenolpyruvate; cpd00020: pyruvate; cpd00067: H^+; cpd00001: H_2O; cpd00002: ATP; cpd00009: phosphate; cpd00008: ADP; cpd00003: NAD; cpd00004: NADH; cpd00082: fructose.

2.5.4 Supplementary Tables

Table 2.S1. Genomes of the *Klebsiella* dataset.

	strain	accession
1	*Klebsiella africana* 200023T	GCA_020526085.1
2	*Klebsiella pneumoniae* 03-9138-2	GCA_020526045.1
3	*Klebsiella pneumoniae* 04A025	GCA_020526025.1
4	*Klebsiella pneumoniae* BJ1	GCA_020526065.1
5	*Klebsiella pneumoniae* CG43	NC_022566.1 + NC_005249.1
6	*Klebsiella pneumoniae* CIP 52.145	GCA_000968155.1
7	*Klebsiella pneumoniae* KP13	GCA_000512165.1
8	*Klebsiella pneumoniae* MGH 78578	GCF_000016305.1
9	*Klebsiella pneumoniae* NJST258-1	GCF_000598005.1
10	*Klebsiella pneumoniae* NTUH K2044	GCF_000009885.1
11	*Klebsiella pneumoniae* SA1	GCA_020525885.1
12	*Klebsiella pneumoniae* SA12	GCA_020525485.1
13	*Klebsiella pneumoniae* SB1067	GCA_020525845.1
14	*Klebsiella pneumoniae* SB1139	GCA_020525505.1
15	*Klebsiella pneumoniae* SB1170	GCA_020525645.1
16	*Klebsiella pneumoniae* SB2390	GCA_020526005.1
17	*Klebsiella pneumoniae* SB611	GCA_020525985.1
18	*Klebsiella pneumoniae* SB612	GCA_020525965.1
19	*Klebsiella pneumoniae* SB615	GCA_020525945.1
20	*Klebsiella pneumoniae* SB617	GCA_020525905.1
21	*Klebsiella pneumoniae* T69	GCA_020525865.1
22	*Klebsiella quasipneumoniae* subsp. *quasipneumoniae* 01A030T	GCA_020525925.1
23	*Klebsiella quasipneumoniae* subsp. *quasipneumoniae* 18A069	GCA_020525805.1
24	*Klebsiella quasipneumoniae* subsp. *quasipneumoniae* SB1124	GCA_020525825.1
25	*Klebsiella quasipneumoniae* subsp. *quasipneumoniae* SB98	GCA_020525785.1
26	*Klebsiella quasipneumoniae* subsp. *quasipneumoniae* U41	GCA_020525725.1
27	*Klebsiella quasipneumoniae* subsp. *similipneumoniae* 07A044T	GCA_020525685.1
28	*Klebsiella quasipneumoniae* subsp. *similipneumoniae* 09A323	GCA_020525585.1
29	*Klebsiella quasipneumoniae* subsp. *similipneumoniae* 12A476	GCA_020525525.1
30	*Klebsiella quasipneumoniae* subsp. *similipneumoniae* CIP 110288	GCA_020525565.1
31	*Klebsiella quasipneumoniae* subsp. *similipneumoniae* SB610	GCA_020525705.1
32	*Klebsiella quasivariicola* 08A119	GCA_020525665.1
33	*Klebsiella variicola* subsp. *tropica* CDC 4241-71	GCA_020525625.1
34	*Klebsiella variicola* subsp. *variicola* 01A065	GCA_020525605.1
35	*Klebsiella variicola* subsp. *variicola* 342	GCF_000019565.1
36	*Klebsiella variicola* subsp. *variicola* At-22	GCF_000025465.1
37	*Klebsiella variicola* subsp. *variicola* F2R9T	GCA_020525545.1

Table 2.S2. Genomes of the *Ralstonia* dataset.

	strain	accession
1	*Ralstonia pseudosolanacearum* GMI1000	GCA_029220065.1
2	*Ralstonia pseudosolanacearum* PSS4	GCA_029220045.1
3	*Ralstonia pseudosolanacearum* RUN2340	GCA_029220025.1
4	*Ralstonia solanacearum* CFBP2957	GCA_029220005.1
5	*Ralstonia solanacearum* K60	GCA_029219985.1
6	*Ralstonia solanacearum* BA7	GCA_029219965.1
7	*Ralstonia solanacearum* UW551	GCA_029219945.1

8	*Ralstonia solanacearum* MOLK2	GCF_000009125.1
9	*Ralstonia syzygii* PSI07	GCF_002251695.1
10	*Ralstonia syzygii* R24	GCF_002251655.1
11	*Ralstonia syzygii* BDBR229	GCF_000283475.1

Table 2.S3. Genomes of the *Pseudomonas* dataset.

	strain	accession
1	*Pseudomonas chlororaphis* 30-84	GCF_000281915.1
2	*Pseudomonas chlororaphis* ATCC17415	GCF_003851145.1
3	*Pseudomonas chlororaphis* TAMOak81	GCF_003850605.1
4	*Pseudomonas chlororaphis* subsp. *piscium* ATCC17411	GCF_003850385.1
5	*Pseudomonas chlororaphis* subsp. *piscium* ATCC17809	GCF_003850365.1
6	*Pseudomonas chlororaphis* subsp. *piscium* ChPhzS135	GCF_003850485.1
7	*Pseudomonas chlororaphis* subsp. *piscium* DTR133	GCF_003850425.1
8	*Pseudomonas chlororaphis* subsp. *piscium* SLPH10	GCF_003850405.1
9	*Pseudomonas chlororaphis* subsp. *piscium* ToZa7	GCF_003850585.1
10	*Pseudomonas chlororaphis* subsp. *piscium* ChPhzTR44	GCF_003850525.1
11	*Pseudomonas chlororaphis* subsp. *piscium* ChPhzS140	GCF_003850505.1
12	*Pseudomonas chlororaphis* subsp. *piscium* PCL1607	GCF_003850465.1
13	*Pseudomonas chlororaphis* subsp. *piscium* PCL1391	GCF_003850445.1
14	*Pseudomonas chlororaphis* subsp. *piscium* DSM 21509	GCF_003850345.1
15	*Pseudomonas chlororaphis* B25	GCF_003851985.1
16	*Pseudomonas chlororaphis* Pb-St2	GCF_003851785.1
17	*Pseudomonas chlororaphis* subsp. *aurantiaca* CW2	GCF_003851225.1
18	*Pseudomonas chlororaphis* subsp. *aurantiaca* 449	GCF_003851205.1
19	*Pseudomonas chlororaphis* subsp. *aurantiaca* 464	GCF_003851805.1
20	*Pseudomonas chlororaphis* subsp. *aurantiaca* DSM 19603	GCF_003851835.1
21	*Pseudomonas chlororaphis* subsp. *aurantiaca* PCM2210	GCF_003851305.1
22	*Pseudomonas chlororaphis* subsp. *aurantiaca* Q16	GCF_003851345.1
23	*Pseudomonas chlororaphis* subsp. *aurantiaca* M12	GCF_003851165.1
24	*Pseudomonas chlororaphis* subsp. *aurantiaca* K27	GCA_003851285.1
25	*Pseudomonas chlororaphis* subsp. *aurantiaca* M71	GCF_003851265.1
26	*Pseudomonas chlororaphis* subsp. *aureofaciens* P2	GCF_003851365.1
27	*Pseudomonas chlororaphis* subsp. *aureofaciens* C50	GCF_003851385.1
28	*Pseudomonas chlororaphis* subsp. *aureofaciens* DSM 6698	GCF_003851905.1
29	*Pseudomonas chlororaphis* subsp. *aureofaciens* ChPhzS23	GCA_003851425.1
30	*Pseudomonas chlororaphis* subsp. *aureofaciens* 66	GCF_003851405.1
31	*Pseudomonas chlororaphis* subsp. *aureofaciens* ChPhzS24	GCF_003851445.1
32	*Pseudomonas chlororaphis* subsp. *aureofaciens* ChPhzTR38	GCF_003852005.1
33	*Pseudomonas chlororaphis* subsp. *aureofaciens* ChPhzTR39	GCF_003851925.1
34	*Pseudomonas chlororaphis* subsp. *aureofaciens* ChPhzTR18	GCF_003851955.1
35	*Pseudomonas chlororaphis* subsp. *aureofaciens* PA23	GCF_000698865.1
36	*Pseudomonas chlororaphis* subsp. *aureofaciens* O6	GCF_000264555.1

Table 2.S4. Comparison between experimental and simulated binarized Biolog® PM screenings. Rows are substrates, columns are strains. Cells follow the format "{E}-{T1}-{T2}-{T3}-{T4}", where {E} is the experimental binary growth output (0: no growth; 1: growth); {T1}, {T2}, {T3}, {T4} are the matches of Gempipe, CarveMe, gapseq, and Bactabolize, respectively ("TP": true positive; "TN": true negative; "FP": false positive; "FN": false negative; "in": infeasible FBA solution). *Table too large for printing - please consult at the following link* [*https://docs.google.com/spreadsheets/d/1nshybycsTugsoig29BBSJIwVdXKBxKyP/edit?usp=sharing&ouid=102382789494211499514&rtpof=true&sd=true*](https://docs.google.com/spreadsheets/d/1nshybycsTugsoig29BBSJIwVdXKBxKyP/edit?usp=sharing&ouid=102382789494211499514&rtpof=true&sd=true)

Table 2.S5. Recovered genes in strain-specific GSMMs. Cells contain gene IDs, rows contain gene cluster IDs where genes belong to; columns contain strains. *Table too large for printing - please consult at the following link https://docs.google.com/spreadsheets/d/1CwbJkg7pwvMqFrRcoGU9R7SdrrpUH0Qc/edit?usp=sharing&ouid=102382789494211499514&rtpof=true&sd=true*

Table 2.S6. Reactions modeled in strain-specific GSMMs as consequence of gene recovery. Rows contain reaction IDs, columns contain strains, cells describe reaction presence (1) or absence (0) in a particular strain. *Table too large for printing - please consult at the following link https://docs.google.com/spreadsheets/d/143dpJxeDcXW_RqLyA_tUIf_ZpQ-4hAvF/edit?usp=sharing&ouid=102382789494211499514&rtpof=true&sd=true*

Chapter 3 | Formulating a growth medium for the endosymbiont *Candidatus* Erwinia dacicola: metabolic modeling and genomic insights from *Erwinia aphidicola*

3.1 Introduction

Insects have different types of symbiotic interactions with microorganisms. These associations can be facultative when they are not essential for survival and can be transient, sometimes acquired from the environment or other insects [220]. In contrast, the obligate symbionts are essential for the host insect's survival, playing a key role in the physiology of their hosts, contributing to fitness, nutrition, detoxification of compounds and reproduction [221–223]. These symbionts are typically vertically transmitted (from mother to offspring), and have a long evolutionary history with their host [220].

The genomes of insects' obligate endosymbionts are often highly reduced, typically around 1 Mbp or less, with high AT content [224]. Indeed, they are missing many genes and pathways compared to those of their free-living relatives [225–228]. These include functions for reproduction, cell membrane / wall construction, DNA repair, secondary metabolism, motility, and essential nutrient biosynthesis [229–232]. Genome reduction results from the redundancy of genes and pathways between the symbiont and its host, or among different symbionts, thus the gradual reduction of genetic material affects genes no more essential for the microorganism in a strict symbiotic relationship [230,233]. Due to these characteristics, obligate endosymbionts are usually unculturable or difficult to culture [229,234] delineating a common pattern of convergent genome evolution [235].

Candidatus Erwinia dacicola exemplifies the intimate and essential relationship between an insect host and its obligate symbiont. This bacterium can be found in the esophageal bulb, gut, crop, rectal sacs, ovipositor and larval midgut of *Bactrocera oleae* [236,237], an important pest of olive orchards. This insect, commonly known as the olive fly, lays its eggs inside green, unripe olive drupes: when growing, larvae produce holes that lead to secondary infections, causing severe economic losses in the olive markets [238]. *Ca.* Erwinia dacicola is the

predominant endosymbiont of *B. oleae*, and it could be the major contributor to the gut bacteria-derived advantages in its host [239,240]. These benefits include the provision of essential amino acids which are crucial for the fly's nutrition, converting non-essential amino acids and urea from bird droppings [241,242]. Importantly, *Ca.* Erwinia dacicola was proven to be strictly required for *B. oleae* larval development in unripe olives [239]. Consequently, the inhibition of this symbiotic relationship could be the basis for a pest control mechanism [243].

Several growth assays in solid and liquid media have been tested to isolate *Ca.* E. dacicola, but all attempts have been unsuccessful. Nutritive formulations included Brain Heart Infusion (BHI) [244], malt agar, MacConkey Agar, MRS Agar, LB medium, tryptone extract and yeast extract [236], Kings medium B, Nutrient broth agar, selective media for the genus *Erwinia* and crushed olive mesocarp with agar [245]. Additionally, NSA (Nutrient Sucrose Agar) medium with cycloheximide was used for samples from twigs, leaves, and olives affected by the olive fly [237].

Although Ca. E. dacicola has not yet been cultured in vitro, six *Ca.* E. dacicola metagenome-assembled genomes from various geographical provenances are available and, interestingly, *Ca.* E. dacicola has a high similarity with free-living cultivable plant pathogens like *Erwinia aphidicola*, *Erwinia persicina* and *Erwinia rhapontici*, based on phylogenetic and phylogenomic analyses [236,246]. This suggests a relatively recent adaptation of the endosymbiont to its host [230,247].

In the challenge of culturing an endosymbiont, a system biology approach provided by genome-scale metabolic modeling can be helpful [248]. In this modeling framework, functional gene annotation of a bacterial genome is processed to create a stoichiometrically-coherent network of metabolic reactions [249]. This also includes a biomass assembly reaction, which describes the amount of each metabolite contributing to dry cell weight [250]. The network can then be constrained with experimental data to simulate steady-state reaction fluxes inside the cell under different nutritive conditions. The resulting genome-scale metabolic model (GSMM) has predictive power that ranges from qualitative to quantitative, depending on how well the network is curated and on the type and amount of experimental data used to constrain the network [251].

Among their many applications [252,253], GSMMs have also been used to study insect endosymbionts, in particular: (i) the fragility of their metabolism, and its dependence on the host, which in many cases exerts control by substrate

provisioning [254,255]; (ii) their nutritional requirements and the transport systems required to obtain diverse host-provided substrates [256–258]; (iii) the metabolic complementation and metabolic segregation (and relative efficiency) when other co-primary symbionts are present [259–261]; (iv) the energy generation in case of highly restricted-genomes [262]; (v) the influence of facultative endosymbionts over primary ones, in terms of competition for resources and forced production of substrates [263,264].

Gap-filling in insect endosymbiont models has also been investigated [265]. This is a process where missing metabolic reactions (gaps) are added to a GSMM based on literature and/or genomic evidence, in order to enable the *in silico* formation of biomass or other metabolic objectives [266]. In general, gaps that in free-living species have to be gap-filled by means of algorithms or (preferably) manual curation, in endosymbionts could provide hints to explain their relationship with the host [258]. By extension, GSMMs can also provide valuable insights for defining a growth medium recipe. However, GSMMs are by definition based on a genome, so they are strain-specific. In order to obtain the metabolic features that define a species, strain variability must be managed. To cope with this issue, it is possible to extract and work with just the set of reactions that are always present in a species, according to its available genomes.

In this scenario, the aim of this study was to formulate a growth medium for the endosymbiont *Ca.* E. dacicola through a genome-scale metabolic modeling approach.

To this purpose, a phylogenomic analysis was conducted to define in detail the relationship among *Ca.* E. dacicola and other cultivable species of *Erwinia*, to determine the most closely related free-living species. Afterwards, species-level comparative metabolic modeling was performed, using the closest relative as a reference for *Ca.* E. dacicola. A chemically-defined medium (CDM) for *Erwinia* species [267] was included in models as the starting growth medium to be supplemented with additional compounds expected to fulfill the nutritional requirements of the endosymbiont. Further, antibiotic resistance genes were searched in genomes, allowing the selection of an antibiotic to be used as selective agent in cultivation trials. The closest relative *E. aphidicola* was used to validate the antibiotic resistance observed *in silico* and to determine the appropriate antibiotic concentrations for cultivation trials through phenotypic minimum inhibitory concentration (MIC) testing.

Together, comparative metabolic modeling and genomics brought to the formulation of a CDM to cultivate *Ca.* E. dacicola in vitro, which was tested

experimentally. The presence of *Ca.* E. dacicola in the grown cultures was detected through species-specific PCR, and culture purity was verified using PCR-DGGE (Denaturing Gradient Gel Electrophoresis) to confirm the exclusive presence of endosymbiont.

3.2 Results

3.2.1 Phylogenomic relationship of *Ca.* E. dacicola

141 genome assemblies were collected for *Ca.* E. dacicola and its known [236,246] free-living close relatives *E. aphidicola*, *E. persicina* and *E. rhapontici* (**Table 3.S1**). Genomes were annotated and filtered to retain only good-quality assemblies. Quality filtering was based on the expected number of universal single-copy orthologs in the Enterobacterales order, and resulted in 2 good-quality assemblies of *Ca.* E. dacicola, 35 of *E. aphidicola*, 16 of *E. persicina*, and 7 of *E. rhapontici* (**Figure 3.S1A**). When looking at the cumulative length of contigs of the quality-filtered genomes, *Ca.* E. dacicola showed the expected genome reduction of insect endosymbionts (**Figure 3.S1B**). Moreover, its assemblies resulted highly fragmented, in-line with the short-reads metagenomics nature of its sequencings available to date (**Figure 3.S1C**).

Genes predicted in the quality-filtered genomes were used to compute a pan-genome. The multiple sequence alignment of the identified core genes was used to create a phylogenomic tree, which showed *E. aphidicola* as the closest free-living relative of the endosymbiont *Ca.* E. dacicola (**Figure 3.1A**), having approximately 1156 core-genes in common (**Figure 3.1B**). Given its phylogenomic proximity, *E. aphidicola* was chosen as the reference free-living species to be compared to *Ca.* E. dacicola in subsequent analysis, enabling a better understanding of the metabolic adaptations specific to the endosymbiont. Interestingly, strain ZSR3 seemed to be unrelated to other strains of the same species: for this reason, it was excluded from the *E. persicina* core gene set computation.

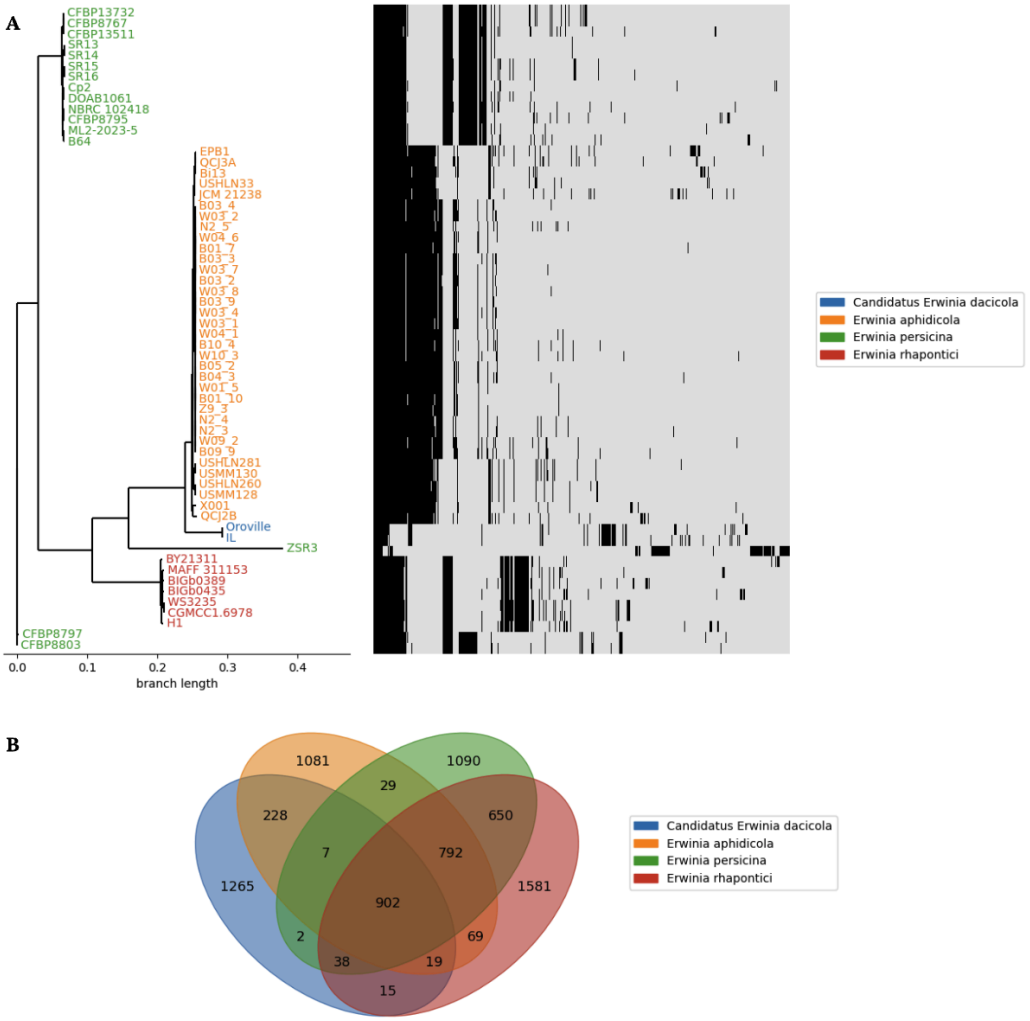

Figure 3.1. Pangenome analysis. **A)** Phylogenomic tree produced from a multiple sequence alignment of the core genes shared among the quality-filtered genomes. On the right, each bar represents the presence (black) or absence (gray) of a gene of the entire pan-genome computed for the quality-filtered genomes in input. **B)** Venn diagram of the core genes of each species, derived from the quality-filtered genomes. The set of core genes of *E. persicina* was computed excluding the outlier strain ZSR3.

3.2.2 Comparative metabolic modeling for devising growth requirements

The reconstruction of species-specific GSMMs was carried out with Gempipe (see **Chapter 2**). No automated gap-filling was applied during their reconstruction, in contrast to other state-of-art reconstruction pipelines [268–270].

First, a draft pan-GSMM was created, representing the entire set of metabolic reactions encoded by the quality-filtered set of genomes. Then, the draft pan-GSMM was manually curated using the Gempipe API, gap-filling for biomass production on a *Erwinia*-specific CDM [267] (**Table 3.S2**).

Once gap-filled, the pan-GSMM was used to derive strain-specific GSMMs, one for each quality-filtered genome. In this process, all gene clusters without strain-specific member genes are subtracted from a copy of the draft pan-GSMM, causing the removal of reactions. At this point, the knowledge of the reactome of each strain enabled the determination of the core reactome of species. This would not have been possible at a gene cluster level, as gene clusters are created based on high sequence identity (90%, see **Chapter 2**), meaning that two sufficiently dissimilar ortholog genes (same function) could be grouped in two different clusters. Therefore, species-specific GSMMs were created, composed by the core set of reactions always present in all the strain-specific GSMMs belonging to the same species.

Species-specific metabolic models were composed by 1642 reactions in common between all the input species, plus 31, 736, 701, and 877 reactions componing the core set *Ca.* E. dacicola, *E. aphidicola*, *E. persicina* and *E. rhapontici*, respectively (**Figure 3.2**).

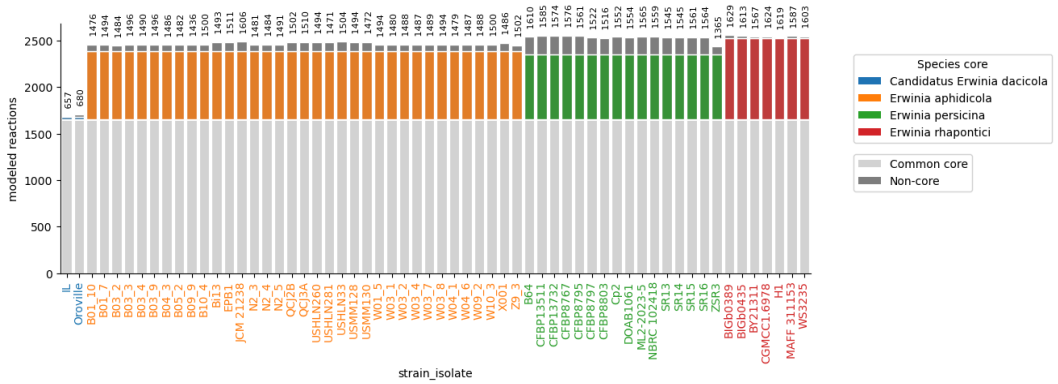

Figure 3.2. Number of modeled reactions for each quality-filtered genome. Strain-specific reactions are colored in light gray; species-specific reactions are colored according to the species; reactions in common between all the genomes are colored in dark gray. On top of the bars is reported the number of strain-specific modeled genes.

Given the input genomes, Gempipe took care not only of the reconstruction of every GSMM, but also of the previous steps of genome filtering and annotation, on which the phylogenomics analysis depended. Therefore, while species-specific GSMM were built for all the three free-living species considered

at the start, the following comparison was focused exclusively between the species-specific GSMMs for *E. aphidicola* (core-Eap) and *Ca.* E. dacicola (core-Eda), due to the indication provided by the phylogenomic tree. These two models were again manually gap-filled for the production of biomass on the CDM, inspecting each blocked biomass precursor one by one, and introducing the missing reactions with a genome-based evidence.

When simulating growth on the CDM, biomass formation was predicted for core-Eap, while it was prevented for core-Eda, due to a list of blocked biomass precursors that it was not possible to solve:

- 10-formyltetrahydrofolate (10fthf),
- phosphatidylethanolamine (pe160/pe161),
- thiamine diphosphate (thmpp),
- Kdo2-Lipid A (kdo2lipide4), and
- menaquinol-8 (mql8).

Of these, the first two are reported in 3.5.1 Supplementary Results as they were blocked likely due to technical issues, while the others are reported below.

The thiamine biosynthetic pathway resulted as heavily incomplete in core-Eda: genes underlying the key reactions phosphomethylpyrimidine kinase (PMPK, EC 2.7.4.7) and thiamine-phosphate diphosphorylase (TMPPP, EC 2.5.1.3) were absent (**Figure 3.3A**, **3.S2**). However, like in core-Eap, a thiamine transporter was present (THMabcpp), corresponding to ThiBPQ [271] (TC code 3.A.1.19.1) and composed by the orthologs [272] K02062, K02063, and K02064. This suggested an active uptake of the vitamine coming from the host or other insect gut bacteria.

The biosynthesis of Kdo2-Lipid A starts from UDP-N-acetylglucosamine, to which two varying 3-OH fatty acid chains are added [273]. The CarveMe gram negative universe [268] models the 3-OH fatty acids as the (R)-3-hydroxytetradecanoyl-CoA (R_3hmrscoa), which is derived from precursors coming form the fatty acid beta-oxidation, trans-tetradec-2-enoyl-CoA (td2coa) and 3-oxotetradecanoyl-CoA (3otdcoa). Despite the Kdo2-Lipid A biosynthetic pathway being complete in core-Eda, the beta-oxidation pathway was totally missing (**Figure 3.3B**, **3.S3**), theoretically preventing the formation of these 3-OH fatty acids precursors.

The pathway for the biosynthesis of menaquinol-8 was deeply gapped due to many missing reactions (**Figure 3.3C**, **3.S4**). Most of these reactions are usually encoded by the MenFDHBCE operon [274], which was absent in quality filtered *Ca.* E. dacicola genomes. In particular, important missing reactions were: (1) isochorismate synthase (ICHORS, MenF, EC 5.4.4.2), (2) 2-succinyl-

5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (SEPHCHCS, MenD, EC 2.2.1.9), (3) 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase (SHCHCS3, MenH, EC 4.2.99.20), (4) o-succinylbenzoate synthase (SUCBZS, MenC, EC 4.2.1.113), (5) o-succinylbenzoate-CoA ligase (SUCBZL, MenE, EC 6.2.1.26), (6) naphthoate synthase (DHNCOAS, MenB, EC 4.1.3.36), (7) 1,4-dihydroxy-2-naphthoate polyprenyltransferase (DHNAOT4, MenA, EC 2.5.1.74). In contrast, the reaction 1,4-dihydroxy-2-naphthoyl-CoA hydrolase (DHNCOAT, MenI, EC 3.1.2.28) was present.

As the biosynthetic routes for the biomass precursors thiamine diphosphate, Kdo2-Lipid A, and menaquinol-8 were blocked likely due to biological issues, the corresponding metabolites were tested in subsequent analysis as growth factors to support the growth of *Ca.* E. dacicola on the CDM.

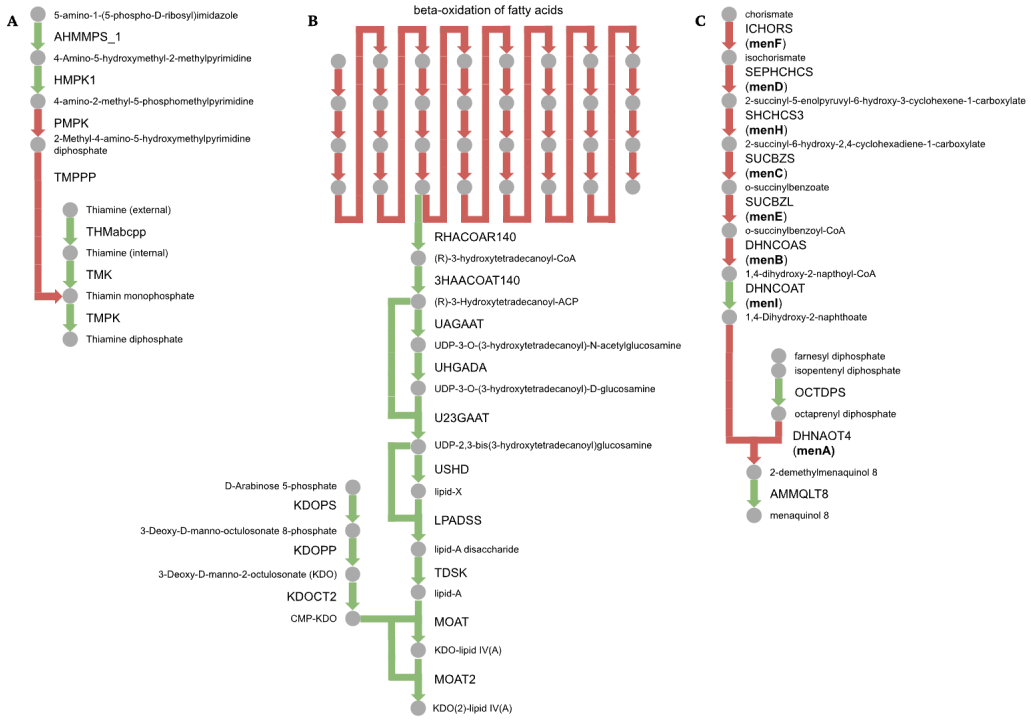

Figure 3.3. Biosynthetic routes leading to putative biomass precursors *Ca.* E. dacicola and *E. aphidicola*. All reactions (arrows) are present in *E. aphidicola*; green and red reactions are present or absent in *Ca.* E. dacicola, respectively. **A)** biosynthesis of thiamine diphosphate (thmpp). **B)** biosynthesis of Kdo2-Lipid A (kdo2lipid4). **C)** biosynthesis of menaquinol-8 (mql8). The MenFDHBCE operon and the MenI and MenA genes are annotated. Reaction IDs are reported in **Table 3.S4**.

3.2.3 Definition of an antibiotic for a selective medium

In order to identify antibiotics to be used as selective medium components, shared antibiotic resistance genes were searched among *Ca.* E. dacicola and *E. aphidicola* coding sequences. Then, three strains of *E. aphidicola* were used for phenotypic validation and assumed as a proxy for *Ca.* E. dacicola resistance patterns.

Among the antibiotic resistance genes in common between the two species, two were associated with ampicillin resistance, specifically *mrdA* and *ftsI*. These genes encode for peptidoglycan transpeptidases, also known as Penicillin Binding Proteins (PBPs), enzymes responsible for bacterial cell wall synthesis. *mrdA* and *ftsI*, encode mutated variants of PBP2 and PBP3, respectively, which exhibit low affinity for β-lactam antibiotics [275–278]. This genetic mechanism of resistance is consistent with existing literature indicating that the endosymbiont *Ca.* E. dacicola probably displays an intrinsic resistance to β-lactam antibiotics. Further, it was previously reported that the addition of β-lactam antibiotics (streptomycin, piperacillin, ampicillin, and rifampicin) to the diets of adult flies is ineffective in eliminating the endosymbiont from the digestive tracts of *B. oleae* [241,245,279].

Based on these findings, ampicillin was considered an useful selective agent to be added to a growth medium for *Ca.* E. dacicola, as it could suppress the growth of other microorganisms present in the esophageal bulbs, supporting the growth of the endosymbiont. Therefore, the MIC of ampicillin was evaluated on three strains of *E. aphidicola* to determine the concentration of ampicillin to be used in *Ca.* E. dacicola cultivation trials.

All strains exhibited resistance to the antibiotic concentrations used: ampicillin MIC of 32 mg/L, 64 mg/L, 128 mg/L and were found in *E. aphidicola* LMG 24877^{T}, LMG 26027 and LMG 5341, respectively. Based on these results, 32 mg/L of ampicillin was applied in the cultivation trials for *Ca.* E. dacicola.

3.2.4 Testing of medium recipes for *Ca*. E. dacicola

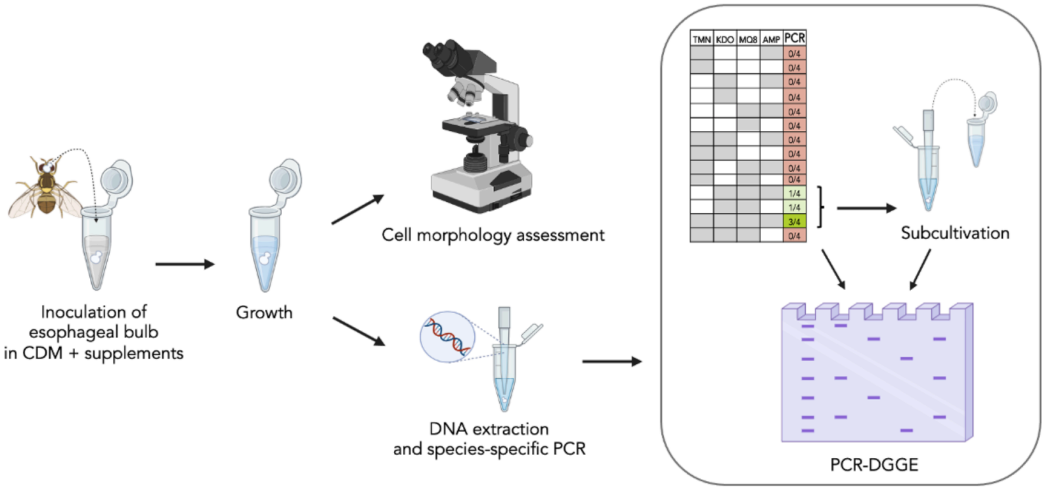

Figure 3.4. Experimental design of cultivation trials.

Based on results presented above, 14 different recipe variants based on the known *Erwinia*-specific CDM [280] (**Table 3.1**) were examined for cultivating *Ca.* E. dacicola (**Figure 3.4**). Specifically, all the combinations of the putative growth factors (thiamine, Kdo2-Lipid A, and menaquinol-8) were evaluated in presence or absence of ampicillin. Each recipe variant was tested in biological quadruplicate for a total of 56 growth trials, inoculating one dissected esophageal bulb per trial.

Table 3.1. Compound concentrations used in growth trials, taken from the *Erwinia*-specific CDM defined in [267] and considering the putative growth factors and antibiotic described in this work (underlined).

component	formula	g/mol	100 mL
glucose	$C_6H_{12}O_6$	180.16	0.3 g
monopotassium phosphate	KH_2PO_4	136.09	136 mg
dipotassium phosphate	K_2HPO_4	174.18	174 mg
magnesium sulphate heptahydrate	$MgSO_4\ 7H_20$	246.48	3 mg
aspartic acid	$C_4H_7NO_4$	133.10	0.28 g
boric acid	H_3BO_3	61.83	0.5 ug
calcium carbonate	$CaCO_3$	100.09	10 ug
copper sulfate pentahydrate	$CuSO_4\ 5H_2O$	249.69	1 ug
ammonium ferrous sulfate hexahydrate	$FeSO_4(NH_4)_2SO_4\ 6H_2O$	392.14	50 ug
potassium iodide	KI	166.00	1 ug
manganese sulfate monohydrate	$MnSO_4\ H_2O$	169.02	2 ug
molybdenum trioxide	MoO_3	143.94	1 ug
zinc sulfate heptahydrate	$ZnSO_4\ 7H_2O$	287.56	5 ug
thiamine hydrochloride	$C_{12}H_{17}ClN_4OS\ HCl$	337.27	0.1 mg
Kdo2-Lipid A	$C_{110}H_{198}N_2Na_4O_{39}P_2$	2326.7	0.1 mg
menaquinone-8	$C_{51}H_{72}O_2$	717.13	0.1 mg
ampicillin sodium salt	$C_{16}H_{18}N_3NaO_4S$	371.39	3.2 mg

Growth was observed under light microscope in 48 out of 56 trials (**Table 3.2** and **3.S3**) and, for each of the 48 growth-positive trials, a species-specific PCR was performed to detect presence of *Ca*. E. dacicola after careful sampling of the supernatant without moving the esophageal bulb. Moreover, as a control, the DNA extracted from the culture of the esophageal bulb inoculated in CDM without supplements was used.

Remarkably, the only combination of putative growth factors that was almost always positive to growth and always positive to the *Ca*. E. dacicola-specific PCR was that including thiamine, Kdo2-Lipid A and menaquinone-8, all together with ampicillin (complete medium). Additionally, the bacterium was sporadically detected also in the medium with Kdo2-Lipid A and menaquinone-8 (with or without ampicillin). This was also a confirmation that 32 ml/L of ampicillin is a suitable concentration for this experiment.

Table 3.2. Synthesis of the cultivation trials. "PCR": presence of the *Ca.* E. dacicola-specific band in species-specific PCR. See **Table 3.S3** for more details.

putative growth factors added to CDM [280]	with ampicillin			without ampicillin		
	trial series code	observed growth	PCR	trial series	observed growth	PCR
TMN	A	4/4	0/4	B	4/4	0/4
KDO	C	4/4	0/4	D	4/4	0/4
MQ8	E	1/4	0/1	F	4/4	0/4
TMN + KDO	G	3/4	0/3	H	4/4	0/4
TMN + MQ8	I	3/4	0/3	J	3/4	0/3
KDO + MQ8	K	3/4	**1/3**	L	4/4	**1/4**
TMN + KDO + MQ8	M	3/4	**3/3**	N	4/4	0/4
	total	21/28	4/21	total	27/28	1/27

Ca. E. dacicola-positive cultures were then sampled to perform subcultures using the complete medium. Unfortunately, species-specific PCR of subcultures failed to detect the presence of the endosymbiont (**Figure 3.5A**).

At this point, PCR-DGGE analysis was performed to evaluate the purity of cultures (**Figure 3.5B** and **Table 3.3**). Interestingly, *Ca*. E. dacicola-positive cultures and respective subcultures were associated with the presence of a light band (4B) in all samples, corresponding to *Ca.* E. dacicola and consistently with the content of the esophageal bulb used as positive control (band 4). However, other bands not related to *Ca.* E. dacicola were also observed. Sequencing of the bands revealed the presence of *Serratia* spp. in all samples (bands 1, 2 and 3), showing that the cultures were not pure. In addition, the presence of *Serratia* spp. and *Winslowiella iniecta* was detected (bands 16 and 17) in the culture derived from incubating the esophageal bulb in CDM without supplements.

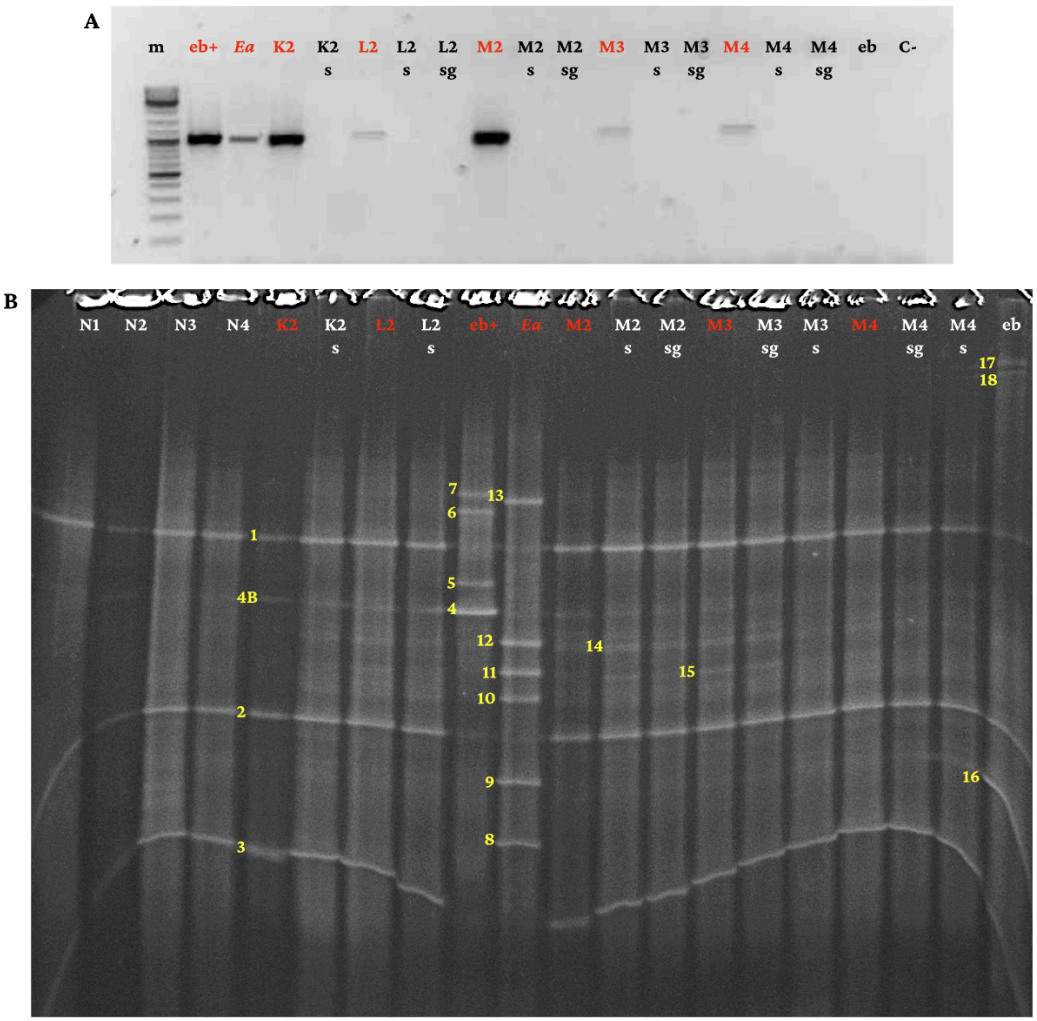

Figure 3.5. Molecular analysis. At the top of the lanes, sample codes are reported (see **Table 3.S3**). Samples positive to the *Ca.* E. dacicola-specific PCR are shown in red. Codes "eb+" and "eb" represent esophageal bulbs, the latter grown in CDM without supplements. Code "*Ea*" represents *E. aphidicola* LMG 24877^{T}. Samples marked with "s" and "sg" are subcultures from fresh culture or glycerol stock, respectively. **A)** *Ca.* E. dacicola-specific PCR (primers also amplify *E. aphidicola*). "m": ladder. "C-": PCR negative control. **B)** DGGE analysis of amplified 16S rRNA gene (PCR) products of DNA extracted from the cultures. Numbered bands in the gel represent those that were excised, re-amplified, and identified.

Table 3.3. Identification of bands excised from PCR-DGGE. "S%" and "V" are the sequence similarity and variation ratio compared to the reference, respectively. "n.a." indicates that band quality or sequence quality was too low for the identification.

band	most similar species	S%	V
1	*Serratia nematodiphila* *Serratia surfactanfaciens* *Serratia ureilytica*	99.28 99.28 99.28	1/139 1/139 1/139
2	*Serratia nematodiphila* *Serratia surfactantfaciens*	100 100	0/153 0/153
3	*Serratia ureilytica* *Serratia rubidaea*	100 100	0/138 0/138
4	*Candidatus* Erwinia dacicola (Erw_SC)	100	0/146
4B	n.a.		
5	*Candidatus* Erwinia dacicola (Erw_SC)	98.56	2/139
6	*Candidatus* Erwinia dacicola (Erw_SC)	100	0/139
7	*Candidatus* Erwinia dacicola (Erw_SC)	99.28	1/138
8	*Erwinia aphidicola*	99.07	1/107
9	n.a.	-	-
10	n.a.	98.55	2/138
11	*Erwinia aphidicola*	99.26	1/135
12	*Erwinia aphidicola*	100	0/139
13	*Erwinia aphidicola*	100	0/139
14	n.a.		
15	n.a.		
16	*Winslowiella iniecta*	100	0/139
17	*Winslowiella iniecta*	100	0/138
18	*Serratia ureilytica* *Serratia rubidaea*	98.29 98.29	2/117 2/117

3.3 Methods

3.3.1 Phylogenomics and metabolic modeling

Gempipe v1.37.3 (see **Chapter 2**) was used to download, filter and annotate genomes, as well as to reconstruct GSMMs. "gempipe recon" was launched on 5th September 2024 using parameters *-t 252393,68334,55211,55212 -b enterobacterales_odb10 --buscoM 4% --buscoF 2% --ncontigs 1000000 --N50 0 -s neg --sbml*. Briefly, all the available genome assemblies for *Ca.* E. dacicola (taxid 252393), *E. aphidicola* (taxid 68334), *E. persicina* (taxid 55211) and *E. rhapontici* (taxid 55212) were automatically retrieved from NCBI [281], and their coding sequences were predicted using Prokka v1.14.6 [282]. Then, coding sequences were passed to BUSCO v5.4.0 [283] to discard bad-quality genomes based on the presence of expected universal single-copy orthologs. Specifically, bad-quality was arbitrarily defined as >4% of missing or >2% of fragmented orthologs, on a total of 440 universal single-copy orthologs provided by the *enterobacterales_odb10* database. The additional metrics N50 and number of

contigs were computed using SeqKit v2.2.0 [202] but were not used for genome filtering.

A pan-genome analysis was performed on the quality-filtered genomes using Roary v3.13.0 [284] with options *-i 85 -s -e -n -cd 95*, giving in input the .gff annotations provided by Gempipe and internally generated by Prokka [282]. The requested multiple sequence alignment (MSA) of the core-genes (file "core_gene_alignment.aln") was used to create a phylogenomic tree using RAxML-NG v1.1.0 [285] with options *--model GTR+FO+G4m --all --tree pars{10} --bs-trees 100*, using GCA_014839105.1 as outgroup. The tree was visualized using custom code relying on the Phylo module of Biopython v1.80 [286] and matplotlib[16] v3.7.0. A venn diagram showing the intersection of species-specific core-genes was produced using matplotlib, computing the set of Roary-defined orthogroups always present in all the strains of the same species. In this process, the outlier strain ZSR3 was excluded when computing the set for *E. persicina*.

To create GSMMs, "gempipe recon" was used first. Briefly, coding sequences of the quality-filtered genomes were grouped into gene clusters based on aminoacidic sequence identity. During this process, a multi-step gene recovery was applied to cope with eventual issues arising during genome assembly or gene calling (**Figure 3.S5**). Once the unannotated genome-specific genes were recovered, they were introduced into their respective clusters. Next, the representative sequences of the gene clusters were aligned on the BiGG-derived [287] gene database used in CarveMe v1.5.2 [268], another state-of-art tool for the reconstruction of GSMMs. Reactions were copied from the CarveMe gram-negative reaction universe [268] based on alignment scores, applying strict rules on the generation of GPRs. This resulted in a reference-free reconstruction of a draft pan-GSMM, which models gene clusters instead of genes like in usual GSMMs, and could be seen as the collection of all the metabolic reactions encoded by the quality-filtered genomes in input. No automated gap-filling was applied during its reconstruction. The detailed implementation methods of Gempipe are described in **Chapter 2**.

The draft pan-GSMM was then manually curated using the Gempipe API, gap-filling for biomass production when growing on a *Erwinia*-specific CDM [267] (**Table 3.1**). The medium was represented by setting the exchange reactions as solute concentrations, following the best practices delineated by Marinos et al. [288] (**Table 3.S2**). Given the lack of knowledge on the biomass composition of *Erwinia*, the biomass definition provided by Gempipe was used,

[16] https://github.com/matplotlib/matplotlib

which is inherited from the gram-negative CarveMe [268] universal GSMM. Constrained metabolic reactions were relaxed and the absence of energy-generating cycles (EGCs) [289] was verified. All the growth simulations were performed via the flux-balance analysis (FBA) provided by the COBRApy v0.29.0 [290] package, with the objective function set to maximizing the biomass production. CPLEX v22.1.1 [291] was used as the linear optimization solver.

The gap-filled pan-GSMM was finally inputted into "gempipe derive" with options *--sbml --skipgf*, producing strain- and species-specific GSMMs with no automated gap-filling. Briefly, strain-specific GSMMs are produced making a copy of the pan-GSMM and subtracting the gene clusters with no member genes in a strain, leading to the eventual loss of reactions [292]. Species-specific GSMMs are defined by the set of reactions that appear in all the strain-specific GSMMs belonging to the same species. The metabolic gaps that prevented the *in silico* growth of the *Ca.* E. dacicola species-specific GSMM on the CDM were manually investigated using the Gempipe API, to detect possible growth factors to provide as medium supplements.

The general gap-filling approach adopted was the following. For the draft pan-GSMM, the function *perform_gapfilling* of the Gempipe API was called, specifying the ID of a blocked biomass precursor and the gram negative universal model as the source of potential gap-filling reactions. Then, the suggested reactions were contextualized on an *ad hoc* Escher [293] map showing all the possible connections encoded in the gram negative universal model. Suggested reactions were then identified on the KEGG Reaction [294] database, and their underlying KEGG Orthologs [272] (KO codes) were searched using the function *query_pam* of the Gempipe API. If no gene clusters were obtained, the same procedure was repeated querying for other attributes, such as the expected EC codes or PFAM domains [295]. If no candidate gene clusters were found, alternative pathways were considered using the KEGG maps as reference.

For the gap-filling of species-specific GSMM, the gap-filled pan-GSMM was used as the source of reactions in the *perform_gapfilling* function, and an *ad hoc* Escher [293] map was once again used to contextualize suggestions. Then, for each suggested reaction, the underlying gene clusters were extracted from the GPR in the pan-GSMM, and additional, unmodeled gene clusters with similar functional annotation were searched using the *search_similar* function of the Gempipe API.

3.3.2 Antimicrobial resistance prediction

The search for antibiotic resistance genes (ARGs) was conducted by analyzing the aminoacidic coding sequences of the type strain of *E. aphidicola* (LMG 24877^T, accession ASM2416951v1), *Ca.* E. dacicola IL (ASM175685v1) and Oroville (ASM305828v2). The identification of ARGs was carried out using BLASTp [50] along with the CARD-RGI v6.0.2 [296], ARG-ANNOT v6.0 [297], and ResFinder v2.1.0 [298] sequence databases. High scoring pairs (HSPs) were filtered for query coverage of at least 80% and sorted by percentage sequence identity. Top hits obtained with the three databases were manually integrated to obtain a consensus, and the resistance genes shared between *E. aphidicola* and *Ca.* E. dacicola were individually aligned on the NR (non-redundant) GenBank NCBI sequence database[17] for further confirmation.

3.3.3 MICs of antibiotic compound

The MICs of antibiotics were tested on three *E. aphidicola* collection strains (LMG 24877^T, LMG 26027, and LMG 5341), following a protocol described by Wiegand *et al*. with modifications [299]. The cells were grown on Nutrient agar plates (5 g/L peptone Sigma-Aldrich Cat#77199; 3 g/L meat extract Sigma-Aldrich Cat#70164-500G; 15 g/L agar Sigma-Aldrich Cat#05039) at pH 7 and incubated aerobically at 28 °C for 72 hours. After incubation, three distinct colonies from each strain were picked, inoculated into 40 mL of Nutrient broth and incubated at 28 °C for 72 hours. The resulting cells were used at a standardized inoculum for phenotypic tests.

The test was performed in triplicate for each strain, with the final cell count standardized to 5 x 10^5 colony-forming units (CFU)/mL. Inocula and antibiotic dilutions were prepared in 15 mL tubes and 100 µL of each was transferred to 96-well plates. Tested concentrations of ampicillin (Sigma-Aldrich Cat#A0166) ranged from 8 to 1024 mg/mL, in a series incremented by 8 mg/mL. Negative controls (ampicillin without cells) and positive controls (cells without ampicillin) were included for each replicate. Plates were incubated for 24 hours at 28 °C, and the optical density at 600 nm was determined using a spectrophotometer.

3.3.4 Origin of *B. oleae* and rearing condition

Olive fly pupae were collected in November 2023 from oil mills in the Veneto Region (Northern Italy) and routinely transferred to the laboratory within 24 hours. Emerged flies were reared in 30 × 30 × 30 cm³ net cages at a temperature

[17] https://blast.ncbi.nlm.nih.gov/Blast.cgi

of 23 ± 2 °C and a relative humidity below 60%. Both female and male flies were reared in the same cage and fed with a dry diet consisting of saccharose and yeast extract (4:1). Water was constantly available on a sponge wick and refreshed every 7 days.

3.3.5 Extraction of esophageal bulbs

B. oleae adults were collected using sterile tweezers and euthanized by placing them at -20 °C for approximately 10 min. The dissections were performed as described by Capuzzo *et al.* [236] with modifications, under a stereomicroscope in a laminar flow hood, utilizing a Bunsen burner. Each adult fly was placed in a drop of sterile saline solution (0.9% w/v NaCl Sigma-Aldrich Cat#S9888) on a sterilized slide and sterile pointed tweezers (No. 5) were used for the dissections. The specimens were decapitated, and the body was removed, leaving only the head. The head was then carefully opened, ensuring the internal organs were not damaged, to extract the esophageal bulb, which was subsequently inoculated into the culture media.

3.3.6 Cultivation trials

Extracted esophageal bulbs were inoculated into 500 uL of various culture media (**Table 3.2**) based on a CDM that supports the growth of free-living *Erwinia* species [267]. Standing on metabolic modeling results, thiamine, Kdo2-Lipid A and menaquinone-8 were hypothesized to promote growth of *Ca.* E. dacicola and thus cultivation trials were set up with the addition of these compounds to the CDM, individually and in combination. Based on the antibiotic resistance genes search results, each growth trial was tested both with and without ampicillin. For each of the 14 growth conditions, four biological replicates were performed, for a total of 56 esophageal bulbs inoculated. Two controls were used in each condition: (i) *E. aphidicola* and (ii) no inoculum.

To prepare the *Erwinia*-specific CDM [267] (**Table 3.1**) the following components were dissolved in 100 mL of distilled water: 3 mg of $MgSO_4 \cdot 7H_2O$ (Sigma-Aldrich Cat#M1880), 174 mg of K_2HPO_4 (Sigma-Aldrich Cat#P8281), 136 mg of KH_2PO_4 (Sigma-Aldrich Cat#P-0662), 280 mg of aspartic acid (Sigma-Aldrich Cat#11189), 0.5 µg of H_3BO_3 (Sigma-Aldrich Cat#1.00162), 10.0 µg of $CaCO_3$ (Fluka Analytical Cat#21069), 1.0 µg of $CuSO_4 \cdot 5H_2O$ (Sigma-Aldrich Cat#209198), 50.0 µg of $FeSO_4(NH_4)2SO_4 \cdot 6H_2O$ (AnalaR NORMAPUR® Cat#10022), 1.0 µg of KI (Sigma-Aldrich Cat#171211), 2.0 µg of $MnSO_4 \cdot H_2O$ (Sigma-Aldrich Cat#940178), 1.0 µg of MoO_3 (Sigma-Aldrich Cat#1.00400), 5.0 µg of $ZnSO_4 \cdot 7H_2O$ (Sigma-Aldrich Cat#Z0251); these were added to 300 mg of

glucose (Sigma-Aldrich Cat#G8270) (autoclaved separately) by filtration through 0.20 µm filters to ensure sterility. pH was adjusted to approximately 7.0 using NaOH.

Concentration of ampicillin sodium salt (Sigma-Aldrich Cat#A0166) was derived from MICs experiments and set to 3.2 mg / 100 mL. Concentration of putative growth factors was supported by literature and set to 0.1 mg / 100 mL for thiamine hydrochloride (Fluka Analytical Cat#95160) [300], Kdo2-Lipid A (AdipoGen Life Sciences Cat#AG-CU1-0001-M001) and menaquinone-8 (SiChem Cat#SC-1230) [301].

When growth was observed in a medium, 200 µL of the culture were stored in 25% glycerol at -80°C, 200 µL were used to collect the pellet for DNA extraction, and 10 µL were used for subcultivation on the same medium if the species-specific PCR test was positive (see below). The remaining volume was examined under a light microscope to assess cellular morphology.

3.3.7 Molecular detection of *Ca.* E dacicola

Genomic DNA was extracted from the cultures by first carefully pipetting 20 µL from the supernatant, ensuring that the esophageal bulb was not in the solution. In addition, the DNA extracted from the culture of the esophageal bulb inoculated in CDM without supplements was used as a control. The sample was then centrifuged for 10 minutes at 13,000 rpm, and the supernatant was subsequently removed. Then, 10 µL of sterile double-distilled water was added to the remaining pellet and the mixture was incubated at 95 °C for 10 minutes in a thermal cycler. Following incubation, the sample was centrifuged for 10 minutes at 13,000 rpm. The resulting supernatant was used as the DNA template for PCR.

DNA yield and purity were determined with a NanoDrop ND1000 UV–Vis Spectrophotometer (Thermo Scientific). The concentration of the DNA was standardized to 10 ng/µL for use in the PCR reactions.

To detect the presence of *Ca.* Erwinia dacicola, a multiplex PCR was performed amplifying the 16S rRNA gene. Universal and *Ca.* E. dacicola-specific primers were used to rapidly (i) check whether the DNA was amplifiable and (ii) detect the presence of the endosymbiont. The universal primers, designed by Baker *et al.* [302], were E8F (AGAGTTTGATCCTGGCTCAG) and E1541R (AAGGAGGTGA TCCANCCRCA), producing an amplicon of approximately 1500 bp. The species-specific primer, designed by Livadaras *et al.* [303], was DACF (GAAG GCGAAGAGGTTAATAACCTTTTT), and was used in conjunction with the

universal primer E1541R. Together, these two primers targeted a portion of the 16S rRNA gene specific to *Ca.* E. dacicola, producing an amplicon of approximately 1000 bp.

The amplification was performed in 20 µL (final volume) of reaction mixture containing 1 unit of GoTaq® G2 Flexi DNA Polymerase (Promega), 1× Green GoTaq® G2 Flexi buffer, 1.5 mM $MgCl_2$, 300 µM each deoxynucleoside triphosphate, 0.5 µM each primer, and 10 ng of genomic DNA. PCR parameters were 94 °C for 5 min, 35 cycles of 94 °C for 60 s, 59 °C for 30 s, 72 °C for 60 s, and a final extension at 72 °C for 5 min.

PCR products were run on a 2% w/v agarose gel in 1× TAE buffer (40 mM Tris Sigma-Aldrich Cat#T5941, 20 mM acetic acid Sigma-Aldrich Cat#a6283, and 0.4 mM EDTA Sigma-Aldrich Cat#E9884) stained with Atlas Clearsight (Bioatlas) at 110 V for 40 min. The gel was loaded with the molecular ladder O'Gene Ruler DNA (Thermo Scientific) to verify the correct amplification and to determine the presence of *Ca.* E. dacicola based on the presence of the targeted amplicon.

3.3.8 PCR-DGGE

DNA extracted from the cultures that resulted positive to *Ca.* E. dacicola-specific PCR, and from relative subcultures, was analyzed using PCR-DGGE to verify the culture purity. As controls, DNA extracted from the growth trial with the esophageal bulb inoculated in CDM without supplements, DNA extracted directly from the esophageal bulb, and DNA extracted from a culture of *E. aphidicola* LMG 24877^T were included.

Approximately 250 nucleotides of the 5'-end region of the 16S rRNA gene were amplified by PCR using universal primers HDA-1 (CGCCCGGGGCGCGCCCCG GGCGGGGCGGGGGCACGGGGGGACTCCTACGGGAGGCAGCAT) containing GC-clamp (underlined) and HDA-2 (GTATTACCGCGGCTGCTGGCAC), following the protocol of Walter *et al.* [304].

Separation of GC-clamped amplicons was carried out in a D-Code™ Universal Mutation Detection System (Bio-Rad). PCR products were loaded onto a 40–60% denaturing gel and run for 16 h at 50 V at a constant temperature of 60 °C. The gel was stained with a solution containing EuroSafe Nucleic Acid Stain (Euroclone) and bands were observed by UV transillumination (UVITEC Gel Documentation System, Cleaver Scientific).

Selected bands were aseptically excised with a sterile scalpel and placed in 15 µL of TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) and stored at -4 °C. Then, DNA

was eluted from the gel fragment and 5 µL of the suspension was used as template for reamplification using the HDA-1/HDA-2 primer pair without the GC-clamp. The PCR products were purified and sent for sequencing at Eurofins Genomics (Ebersberg, Germany). The sequencing data were searched against bacterial type strains 16S rRNA gene sequences in the EzBioCloud database (www.ezbiocloud.net, accessed on 22 April 2024).

3.4 Discussion

In the present study, a systems biology approach was implemented to enable laboratory cultures of the bacterium *Ca*. E. dacicola, main endosymbiont of *B. oleae* [239], a fly which develops inside the olive drupa (*Olea europaea*) causing economic losses worldwide [305]. Specifically, comparative genome-scale metabolic modeling was performed between *Ca.* E. dacicola and *E. aphidicola*, after the latter was shown to be the closest free-living species. Focusing the comparative modeling on the production of biomass on a known *Erwinia*-specific CDM [267], three putative growth factors were identified: thiamine, Kdo2-Lipid A, and menaquinol. Moreover, genome analysis suggested the presence of antibiotic resistance for ampicillin in both the species, which was evaluated *in vitro* using *E. aphidicola* as a proxy for *Ca.* E. dacicola.

Menaquinol is a reduced form of menaquinone, an essential cofactor in bacteria involved in the generation of ATP. Menaquinol transfers electrons to terminal acceptors, leading to the shuttling of protons across membranes, which is required by the ATP synthase complex to operate [306,307]. Menaquinone is known to transfer electrons during microaerobic growth conditions [308]. In *Ca*. E. dacicola, the biosynthetic pathway for menaquinone appeared deeply gapped, having lost the MenFDHBCE operon together with MenA [274] (**Figure 3.3C**). However, other cases of insect endosymbionts with impaired menaquinone biosynthesis are known in literature. *Ca.* Sulcia muelleri, for example, symbiont of sharpshooters (*Homalodisca vitripennis*) along with *Ca.* Baumannia cicadellinicola, have a complete ATP synthase complex while retaining only two genes of the menaquinone biosynthesis: MenA and MenG; in Ca. S. muelleri, the pathway is likely complemented by metabolites coming from *Ca.* Baumannia, the insect host, and/or the plant [309]. Moreover, the same two genes are retained in other strains of *Ca*. Sulcia associated with the keeled treehopper (*Entylia carinata*), but with signs of recent degradation [310]. Also, *Ca*. Liberibacter asiaticus, which uses psyllids (*Diaphorina citri*) to reach its final plant host (usually *Citrus*), does not harbor menaquinone nor ubiquinone biosynthetic genes, so it is believed to import quinones exogenously to fuel its respiratory chain in the microaerophilic conditions which characterize the phloem [311,312].

Another relevant example, not related to insects, is that of *Porphyromonas pasteri* KLE1280, a dental plaque bacteria that was unculturable *in vitro* unless a co-culture with *Escherichia coli* was established. Also in this organism, the entire biosynthetic pathway for menaquinone was lost except for MenA and MenG [301], which are reported to be scattered in the genome in contrast to the other biosynthetic genes that are grouped in the MenFDHBCE operon [274,301]. In addition to heme, the provisioning of menaquinone or 1,4-dihydroxy-2-naphthoate (substrate of the MenA-econded enzyme and probably secreted by *E. coli*) brought to the first axenic culture of *P. pasteri* KLE1280 [301]. Since *Ca.* E. dacicola seemed to have also lost MenA, it was decided to provide menaquinone directly.

Kdo2-Lipid A is an important membrane molecule of gram-negative bacteria, as it provides structural integrity of cell walls and it is required to anchor external lipopolysaccharides (LPS), which are important in host-bacteria recognition and interaction. Its main precursors are UDP-N-acetylglucosamine, CMP-Kdo, and 3-OH fatty acids [273]. Literature reports exist of bacterial symbionts of insects with impaired Kdo2-Lipid A biosynthetic activity. For example, strains of *Rickettsia* [313] can synthesize Kdo2-Lipid A, including the 3-OH fatty acid precursors, but not N-acetylglucosamine-1-phosphate nor D-ribose-5-phopshate, which are predicted to be imported from the host in order to produce the Kdo2-Lipid A precursors UDP-N-acetylglucosamine and CMP-Kdo, respectively [257]. A somewhat opposite situation occurs in some lineages *Pantoea*, symbionts of the *Edessa* spp. stink bugs, where the production of UDP-N-acetylglucosamine is always enabled, but the metabolite cannot be used to produced Lipid A due to metabolic gaps, so the synthesis of Kdo2-Lipid A and LPS is prevented; authors hypothesized that LPS are not needed for an extracellular symbiont like the *Edessa*-associated *Pantoea*, while the sole production of peptidoglycan (which originates from the same precursor UDP-N-acetylglucosamine) should be sufficient to guarantee a minimal cell wall integrity [314]. In *Ca.* E. dacicola, the biosynthesis of Kdo2-Lipid A appeared complete. However, while the relative GSMM could produce UDP-N-acetylglucosamine and CMP-Kdo, the synthesis of 3-OH fatty acids was blocked. According to the CarveMe universal model [268] used as reference, the 3-OH fatty acids inserted in Kdo2-Lipid A should derive from the fatty acid beta-oxidation pathway, which appeared completely lost in *Ca.* E. dacicola (**Figure 3.3B**). Examples of insect-associated bacteria lacking beta-oxidation are also reported in literature. *Ca.* Portiera aleyrodidarum, obligate endosymbiont of whiteflies (*Bemisia tabaci*), is known to lack the beta-oxidation of fatty acids [262]. Moreover, some *Buchnera* strains, common endosymbionts of aphids (*Acyrthosiphon pisum*), are known to lack beta-oxidation and even

fatty acid biosynthesis, likely showing structural cell fragility while acquiring membrane phospholipids from the host [315]. Since 3-OH fatty acids vary in the structure of Kdo2-Lipid A [273], it was decided to directly provide Kdo2-Lipid A in the variant found in *E. coli* K12 during the cultivation trials of *Ca.* E. dacicola.

Thiamine is an important cofactor in every domain of life, which likely emerged during the shift from anaerobic to aerobic metabolism in early evolutionary history, and which is involved in many key pathways including glycolysis, pentose phosphate cycle and tricarboxylic acid cycle [316]. *Ca.* E. dacicola appeared incapable of synthesizing thiamine due to the absence of important biosynthetic genes, including thiE (EC 2.5.1.3) and thiD (EC 2.7.4.7), while its genome seemed to encode a dedicated transporter (**Figure 3.3A**). In literature, other examples of insect-associated bacteria with thiamine auxotrophy are known. For instance, several strains of *Arsenophonus*, symbiont of the louse flies (Hippoboscidae family), present gaps in the thiamine biosynthetic pathway [317]; *Rickettsia* cannot synthesize thiamine nor its phosphate derivatives [257]; *Ca.* Ishikawaella capsulata and *Ca.* Tachikawaea gelatinosa, symbionts of *Edessa* spp., are unable to produce thiamine [314]. A relevant example by similarity is that of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies (Glossinidae family). This organism has many genes of the thiamin biosynthetic pathway that are missing or mutated into pseudogenes, including thiE and thiD. However, like in *Ca.* E. dacicola, a transporter is present in *S. glossinidius* to uptake the thiamine released by the primary endosymbiont *Wigglesworthia glossinidia* [318]. In order to cultivate *Ca.* E. dacicola, given the presence of a dedicated transporter, thiamine was added directly to the CDM.

Among all the possible combinations (**Table 3.2**), the inclusion into the *Erwinia*-specific CDM of all the three putative growth factors identified (menaquinone, Kdo2-Lipid A, and thiamine), along with ampicillin, proved to be successful in sustaining cultures of *Ca.* E. dacicola after the immersion of an esophageal bulb in solution. However, the same medium recipe was not sufficient for subculturing, as *Ca.* E. dacicola was not detected through common PCR with dedicated primers. Intriguingly, applying PCR-DGGE one the same subcultures, light bands for *Ca.* E. dacicola were reported.

PCR-DGGE also reported the presence of other bacteria being in co-culture with *Ca.* E. dacicola, including *Serratia* spp. A strain of this genus, most similar to *Serratia marcescens* subsp. *sakuensis*, was already isolated from the digestive tract of *B. oleae* [319] and other *Bactrocera* species, including *B. cacuminata*, *B. tryoni* [320] and *B. dorsalis* [321–323]. Moreover, olives are a known reservoir for this genus [324], so the detection of *Serratia* here could be expected. Interestingly, *Serratia* has been indicated as pathogenic for *B. oleae* when present in high

viable amounts, potentially replacing beneficial microorganisms in the gut or producing toxic molecules [319]. In future attempts of isolating *Ca.* E. dacicola, ciprofloxacin could be introduced as an additional antibiotic to the recipe here used, as genome analysis suggested that *Ca.* E. dacicola is resistant also to ciprofloxacin (data not shown), while *Serratia* is not [325].

Metabolic models used in this study were focused on a generic gram-negative biomass assembly reaction [44], as biomass composition data for close *Erwinia* species were not available. However, it should be noted that the adoption of an alternative biomass equation could have led to the identification of different growth factors. Indeed, the failure in subculturing could be due to additional molecules required by *Ca.* E. dacicola for growing and not captured by this generic biomass definition, possibly released by the esophageal bulb (which is not included in subcultures). In addition, molecules that are not directly channeled to biomass, but elicit its formation by transcriptional regulation, are outside the scope of application of this kind of metabolic models. However, the eventual presence of additional growth factors could be addressed in future studies by means of a metabolomic analysis of the bulb.

While other studies already used GSMMs to characterize the metabolism of insect endosymbionts, and in particular the metabolic gaps in relation to hosts, here the modeling focus was shifted from the nutritive cross-feeding to the *in vitro* production of bacterial biomass, which requires the definition of a growth medium recipe. To this aim, even "draft" GSMMs produced by automated tools [253] can still be useful to start interpreting the metabolism of an organism [326], as was demonstrated here by isolating putative growth factors.

However, by design, GSMMs are based on a single genome, and thus they are strain-specific. Unfortunately, it is hard to know "a priori" which genome assembly, among those available, will be the most similar to the endosymbiont extracted from fresh fly samples. Working with a typical (strain-specific) GSMM could bias predictions without providing hints of general applicability, so there was the need to manage strain variability. Here, species-specific GSMMs were used, defined by the set of reactions that are always present in all the strain-specific GSMMs belonging to a species. This kind of model is useful not only to get the core set of metabolic reactions that distinguish one species from another; it also addresses the strain variability (at least in terms of reactome), particularly relevant in the challenge of culturing a yet-uncultured bacteria disposing only of metagenome-assembled genomes (MAGs).

In the case of *Ca.* E. dacicola, however, the number of available MAGs was scarce, with only 6 available assemblies (**Table 3.S1**), characterized by a rather

variable cumulative length and gene content (**Figure 3.S1A** and **B**, respectively). While an higher number of genomes would have allowed a definition of core genes (and thus core metabolic reactions) based on their frequency of occurrence, here it was opted to arbitrarily filter genomes based on the expected number of single-copy orthologs (BUSCO orthologs) [283,327], and to classify a reaction as belonging to the species core when encoded by both the two remaining MAGs (GCA_003058285.2: "Oroville" assembly; GCA_001756855.1: "IL" assembly). While the thresholds selection was opinionated, it was justified by the assumption that *Ca.* E. dacicola represents a relatively recent endosymbiotic lineage, with the phylogenomically closest species *E. aphidicola* being free-living. Therefore, with a gene decay expected to be still at the beginning of the symbiosis history [230], the number of expected BUSCO orthologs was assumed to be not far from that of *E. aphidicola*.

This assumption, however, implicitly presupposes that genome assemblies are not contaminated with exogenous sequences. Indeed, there is the chance that MAGs here considered of good quality (i.e., passing the BUSCO thresholds) are actually contaminated with sequences from other species, for example those co-occurring in the esophageal bulb. In this case, the number of metabolic gaps of *Ca.* E. dacicola would have been higher, with pathway likely complemented by other co-symbionts, as already demonstrated in other insect symbiotic relationships [259,260,263]. For this reason, future studies focusing on *Ca.* E. dacicola cultivation would benefit from better genomes, at least in terms of contiguity, like for example those produced by nanopore-based systems [328].

Gempipe, the GSMM reconstruction tool here used, provides extensive procedures of gene recovery to cope with possible errors introduced during the genome assembling or gene calling. As *Ca.* E. dacicola is expected to contain pseudogenes [230], concerns may arise regarding the use of the gene recovery feature, as recovered genes may actually be pseudogenes with loss of function. However, the number of recovered genes was low and in line (**Figure 3.S5**) with that of the other free-living species here considered (*E. aphidicola*, *E. persicina* and *E. rhapontici*). In addition, the gene recovery feature of Gempipe accounts for the eventual presence of premature stop codons, which is one of the main causes of pseudogenization [329]. Moreover, to be sure that gene recovery was not affecting results, *in silico* analysis were repeated using models built running Gempipe with the *--norec* option, which bypass the gene recovery; in these conditions, no additional putative growth factors were identified and conclusions remained identical.

The metabolic modeling-based identification of putative growth factors presented here was based on a manual curation of the GSMMs produced.

However, since GSMMs were applied to identify growth requirements, the manual curation was exclusively focused on metabolic pathways leading to blocked biomass precursors. This means that other metabolic gaps, left by the automated reconstruction tool, are likely present. Filling all these technical gaps would highlight just those of biological origin, representing the foundation of the symbiotic relationship and showing the progression of the symbiotic lifestyle [230]. However, to close these gaps and thus improve the quality of the models, not only better genomes are needed, but also better modeling databases. In Gempipe, as well as in CarveMe [268], the reference database for metabolic genes is BiGG [287,330]. This database actually is a collection of manually curated GSMMs, from which a catalog of modeled reactions and genes is derived. The bacterial models contained in the current version of BiGG (v1.6), not belonging to the *Escherichia* / *Shigella* species complex, are just 22. Therefore, the covered biodiversity of gene sequences is rather reduced, and thus reconstructions of non-model organisms, when based on the BiGG genes, will likely be characterized by several metabolic gaps of technical (non biological) origin, that should be manually curated.

To conclude, the present work proposed a system biology approach to define a growth medium recipe for cultivating the insect endosymbiont *Ca.* E. dacicola. The approach started with a phylogenomic analysis for the identification of the closest free-living species, which was then utilized as a reference for all the dowstram analysis, in particular (i) to drive comparative metabolic modeling for the identification of growth factors, and (ii) to identify common antibiotic resistance for selective growth, experimentally evaluated on the free-living as a proxy for the endosymbiont. This approach was thought to be effective, given the relatively recent endosymbiosis event in the evolutionary history of the genus *Erwinia*, which resulted in *Ca.* E. dacicola sharing much of its genome with that of the free-living *E. aphidicola*; indeed, it remained relatively large compared to genomes of other insect endosymbionts, including those within the same genus, such as the aphid-associated *Ca.* E. haradaeae, whose genome is already much more restricted (~1.1 Mbp) [233].

While the biological role of *Ca.* E. dacicola in *B. oleae* is still debated [4,240], this endosymbiont is known to be strictly required for the larval development in green, unripe olives [239]. In a context where *B. oleae* is increasingly widespread and European regulations are focused on reducing chemical pesticides, the impairing of this symbiotic relation seems a promising principle for pest control. Indeed, bioactive molecules that inhibit or disturb a bacteria-host interactions, critical for the fitness of the insect, have been already described as “symbiocides” [243]. However, in order to develop effective symbiocides, and to study the biology of *Ca*. E. dacicola in general, laboratory culture of this

endosymbiont would be desired. In the present work, even if a stable culture or co-culture of *Ca.* E. dacicola was not obtained, a new step forward has been made in that direction. Moreover, the approach proposed here is generalizable to other insect endosymbionts and to other fastidious bacteria in general, and provides a useful basis where to start from.

3.5 Supplementary Material

3.5.1 Supplementary Results

Blocked biomass precursors likely due to technical issues

The biosynthetic pathway for the tetrahydrofolate resulted complete in core-Eda. Anyway, the biosynthesis of this vitamin involves the production of glycolaldehyde (gcald), which resulted as a dead-end metabolite in core-Eda, as it could not be consumed. In core-Eap, glycolaldehyde was consumed by the 4-hydroxy-L-threonine aldolase (4HTHRA), encoded by several gene clusters with no members among the input *Ca.* E. dacicola genomes. As other ways to consume glycolaldehyde were not available based on functional annotation of gene clusters, the presence of a transporter was hypothesized, transferring glycolaldehyde outside the endosymbiont cell.

The palmitoyl-(acyl-carrier protein) (palmACP) is defined in the CarveMe gram negative universe [268] as one of the fundamental precursors in the phosphatidylethanolamine biosynthesis. This precursor is produced by the long-chain fatty acid:acyl-carrier protein ligase (AACPS3), loading a palmitic acid (C16:0) on an acyl-carrier protein (ACP), a reaction that missed from core-Eda preventing the phosphatidylethanolamine production. Even if the KEGG orthologs [272] (K01909 / K05939) for this reaction (R01406) were not available, since (i) core-Eda resulted with complete fatty acids and phosphatidylethanolamine synthetic pathways, and (ii) the biosynthetic gene for ACP was present in quality-filtered *Ca.* E. dacicola genomes, then it was decided to introduce AACPS3 without GPR.

3.5.2 Supplementary Figures

Figure 3.S1. Metrics and quality filtering of genomes. **A)** Quality filtering of genome assemblies based on the percentage of expected universal single-copy orthologs in the Enterobacterales order, as defined in BUSCO [283] ("complete": present and complete orthologs; "fragmented": present but fragmented orthologs; "missing": missing orthologs). Bars belonging to discarded assemblies are represented with muted colors. **B)** Cumulative length of the input genome assemblies. Bars belonging to discarded assemblies are represented with muted colors. **C)** Number of contigs of the input genome assemblies. Bars belonging to discarded assemblies are represented with muted colors.

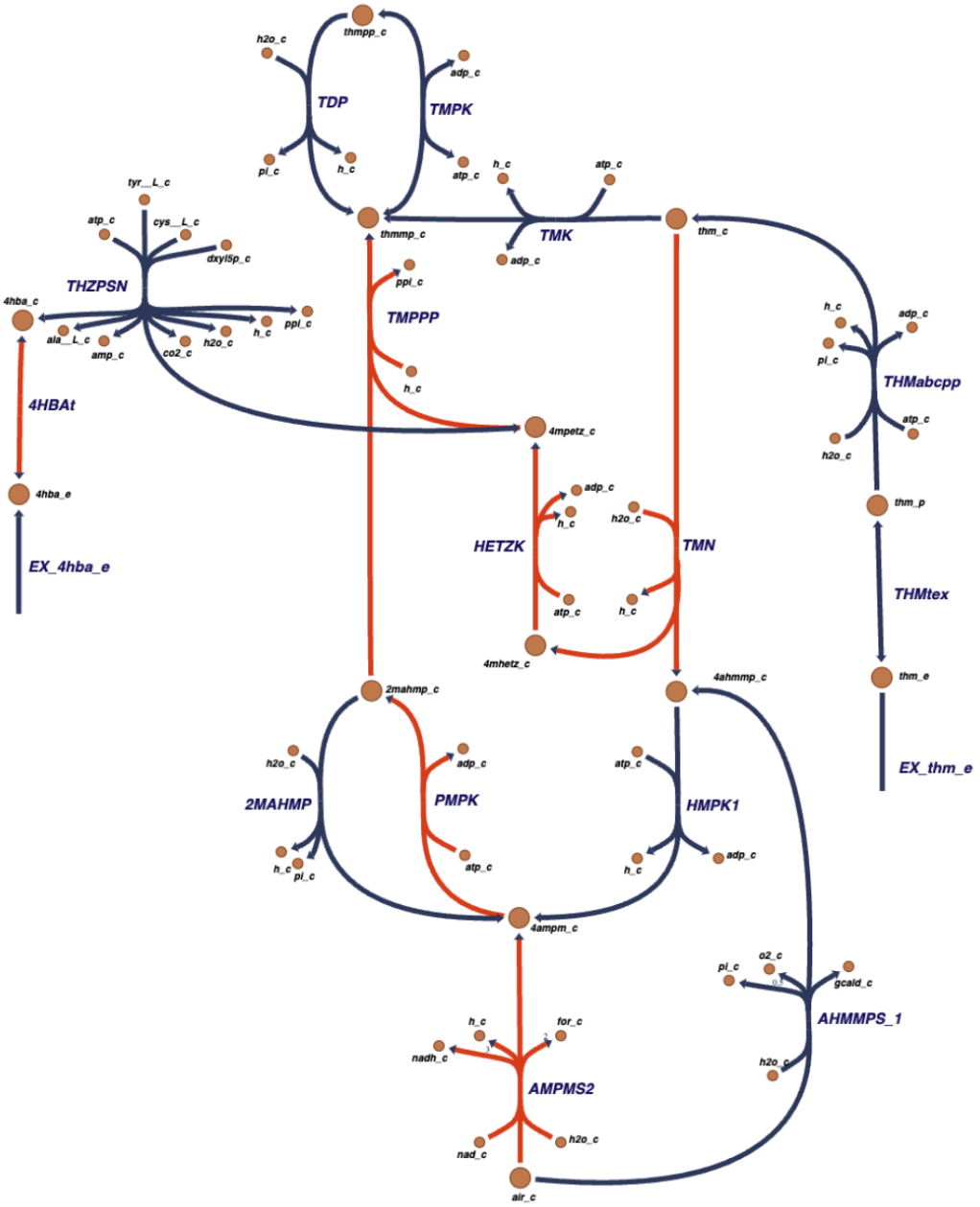

Figure 3.S2. Escher [293] map for the biosynthesis of thiamine diphosphate (thmpp). All the represented reactions are present in core_Eap. Red reactions are missing from core_Eda. IDs are reported in **Table 3.S4**.

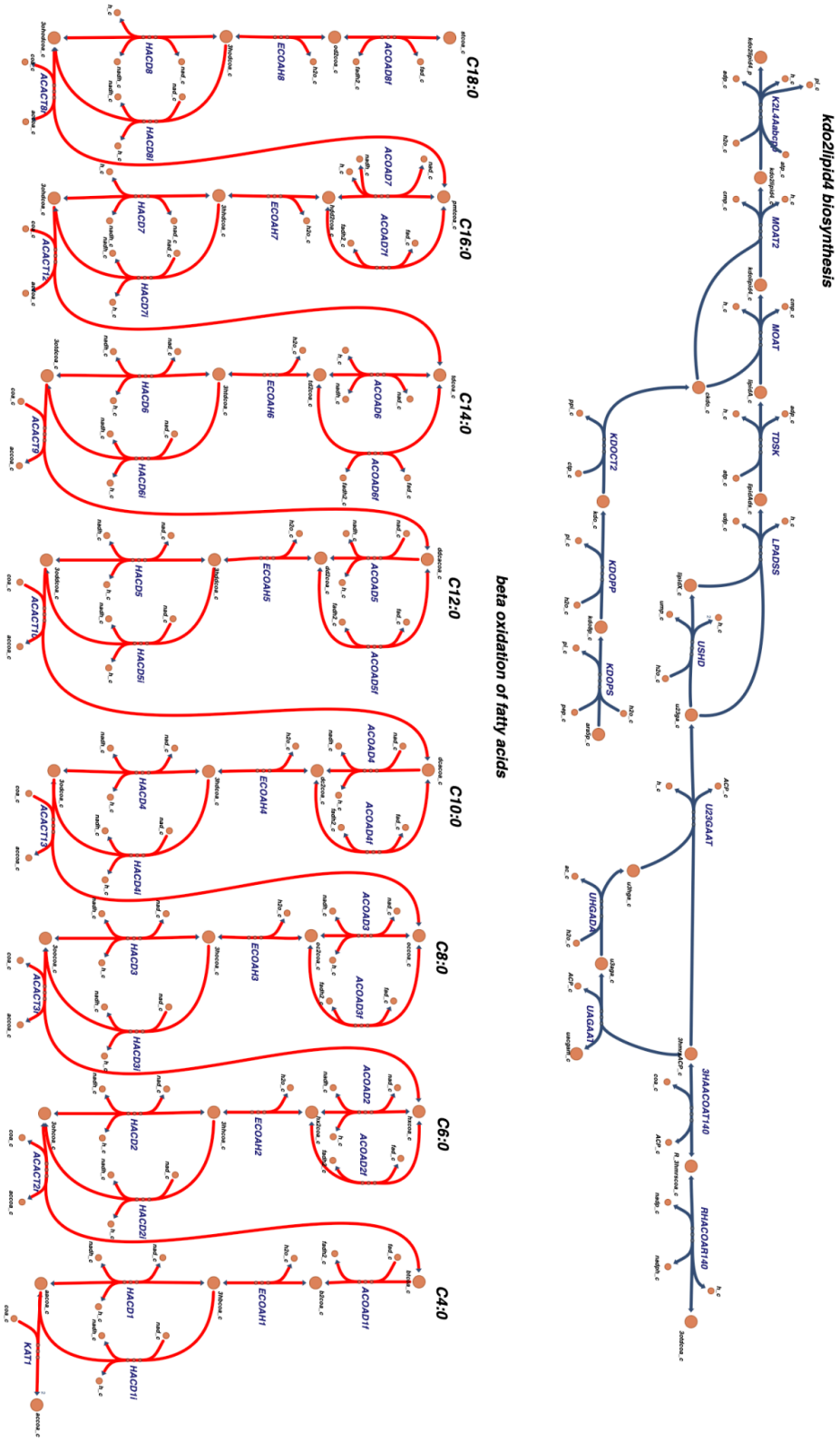

Figure 3.S3. Escher [293] map for the biosynthesis of Kdo2-Lipid A (kdo2lipid4). All the represented reactions are present in core_Eap. Red reactions are missing from core_Eda. IDs are reported in **Table 3.S4**.

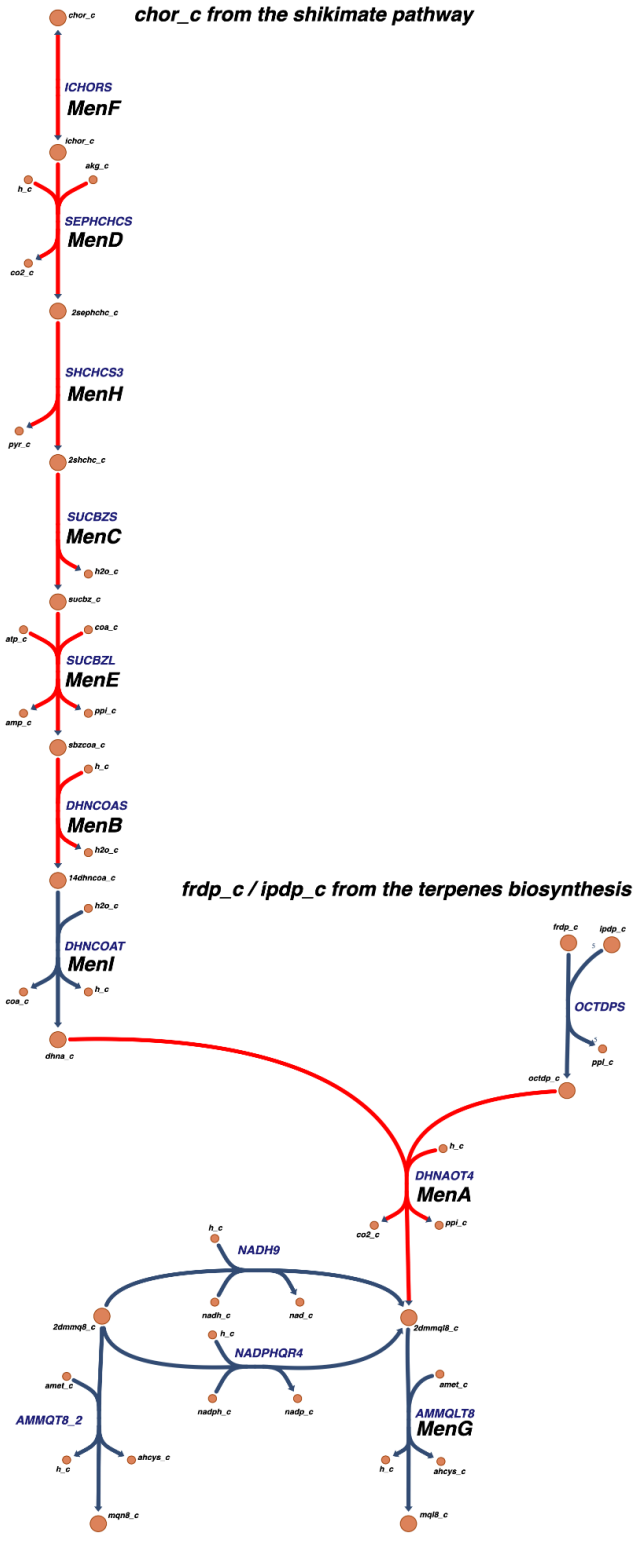

Figure 3.S4. Escher [293] map for the biosynthesis of menaquinol-8 (mql8). All the represented reactions are present in core_Eap. Red reactions are missing from core_Eda. The MenFDHBCE operon and the MenI, MenA and MenG genes are annotated. IDs are reported in **Table 3.S4**.

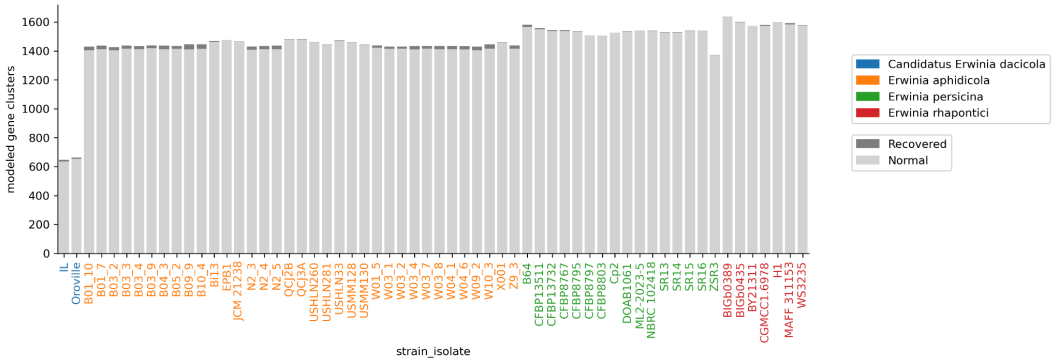

Figure 3.S5. Number of recovered gene clusters per quality-filtered genome, consequence of the strain-specific gene recovery feature of Gempipe.

3.5.3 Supplementary Tables

Table 3.S1. Genome assemblies used in this study.

	accession	taxid	species	strain / isolate
1	GCA_003510065.1	55211	*Erwinia persicina*	UBA9200
2	GCA_003511585.1	55211	*Erwinia persicina*	UBA10743
3	GCA_003507915.1	55211	*Erwinia persicina*	UBA9199
4	GCA_003509515.1	55211	*Erwinia persicina*	UBA9202
5	GCA_003510685.1	55211	*Erwinia persicina*	UBA9201
6	GCA_005233475.1	55211	*Erwinia persicina*	CFBP13511
7	GCA_003485445.1	55211	*Erwinia persicina*	B64
8	GCA_003507845.1	55211	*Erwinia persicina*	UBA9197
9	GCA_003507935.1	55211	*Erwinia persicina*	UBA9198
10	GCA_014839305.1	55211	*Erwinia persicina*	CFBP13732
11	GCA_014838825.1	55211	*Erwinia persicina*	CFBP8797
12	GCA_014839105.1	55211	*Erwinia persicina*	CFBP8803
13	GCA_014357465.1	55211	*Erwinia persicina*	DOAB1061
14	GCA_014838835.1	55211	*Erwinia persicina*	CFBP8795
15	GCA_014839215.1	55211	*Erwinia persicina*	CFBP8767
16	GCA_019844095.1	55211	*Erwinia persicina*	Cp2
17	GCA_024168495.1	55211	*Erwinia persicina*	ZSR3
18	GCA_024436235.1	55211	*Erwinia persicina*	SR16
19	GCA_024436265.1	55211	*Erwinia persicina*	SR14
20	GCA_024397315.1	55211	*Erwinia persicina*	SR15
21	GCA_024436315.1	55211	*Erwinia persicina*	SR13
22	GCA_001571305.1	55211	*Erwinia persicina*	NBRC 102418
23	GCA_037081855.1	55211	*Erwinia persicina*	ML2-2023-5
24	GCA_004364855.1	55212	*Erwinia rhapontici*	BIGb0435
25	GCA_009846845.1	55212	*Erwinia rhapontici*	CGMCC1.6978
26	GCA_012271765.1	55212	*Erwinia rhapontici*	H1
27	GCA_018326325.1	55212	*Erwinia rhapontici*	MAFF 311154
28	GCA_017875455.1	55212	*Erwinia rhapontici*	WS3235
29	GCA_018326345.1	55212	*Erwinia rhapontici*	MAFF 311155
30	GCA_018409035.1	55212	*Erwinia rhapontici*	MAFF 311153
31	GCA_020683125.1	55212	*Erwinia rhapontici*	BY21311
32	GCA_024807835.1	55212	*Erwinia rhapontici*	BIGb0389
33	GCA_014773485.1	68334	*Erwinia aphidicola*	JCM 21238
34	GCA_016925695.1	68334	*Erwinia aphidicola*	18B1
35	GCA_918698235.1	68334	*Erwinia aphidicola*	Bi13
36	GCA_024169515.1	68334	*Erwinia aphidicola*	X001 (24877T)

37	GCA_036865825.1	68334	*Erwinia aphidicola*	W06_1
38	GCA_036865845.1	68334	*Erwinia aphidicola*	W04_7
39	GCA_036865865.1	68334	*Erwinia aphidicola*	W04_5
40	GCA_036865915.1	68334	*Erwinia aphidicola*	W04_1
41	GCA_036865965.1	68334	*Erwinia aphidicola*	W03_4
42	GCA_036866015.1	68334	*Erwinia aphidicola*	W01_5
43	GCA_036866065.1	68334	*Erwinia aphidicola*	N2_2
44	GCA_036866115.1	68334	*Erwinia aphidicola*	B10_3
45	GCA_036866165.1	68334	*Erwinia aphidicola*	B09_4
46	GCA_036866175.1	68334	*Erwinia aphidicola*	B09_3
47	GCA_036866215.1	68334	*Erwinia aphidicola*	B06_8
48	GCA_036866225.1	68334	*Erwinia aphidicola*	B06_6
49	GCA_036866265.1	68334	*Erwinia aphidicola*	B06_2
50	GCA_036866275.1	68334	*Erwinia aphidicola*	B06_1
51	GCA_036866315.1	68334	*Erwinia aphidicola*	B03_8
52	GCA_036866365.1	68334	*Erwinia aphidicola*	B03_3
53	GCA_036866415.1	68334	*Erwinia aphidicola*	B01_5
54	GCA_036865815.1	68334	*Erwinia aphidicola*	W06_10
55	GCA_036865835.1	68334	*Erwinia aphidicola*	W04_9
56	GCA_036865905.1	68334	*Erwinia aphidicola*	W04_10
57	GCA_036865955.1	68334	*Erwinia aphidicola*	W03_5
58	GCA_036866005.1	68334	*Erwinia aphidicola*	W03_1
59	GCA_036866055.1	68334	*Erwinia aphidicola*	N2_3
60	GCA_036866105.1	68334	*Erwinia aphidicola*	B10_4
61	GCA_036866155.1	68334	*Erwinia aphidicola*	B09_7
62	GCA_036866205.1	68334	*Erwinia aphidicola*	B07_5
63	GCA_036866255.1	68334	*Erwinia aphidicola*	B06_3
64	GCA_036866295.1	68334	*Erwinia aphidicola*	B04_3
65	GCA_036866305.1	68334	*Erwinia aphidicola*	B03_9
66	GCA_036866345.1	68334	*Erwinia aphidicola*	B03_5
67	GCA_036866355.1	68334	*Erwinia aphidicola*	B03_4
68	GCA_036866395.1	68334	*Erwinia aphidicola*	B01_8
69	GCA_036866405.1	68334	*Erwinia aphidicola*	B01_7
70	GCA_036865725.1	68334	*Erwinia aphidicola*	W09_4
71	GCA_036865735.1	68334	*Erwinia aphidicola*	W09_3
72	GCA_036865795.1	68334	*Erwinia aphidicola*	W06_3
73	GCA_031981505.1	68334	*Erwinia aphidicola*	CTOTU47228
74	GCA_037149315.1	68334	*Erwinia aphidicola*	USMM130
75	GCA_037149855.1	68334	*Erwinia aphidicola*	USHLN260
76	GCA_036865875.1	68334	*Erwinia aphidicola*	W04_4
77	GCA_036865925.1	68334	*Erwinia aphidicola*	W03_8
78	GCA_036865975.1	68334	*Erwinia aphidicola*	W03_3
79	GCA_036866025.1	68334	*Erwinia aphidicola*	N2_6
80	GCA_036866035.1	68334	*Erwinia aphidicola*	N2_5
81	GCA_036866075.1	68334	*Erwinia aphidicola*	N2_1
82	GCA_036866085.1	68334	*Erwinia aphidicola*	B10_7
83	GCA_036866125.1	68334	*Erwinia aphidicola*	B10_2
84	GCA_036866135.1	68334	*Erwinia aphidicola*	B09_9
85	GCA_036866185.1	68334	*Erwinia aphidicola*	B09_2
86	GCA_036866235.1	68334	*Erwinia aphidicola*	B06_5
87	GCA_036866325.1	68334	*Erwinia aphidicola*	B03_7
88	GCA_036866375.1	68334	*Erwinia aphidicola*	B03_2
89	GCA_036866425.1	68334	*Erwinia aphidicola*	B01_4
90	GCA_037149355.1	68334	*Erwinia aphidicola*	USMM128
91	GCA_036865695.1	68334	*Erwinia aphidicola*	W09_7
92	GCA_036865755.1	68334	*Erwinia aphidicola*	W06_9
93	GCA_036865775.1	68334	*Erwinia aphidicola*	W06_7
94	GCA_036865805.1	68334	*Erwinia aphidicola*	W06_2
95	GCA_036865565.1	68334	*Erwinia aphidicola*	Z9_3
96	GCA_036865605.1	68334	*Erwinia aphidicola*	W10_8

97	GCA_036865635.1	68334	*Erwinia aphidicola*	W10_5
98	GCA_036865615.1	68334	*Erwinia aphidicola*	W10_7
99	GCA_036865645.1	68334	*Erwinia aphidicola*	W10_4
100	GCA_036865675.1	68334	*Erwinia aphidicola*	W10_1
101	GCA_036865715.1	68334	*Erwinia aphidicola*	W09_5
102	GCA_036865745.1	68334	*Erwinia aphidicola*	W09_2
103	GCA_036865765.1	68334	*Erwinia aphidicola*	W06_8
104	GCA_036865575.1	68334	*Erwinia aphidicola*	Z9_1
105	GCA_036865705.1	68334	*Erwinia aphidicola*	W09_6
106	GCA_036865785.1	68334	*Erwinia aphidicola*	W06_6
107	GCA_036865855.1	68334	*Erwinia aphidicola*	W04_6
108	GCA_036865885.1	68334	*Erwinia aphidicola*	W04_3
109	GCA_036865895.1	68334	*Erwinia aphidicola*	W04_2
110	GCA_036865935.1	68334	*Erwinia aphidicola*	W03_6
111	GCA_036865945.1	68334	*Erwinia aphidicola*	W03_7
112	GCA_036865985.1	68334	*Erwinia aphidicola*	W03_2
113	GCA_036865995.1	68334	*Erwinia aphidicola*	W03_10
114	GCA_036866045.1	68334	*Erwinia aphidicola*	N2_4
115	GCA_036866095.1	68334	*Erwinia aphidicola*	B10_5
116	GCA_036866145.1	68334	*Erwinia aphidicola*	B09_8
117	GCA_036866195.1	68334	*Erwinia aphidicola*	B09_1
118	GCA_036866245.1	68334	*Erwinia aphidicola*	B06_4
119	GCA_036866285.1	68334	*Erwinia aphidicola*	B05_2
120	GCA_036866335.1	68334	*Erwinia aphidicola*	B03_6
121	GCA_036866385.1	68334	*Erwinia aphidicola*	B03_10
122	GCA_036866435.1	68334	*Erwinia aphidicola*	B01_2
123	GCA_036866445.1	68334	*Erwinia aphidicola*	B01_10
124	GCA_036874005.1	68334	*Erwinia aphidicola*	W09_1
125	GCA_036865585.1	68334	*Erwinia aphidicola*	W11_1
126	GCA_036865595.1	68334	*Erwinia aphidicola*	W10_9
127	GCA_036865625.1	68334	*Erwinia aphidicola*	W10_6
128	GCA_036865655.1	68334	*Erwinia aphidicola*	W10_3
129	GCA_036865665.1	68334	*Erwinia aphidicola*	W10_10
130	GCA_036865685.1	68334	*Erwinia aphidicola*	W09_8
131	GCA_037045765.1	68334	*Erwinia aphidicola*	QCJ2B
132	GCA_037045825.1	68334	*Erwinia aphidicola*	EPB1
133	GCA_037045795.1	68334	*Erwinia aphidicola*	QCJ3A
134	GCA_037149555.1	68334	*Erwinia aphidicola*	USHLN281
135	GCA_037144385.1	68334	*Erwinia aphidicola*	USHLN33
136	GCA_001689725.1	252393	*Ca.* Erwinia dacicola	Erw_SC
137	GCA_003058285.2	252393	*Ca.* Erwinia dacicola	Oroville
138	GCA_012026695.1	252393	*Ca.* Erwinia dacicola	EdA2
139	GCA_012027475.1	252393	*Ca.* Erwinia dacicola	EdB2
140	GCA_937876675.1	252393	*Ca.* Erwinia dacicola	SRR3742170*
141	GCA_001756855.1	252393	*Ca.* Erwinia dacicola	IL

Table 3.S2. *In silico* representation of the chemically-defined medium for *Erwinia* [280]. Exchange reactions were set as the millimolar concentration of the dissolved species as reported in [288]. Some exchange reactions were not modeled ("/").

	component	formula	exchange reaction	mmol/L
1	glucose	$C_6H_{12}O_6$	EX_glc__D_e	16.6522347
2	potassium	K^+	EX_k_e	29.9730724
3	phosphate	HPO_4^{--}	EX_pi_e	19.9833463
4	magnesium	Mg^{++}	EX_mg2_e	0.12171373
5	sulfate	SO_4^{--}	EX_so4_e	0.12587115
6	borate	BO_3^{---}	/	/
7	calcium	Ca^{++}	EX_ca2_e	0.00099913

8	bicarbonate	HCO_3^-	/	/
9	copper	Cu^{++}	EX_cu2_e	0.00004005
10	iron 2+	Fe^{++}	EX_fe2_e	0.00127505
11	iron 3+	Fe^{+++}	EX_fe3_e	0.00127505
12	ammonia	NH_4^+	EX_nh4_e	0.00127505
13	iodide	I^-	/	/
14	manganese	Mn^{++}	EX_mn2_e	0.00011833
15	molibdate	Mo_4^{--}	/	/
16	zinc	Zn^{++}	EX_zn2_e	0.00017388
17	aspartate	$C_4H_6NO_4^-$	EX_asp__L_e	21.0363403

Table 3.S3. Details of the cultivation trials. "p.f.g.": putative growth factor added to the CDM; "ID": name of the growth trial; "o.g.": observed growth; "PCR": presence of the *Ca.* E. dacicola-specific band in species-specific PCR.

p.g.f.	with ampicillin				without ampicillin			
	ID	o.g.	morphology	PCR	ID	o.g.	morphology	PCR
TMN	A1	+	other morphologies (short rods)	no	B1	+	other morphologies (short rods)	no
	A2	+	mixed culture: rods and cocci	no	B2	+	other morphologies (short rods)	no
	A3	+	other morphologies (short rods)	no	B3	+	other morphologies (short rods)	no
	A4	+	other morphologies (short rods)	no	B4	+	other morphologies (short rods)	no
KDO	C1	+	other morphologies (short rods)	no	D1	+	other morphologies (short rods)	no
	C2	+	other morphologies (short rods)	no	D2	+	other morphologies (short rods)	no
	C3	+	other morphologies (short rods)	no	D3	+	other morphologies (short rods)	no
	C4	+	other morphologies (short rods)	no	D4	+	other morphologies (short rods)	no
MQ8	E1	-	-	-	F1	+	other morphologies (short rods)	no
	E2	+	other morphologies (short rods)	no	F2	+	other morphologies (short rods)	no
	E3	-	-	-	F3	+	mixed culture: rods and cocci	no
	E4	-	-	-	F4	+	mixed culture: rods and cocci	no
TMN + KDO	G1	-	-	-	H1	+	mixed culture: rods and short rods	no
	G2	+	mixed culture: rods and cocci	no	H2	+	mixed culture: rods and short rods	no
	G3	+	mixed culture: rods and cocci	no	H3	+	other morphologies (short rods)	no
	G4	+	mixed culture: rods and short rods	no	H4	+	other morphologies (short rods)	no
TMN + MQ8	I1	+	other morphologies (short rods)	no	J1	+	other morphologies (short rods)	no

	I2	-	-	-	J2	+	other morphologies (short rods)	no
	I3	+	other morphologies (short rods)	no	J3	+	other morphologies (short rods)	no
	I4	+	mixed culture: rods and short rods	no	J4	-	-	-
KDO + MQ8	K1	+	mixed culture: rods and short rods	no	L1	+	mixed culture: rods and short rods	no
	K2	+	mixed culture: rods and short rods	yes	L2	+	mixed culture: rods and short rods	yes
	K3	-	-	-	L3	+	mixed culture: rods and short rods	no
	K4	+	other morphologies (short rods)	no	L4	+	mixed culture: rods and short rods	no
TMN + KDO + MQ8	M1	-	-	-	N1	+	mixed culture: rods and short rods	no
	M2	+	mixed culture: rods and short rods	yes	N2	+	other morphologies (short rods)	no
	M3	+	mixed culture: rods and short rods	yes	N3	+	mixed culture: rods and short rods	no
	M4	+	mixed culture: rods and short rods	yes	N4	+	other morphologies (short rods)	no

Table 3.S4. Metabolite and reaction IDs used in figures containing metabolic maps (alphabetical order). “R”: reaction. “M”: metabolite.

	type	ID	name
1	R	2MAHMP	2-Methyl-4-amino-5-hydroxymethylpyrimidine diphosphatase
2	R	3HAACOAT140	3 Hydroxyacyl ACPCoA Transacylase
3	R	4HBAt	4-hydroxy-benzyl alcohol transport via diffusion
4	R	ACACT10	3-ketoacyl-CoA thiolase
5	R	ACACT12	3-ketoacyl-CoA thiolase
6	R	ACACT13	3-ketoacyl-CoA thiolase
7	R	ACACT2r	Acetyl-CoA C-acyltransferase (butanoyl-CoA)
8	R	ACACT3r	Acetyl-CoA C-acyltransferase (hexanoyl-CoA)
9	R	ACACT8r	Acetyl-CoA acyltransferase (hexadecanoyl-CoA)
10	R	ACACT9	3-ketoacyl-CoA thiolase
11	R	ACOAD1f	Acyl-CoA dehydrogenase (butanoyl-CoA)
12	R	ACOAD2	Acyl-CoA dehydrogenase (hexanoyl-CoA)
13	R	ACOAD2f	Acyl-CoA dehydrogenase (hexanoyl-CoA)
14	R	ACOAD3	Acyl-CoA dehydrogenase (octanoyl-CoA)
15	R	ACOAD3f	Acyl-CoA dehydrogenase (octanoyl-CoA)
16	R	ACOAD4	Acyl-CoA dehydrogenase (decanoyl-CoA)
17	R	ACOAD4f	Acyl-CoA dehydrogenase (decanoyl-CoA)
18	R	ACOAD5	Acyl-CoA dehydrogenase (dodecanoyl-CoA)
19	R	ACOAD5f	Acyl-CoA dehydrogenase (dodecanoyl-CoA)
20	R	ACOAD6	Acyl-CoA dehydrogenase (tetradecanoyl-CoA)
21	R	ACOAD6f	Acyl-CoA dehydrogenase (tetradecanoyl-CoA)
22	R	ACOAD7	Acyl-CoA dehydrogenase (hexadecanoyl-CoA)
23	R	ACOAD7f	Acyl-CoA dehydrogenase (hexadecanoyl-CoA)
24	R	ACOAD8f	Acyl-CoA dehydrogenase (octadecanoyl-CoA)

25	R	AHMMPS_1	4 amino 5 hydroxymethyl 2 methylpyrimidine synthetase
26	R	AMMQLT8	S-adenosylmethione:2-demthylmenaquinole methyltransferase (menaquinone 8)
27	R	AMMQT8_2	S-adenosylmethione:2-demethylmenaquinone methyltransferase
28	R	AMPMS2	4-amino-2-methyl-5-phosphomethylpyrimidine synthetase
29	R	DHNAOT4	1,4-dihydroxy-2-naphthoate octaprenyltransferase
30	R	DHNCOAS	1,4-dihydroxy-2-napthoyl-CoA synthase
31	R	DHNCOAT	1,4-dihydroxy-2-napthoyl-CoA thioesterase
32	R	ECOAH1	3-hydroxyacyl-CoA dehydratase (3-hydroxybutanoyl-CoA)
33	R	ECOAH2	3-hydroxyacyl-CoA dehydratase (3-hydroxyhexanoyl-CoA)
34	R	ECOAH3	3-hydroxyacyl-CoA dehydratase (3-hydroxyoctanoyl-CoA)
35	R	ECOAH4	3-hydroxyacyl-CoA dehydratase (3-hydroxydecanoyl-CoA)
36	R	ECOAH5	3-hydroxyacyl-CoA dehydratase (3-hydroxydodecanoyl-CoA)
37	R	ECOAH6	3-hydroxyacyl-CoA dehydratase (3-hydroxytetradecanoyl-CoA)
38	R	ECOAH7	3-hydroxyacyl-CoA dehydratase (3-hydroxyhexadecanoyl-CoA)
39	R	ECOAH8	3-hydroxyacyl-CoA dehydratase (3-hydroxyoctadecanoyl-CoA)
40	R	EX_4hba_e	Exchange reaction for 4hba
41	R	EX_thm_e	Exchange reaction for thm
42	R	HACD1	3-hydroxyacyl-CoA dehydrogenase (acetoacetyl-CoA)
43	R	HACD1i	3-hydroxyacyl-CoA dehydrogenase (acetoacetyl-CoA)
44	R	HACD2	3-hydroxyacyl-CoA dehydrogenase (3-oxohexanoyl-CoA)
45	R	HACD2i	3-hydroxyacyl-CoA dehydrogenase (3-oxohexanoyl-CoA)
46	R	HACD3	3-hydroxyacyl-CoA dehydrogenase (3-oxooctanoyl-CoA)
47	R	HACD3i	3-hydroxyacyl-CoA dehydrogenase (3-oxooctanoyl-CoA)
48	R	HACD4	3-hydroxyacyl-CoA dehydrogenase (3-oxodecanoyl-CoA)
49	R	HACD4i	3-hydroxyacyl-CoA dehydrogenase (3-oxodecanoyl-CoA)
50	R	HACD5	3-hydroxyacyl-CoA dehydrogenase (3-oxododecanoyl-CoA)
51	R	HACD5i	3-hydroxyacyl-CoA dehydrogenase (3-oxododecanoyl-CoA)
52	R	HACD6	3-hydroxyacyl-CoA dehydrogenase (3-oxotetradecanoyl-CoA)
53	R	HACD6i	3-hydroxyacyl-CoA dehydrogenase (3-oxotetradecanoyl-CoA)
54	R	HACD7	3-hydroxyacyl-CoA dehydrogenase (3-oxohexadecanoyl-CoA)
55	R	HACD7i	3-hydroxyacyl-CoA dehydrogenase (3-oxohexadecanoyl-CoA)
56	R	HACD8	3-hydroxyacyl-CoA dehydrogenase (3-oxooctadecanoyl-CoA)
57	R	HACD8i	3-hydroxyacyl-CoA dehydrogenase (3-oxooctadecanoyl-CoA)
58	R	HETZK	Hydroxyethylthiazole kinase
59	R	HMPK1	Hydroxymethylpyrimidine kinase (ATP)
60	R	ICHORS	Isochorismate synthase
61	R	K2L4Aabcpp	KDO(2)-lipid IV A transport via ABC system (periplasm)
62	R	KAT1	3-ketoacyl-CoA thiolase
63	R	KDOCT2	3-deoxy-manno-octulosonate cytidylyltransferase
64	R	KDOPP	3-deoxy-manno-octulosonate-8-phosphatase
65	R	KDOPS	3-deoxy -D-manno-octulosonic -acid 8-phosphate synthase
66	R	LPADSS	Lipid A disaccaride synthase
67	R	MOAT	3-deoxy-D-manno-octulosonic acid transferase
68	R	MOAT2	3-deoxy-D-manno-octulosonic acid transferase
69	R	NADH9	NADH dehydrogenase (demethylmenaquinone-8 & 0 protons)
70	R	NADPHQR4	NADPH Quinone Reductase (2-Demethylmenaquinone-8)
71	R	OCTDPS	Octaprenyl pyrophosphate synthase
72	R	PMPK	Phosphomethylpyrimidine kinase
73	R	RHACOAR140	3R 3 Hydroxyacyl CoANADP oxidoreductase
74	R	SEPHCHCS	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase

75	R	SHCHCS3	2-succinyl-6-hydroxy-2,4-cyclohexadiene 1-carboxylate synthase
76	R	SUCBZL	O-succinylbenzoate-CoA ligase
77	R	SUCBZS	O-succinylbenzoate-CoA synthase
78	R	TDP	Thiamin pyrophosphatase
79	R	TDSK	Tetraacyldisaccharide 4'kinase
80	R	THMabcpp	Thiamine transport via ABC system (periplasm)
81	R	THMtex	Thiamine transport via diffusion (extracellular to periplasm)
82	R	THZPSN	Thiazole phosphate synthesis
83	R	TMK	Thiamine kinase
84	R	TMN	Thiaminase
85	R	TMPK	Thiamine-phosphate kinase
86	R	TMPPP	Thiamine-phosphate diphosphorylase
87	R	U23GAAT	UDP-3-O-(3-hydroxymyristoyl)glucosamine acyltransferase
88	R	UAGAAT	UDP-N-acetylglucosamine acyltransferase
89	R	UHGADA	UDP-3-O-acetylglucosamine deacetylase
90	R	USHD	UDP-sugar hydrolase
91	M	14dhncoa_c	1,4-dihydroxy-2-napthoyl-CoA
92	M	2dmmq8_c	2-Demethylmenaquinone 8
93	M	2dmmql8_c	2-Demethylmenaquinol 8
94	M	2mahmp_c	2-Methyl-4-amino-5-hydroxymethylpyrimidine diphosphate
95	M	2sephchc_c	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate
96	M	2shchc_c	2-Succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate
97	M	3hbcoa_c	(S)-3-Hydroxybutanoyl-CoA
98	M	3hdcoa_c	(S)-3-Hydroxydecanoyl-CoA
99	M	3hddcoa_c	(S)-3-Hydroxydodecanoyl-CoA
100	M	3hhcoa_c	(S)-3-Hydroxyhexanoyl-CoA
101	M	3hhdcoa_c	(S)-3-Hydroxyhexadecanoyl-CoA
102	M	3hmrsACP_c	(R)-3-Hydroxytetradecanoyl-[acyl-carrier protein]
103	M	3hocoa_c	(S)-3-Hydroxyoctanoyl-CoA
104	M	3hodcoa_c	(S)-3-Hydroxyoctadecanoyl-CoA
105	M	3htdcoa_c	(S)-3-Hydroxytetradecanoyl-CoA
106	M	3odcoa_c	3-Oxodecanoyl-CoA
107	M	3oddcoa_c	3-Oxododecanoyl-CoA
108	M	3ohcoa_c	3-Oxohexanoyl-CoA
109	M	3ohdcoa_c	3-Oxohexadecanoyl-CoA
110	M	3ohodcoa_c	3-Oxooctadecanoyl-CoA
111	M	3oocoa_c	3-Oxooctanoyl-CoA
112	M	3otdcoa_c	3-Oxotetradecanoyl-CoA
113	M	4ahmmp_c	4-Amino-5-hydroxymethyl-2-methylpyrimidine
114	M	4ampm_c	4-Amino-2-methyl-5-phosphomethylpyrimidine
115	M	4hba_c	4-Hydroxy-benzyl alcohol
116	M	4hba_e	4-Hydroxy-benzyl alcohol (external)
117	M	4mhetz_c	4-Methyl-5-(2-hydroxyethyl)-thiazole
118	M	4mpetz_c	4-Methyl-5-(2-phosphoethyl)-thiazole
119	M	ACP_c	Acyl carrier protein
120	M	R_3hmrscoa_c	(R)-3-hydroxytetradecanoyl-coa
121	M	aacoa_c	Acetoacetyl-CoA
122	M	ac_c	Acetate
123	M	accoa_c	Acetyl-CoA
124	M	adp_c	ADP
125	M	ahcys_c	S-Adenosyl-L-homocysteine
126	M	air_c	5-amino-1-(5-phospho-D-ribosyl)imidazole

127	M	akg_c	2-Oxoglutarate
128	M	ala__L_c	L-Alanine
129	M	amet_c	S-Adenosyl-L-methionine
130	M	amp_c	AMP
131	M	ara5p_c	D-Arabinose 5-phosphate
132	M	atp_c	ATP
133	M	b2coa_c	Crotonoyl-CoA
134	M	btcoa_c	Butanoyl-CoA
135	M	chor_c	Chorismate
136	M	ckdo_c	CMP-3-deoxy-D-manno-octulosonate
137	M	cmp_c	CMP
138	M	co2_c	CO2
139	M	coa_c	Coenzyme A
140	M	ctp_c	CTP
141	M	cys__L_c	L-Cysteine
142	M	dc2coa_c	Trans-Dec-2-enoyl-CoA
143	M	dcacoa_c	Decanoyl-CoA (n-C10:0CoA)
144	M	dd2coa_c	Trans-Dodec-2-enoyl-CoA
145	M	ddcacoa_c	Dodecanoyl-CoA (n-C12:0CoA)
146	M	dhna_c	1,4-Dihydroxy-2-naphthoate
147	M	dxyl5p_c	1-deoxy-D-xylulose 5-phosphate
148	M	fad_c	Flavin adenine dinucleotide oxidized
149	M	fadh2_c	Flavin adenine dinucleotide reduced
150	M	for_c	Formate
151	M	frdp_c	Farnesyl diphosphate
152	M	gcald_c	Glycolaldehyde
153	M	h2o_c	H2O
154	M	h_c	H+
155	M	hdd2coa_c	Trans-Hexadec-2-enoyl-CoA
156	M	hx2coa_c	Trans-Hex-2-enoyl-CoA
157	M	hxcoa_c	Hexanoyl-CoA (n-C6:0CoA)
158	M	ichor_c	Isochorismate
159	M	ipdp_c	Isopentenyl diphosphate
160	M	kdo2lipid4_c	KDO(2)-lipid IV(A)
161	M	kdo2lipid4_p	KDO(2)-lipid IV(A) (periplasm)
162	M	kdo8p_c	3-Deoxy-D-manno-octulosonate 8-phosphate
163	M	kdo_c	3-Deoxy-D-manno-2-octulosonate
164	M	kdolipid4_c	KDO-lipid IV(A)
165	M	lipidA_c	2,3,2'3'-Tetrakis(beta-hydroxymyristoyl)-D-glucosaminyl-1,6-beta-D-glucosamine 1,4'-bisphosphate
166	M	lipidAds_c	Lipid A Disaccharide
167	M	lipidX_c	2,3-Bis(3-hydroxytetradecanoyl)-beta-D-glucosaminyl 1-phosphate
168	M	mql8_c	Menaquinol 8
169	M	mqn8_c	Menaquinone 8
170	M	nad_c	Nicotinamide adenine dinucleotide
171	M	nadh_c	Nicotinamide adenine dinucleotide - reduced
172	M	nadp_c	Nicotinamide adenine dinucleotide phosphate
173	M	nadph_c	Nicotinamide adenine dinucleotide phosphate - reduced
174	M	o2_c	O2 O2
175	M	oc2coa_c	Trans-Oct-2-enoyl-CoA
176	M	occoa_c	Octanoyl-CoA (n-C8:0CoA)
177	M	octdp_c	All-trans-Octaprenyl diphosphate

178	M	od2coa_c	Trans-Octadec-2-enoyl-CoA
179	M	pep_c	Phosphoenolpyruvate
180	M	pi_c	Phosphate
181	M	pmtcoa_c	Palmitoyl-CoA (n-C16:0CoA)
182	M	ppi_c	Diphosphate
183	M	pyr_c	Pyruvate
184	M	sbzcoa_c	O-Succinylbenzoyl-CoA
185	M	stcoa_c	Stearoyl-CoA (n-C18:0CoA)
186	M	sucbz_c	O-Succinylbenzoate
187	M	td2coa_c	Trans-Tetradec-2-enoyl-CoA
188	M	tdcoa_c	Tetradecanoyl-CoA (n-C14:0CoA)
189	M	thm_c	Thiamin
190	M	thm_e	Thiamin (external)
191	M	thm_p	Thiamin (periplasm)
192	M	thmmp_c	Thiamin monophosphate
193	M	thmpp_c	Thiamine diphosphate
194	M	tyr__L_c	L-Tyrosine
195	M	u23ga_c	UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine
196	M	u3aga_c	UDP-3-O-(3-hydroxytetradecanoyl)-N-acetylglucosamine
197	M	u3hga_c	UDP-3-O-(3-hydroxytetradecanoyl)-D-glucosamine
198	M	uacgam_c	UDP-N-acetyl-D-glucosamine
199	M	udp_c	UDP
200	M	ump_c	UMP

Chapter 4 | Large-scale metabolic clustering of *Lactiplantibacillus plantarum* strains

4.1 Introduction

Lactic acid bacteria (LABs) are a group of gram-positive bacteria belonging to the *Lactobacillales* order and sharing key features of their physiology. They are facultatively anaerobic, non-motile bacteria, colonizing a wide array of different environments: animal gut, milk, cereals, vegetables, meat, fish, beer, wine, sourdough, etc. Some of them are opportunistic or obligate pathogens, but the majority are safe, with a long tradition of use in food fermentations. Such bacteria, in particular those used in food processing, tend to show genome reduction as a consequence of adaptation to nutrient-rich environments, where carbohydrates are rapidly converted in lactic acid and other organic acids, released to inhibit growth of competitors. This fitness strategy involves a low-efficiency metabolism, as discussed in **Chapter 1**, because energy-containing molecules are secreted and thus "wasted". In the food industry LABs play a key role since they determine taste, texture, flavor, nutritional value and preservability of foods, where preservation features are mainly due to the high amount of lactic acid released. For these reasons, they are often part of starter and non-starter cultures [8,331–336].

The importance of LABs have also been recognized in human health (and animal health in general), with many strains proved to have probiotic activity. Probiotics are living organisms that, if assumed in sufficient amounts, have positive effects for the human. Indeed, they play a crucial role in maintaining the gut microbiota equilibrium, preventing the proliferation of harmful bacteria which would lead to gut dysbiosis. Also non-viable LABs displayed positive effects on human health, and since they are not living cells they are referred to as parabiotics. Fortunately, *Lactobacillales* are often GRAS (generally recognized as safe), a status released by the FDA (Food and Drug Agency), so they are legally authorized to be used in food applications in the United States. Similarly, in the European Union, they often have the QPS status (qualified presumption of safety), which is released by EFSA (European Food Safety Authority). To reach a GRAS/QPS designation, it is needed to prove the organism is not mutagenic, carcinogenic, pathogenic, nor resistant to antibiotics, provided a clear taxonomic identification at the strain level [333,335,337].

Traditionally, LABs were divided into three groups according to their metabolism: (group I) obligate homofermentative, were hexose were fermented to lactate via the EMP (Embden-Meyerhof-Parnas) pathway, and pentoses were not fermented due to the absence of a phosphoketolase; (group II) facultative heterofermentative, were hexose were fermented via the EMP pathway reaching depletion, then pentoses were fermented using an inducible phosphoketolase, leading to acetate in addition to lactate; (group III) obligate heterofermentative, were hexoses and pentoses were both fermented through the phosphoketolase pathway since the fructose-1,6-bisphosphate aldolase (EMP) was missing, leading to ethanol and CO_2 in addition to lactate in case of hexoses [334,336].

However, this traditional division in three groups do not take into consideration the case where pentoses are fermented exclusively to lactate, and it is not in line with the phylogeny of the old genus *Lactobacillus* [338], which indeed was later subjected to taxonomic revision [339]. Therefore, today the distinction is simpler, with just two groups, homofermentative and heterofermentative, depending on the enzyme used to ferment hexoses: the first group performs a homolactic fermentation using a fructose-1,6-bisphosphate aldolase (EMP), the second carries out a heterolactic fermentation using a phosphoketolase [332,338,339]. Moreover, homofermentative LABs in anaerobiosis and glucose limitation may perform a mixed-acid fermentation, adding formic acid among the fermentation products, thanks to the activity of a pyruvate formate lyase [340,341].

Lactiplantibacillus plantarum is a genetically tractable, model LAB species. It is traditionally classified as facultative heterofermentative, meaning that it can lead to both homolactic and heterolactic fermentation profiles. It has a relatively large genome: 3.3 Mb compared to 1.8-2.6 Mb of other LABs [333,342], and it is able to colonize many different environmental niches: animal gut, diary products, vaginal tract, vegetables, etc. This is possible thanks to the plasticity of its genome, which includes genetic determinants needed to grow in various environments. Determinants for the catabolism of carbon sources have been described in *L. plantarum* as “carbohydrate utilization islands”, which are gene clusters needed for growing on a specific carbon source, subjected to lateral gene transfer, and usually placed near the origin of replication in a so-called “lifestyle adaptation island” characterized by a low GC content. Reflecting its ability to survive in many different habitats, genomes of *L. plantarum* are widely rich in transporters, which can be roughly divided in ABC-type (ATP-binding cassettes) and PTSs (phosphotransferase systems). ABCs are mainly involved in transport of peptides and amino acids, while PTSs are mainly involved in transport of carbohydrates and show substrate promiscuity. Like many other

LABs being evolved on nutrient-rich substrates, *L. plantarum* is auxotrophic for several aminoacids and vitamins [3,8,337,342].

One of the most studied strains of *L. plantarum*, and of LABs in general, is *L. plantarum* WCFS1. It is a probiotic strain originally isolated from saliva, able to survive for more than 6 days in the human gut. It has a 3.3 Mb chromosome plus three plasmids, two of them defined as "cryptic" due to their dimensions: 2.4 Kb, 1.9 Kb, and 36,1 Kb. It was the first representative of the old *Lactobacillus* genus to be fully sequenced in 2003 [3,11,342]. In 2006, its metabolism was modeled on a genome-scale by Teusink and colleagues [8], leading to one of the first manually-curated genome-scale metabolic models (GSMMs) for LABs. With single omission experiments, WCFS1 was proven to have ten auxotrophies for amino acids (arginine, glutamate, isoleucine, leucine, methionine, valine, phenylalanine, tryptophan, tyrosine, threonine) and three auxotrophies for vitamins (nicotinic acid, pantothenic acid, riboflavin), some of which were not modeled in the first version of the GSMM, due to the presence of gaps in the network. For example, cysteine was modeled as essential due to an incomplete biosynthetic pathway; however, Wegkamp and colleagues experimentally proved the opposite in 2010, supposing that a small amount of methionine could be somehow converted into cysteine inside the cell [132].

When the WCFS1 model was used to simulate growth on a chemically defined medium (CDM), despite many uptake rates being measured in chemosat and used as constraints, the obtained growth rate was unrealistically high, according to a known FBA limitation discussed in **Chapter 1**. Specifically, *L. plantarum* performs homolactic fermentation, an inefficient metabolism with release of lactate, even in chemostat at low growth rates. Instead, when the same growth conditions were simulated, FBA predicted a flux distribution compatible with a mixed-acid fermentation. This happened because this type of fermentation leads to a higher energetic yield (3 ATPs per glucose) compared to homolactic fermentation (2 ATPs per glucose), energy that can be used to produce more biomass. In other words, mixed acid fermentation was predicted as it leads to better substrate usage efficiency, which is the real objective that FBA mathematically tries to maximize. This is in line with the evolutionary history of *L. plantarum*, adapted to grow on nutrient-rich media: the usage of sugars is less efficient than other bacteria where FBA works well, such as for example lab-adapted strains of *Escherichia coli*. Indeed, the fitness strategy of LABs is different: an inefficient sugar catabolism was selected to quickly produce lactic acid, to inhibit growth of competitors for those nutrient-rich niches [77].

Apart from *L. plantarum* WCFS1, other curated GSMMs for LABs have been reconstructed [9,24]. Despite the limitations of FBA, particularly relevant for LAB species, the conversion from genes to reactions provided by GSMMs offers a way to explore strain variability among a species. Indeed, even if the definition of bacterial species describes its members as *"genomically and phenotypically coherent"* [20], it is known that strains of the same species may display genomic and eventually phenotypic differences. These differences are due to the adaptation to different environmental niches, where possible mutations, gene losses, and/or acquisition of exclusive genes through horizontal gene transfer events can determine different evolutionary patterns [10] and strain-specific metabolic capabilities within a species [3], [19]. This is particularly relevant in industry, as the many industrial applications of LABs are strain dependent. In the food industry, for example, different strains can confer different flavors, texture, and nutritional values to the same edible product, given a fixed production process [173]. Therefore, to optimize a microbial process, systems biology approaches like the genome-based metabolic modeling could be used to select industrially relevant strains in a rational way, as an alternative to traditional strain selection guided by trials and errors, significantly reducing both time and costs [73].

To start with, the content of GSMMs, together with the simulations they provide, should give the opportunity to isolate groups of strains sharing similar metabolic potential, clustering strains according to their metabolic features (see **Chapter 1**). This can be made by accounting for binary features such as the presence / absence of reactions, auxotrophies, and ability to grow on alternative substrates. These data-driven clusters of strains could reflect phylogeny and particular niche adaptations; in addition, when closely related species are processed, they could highlight the metabolic boundaries that distinguish a species from another [24], [25], [26], [27], [28], [29]. For example, if strains belonging to a particular cluster are predicted to have catabolic activity for plant-derived carbon sources like raffinose, panose and pullulan, then the cluster could correspond to a particular clade adopted to grow on plants [9].

Here, a preliminary attempt to explore the metabolic biodiversity of *L. plantarum* through GSMMs is reported. Specifically, the consistency between strain-specific isolation niches and data-driven metabolic clusters was verified.

4.2 Results

4.2.1 Reference sanity check

The GSMM for *L. plantarum* WCFS1 [8] (iLP728) was downloaded and entirely hand-drawn on an Escher map [35,36] to better visualize and understand its structure. Then, small updates were applied regarding the formula and charge of metabolites and reaction balances. This led the updated reference model (iLP728u) to be devoid of unbalanced metabolic reactions.

iLP728u was then constrained with literature data to evaluate its accuracy. First, uptake and secretion rates recorded in a chemostat [8] were used (**Table 4.S1**). The chemostat was originally set up with a chemically defined medium (CDM) specific for *L. plantarum* WCFS1 [3,132], with 25 mM glucose and a dilution rate of ~0.3 h^{-1} [8]. Using flux-balance analysis (FBA), the growth rate was optimized to 0.363 h^{-1}. Then, exchange reactions were set to represent CDM concentrations (**Table 4.S2**) known to yield ~1.0 gDW/L [8]. In this case, FBA predicted a maximum theoretical growth yield of 1.511 gDW/L, in accordance with an unrealistic mixed acid fermentation. In order to force the FBA optimization to resemble a homolactic fermentation, the acetate branch was switched off, this way predicting a yield of 1.012 gDW/L.

4.2.2 Isolation niche definition

1022 genome assemblies of *L. plantarum* were downloaded from NCBI. Since there is no standard procedure on NCBI to define the isolation niche metadata upon the upload of a genome, all genomes metadata were manually reviewed to classify each assembly into arbitrary, standardized isolation niches (**Figure 4.1**).

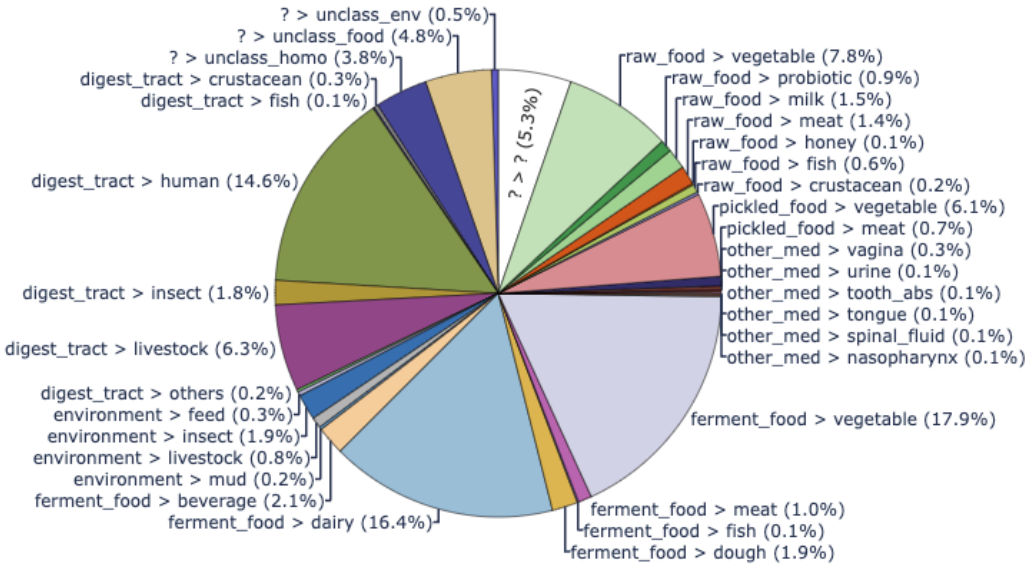

Figure 4.1. Results of the manual classification of 1022 downloaded genome assemblies into categories of isolation niche.

4.2.3 Genomes filtering

Genome assemblies were then filtered based on technical metrics like N50 (**Figure 4.S1A**) and number of contigs (**Figure 4.S1B**), as well as on biological metrics such as the number of expected single-copy orthologs in the Lactobacillales order (**Figure 4.S1C**). This selection removed 85 low-quality assemblies. Then, if multiple good-quality assemblies remained for the same strain, the assembly with the lowest number of contigs was retained, discarding the others. This removed another 25 assemblies, leading to a total of 912 strains respectively associated with a single, good-quality assembly.

The remaining genomes were subjected to a second round of filtering based on taxonomy. An all-vs-all average nucleotide identity (ANI) was computed for the 912 good-quality genomes (**Figure 4.2**). From the dendrogram built on the ANI matrix, two branches containing a total of 20 genomes with a noticeably lower average ANI value were identified. These genomes were primarily associated with a "taxonomy-check-status" (TSC) attribute equal to "failed", indicating that NCBI had flagged them as likely having incorrect species assignments (**Figure 4.3B**).

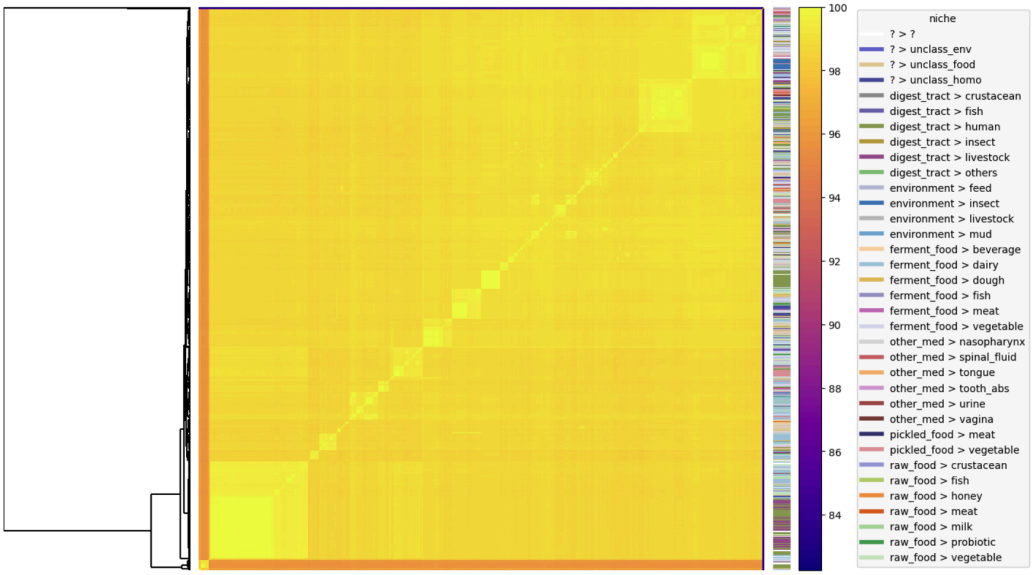

Figure 4.2. ANI analysis of 912 good-quality assemblies associated on NCBI with the species *L. plantarum* (taxid 1590). On the left, a dendrogram based on the ANI matrix. On the right, a colormap indicating the niche metadata assigned to each genome assembly.

Genomes were then filtered based on a 97% ANI threshold with respect to both the WCFS1 and the type strain, precisely removing the genomes belonging to two branches above mentioned (**Figure 4.3A**), leading to a total of 892 genomes of good-quality and confidently belonging to the *L. plantarum* species.

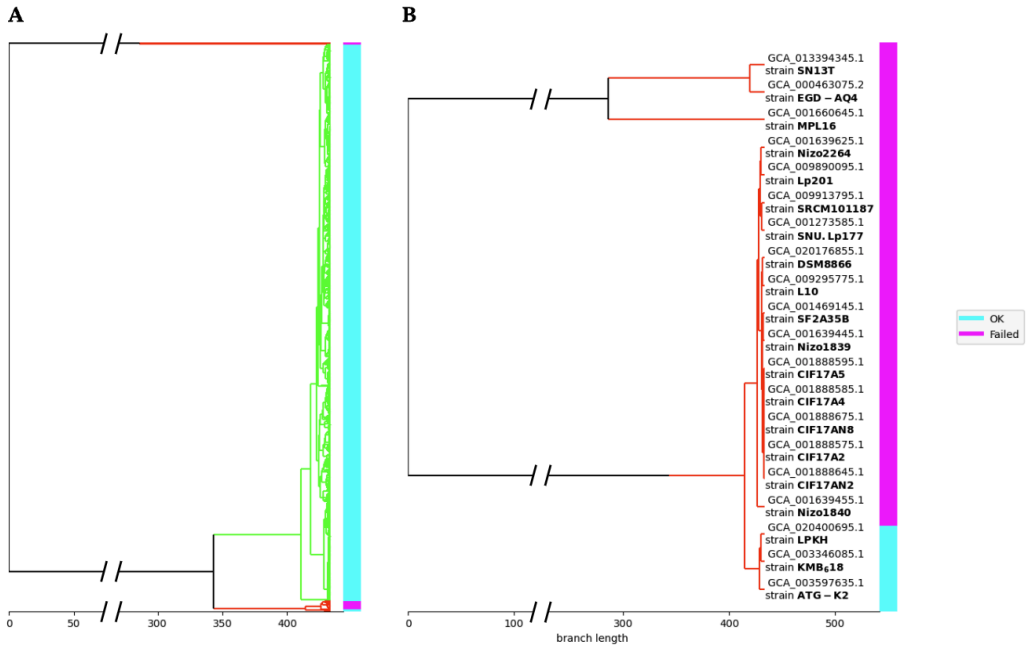

Figure 4.3. Dendrogram based on the ANI matrix for 912 good-quality assemblies. A colorbar reporting the NCBI TCS attribute associated with each genome is shown on the right of each panel. **A)** Dendrogram showing in green the 2nd-level branch with the largest number of genomes. **B)** The same dendrogram, but without the green branch; genome accessions and strain names are shown.

4.2.4 Multi-strain reconstruction

A multi-strain reconstruction of GSMMs was performed with Gempipe (see Chapter 2), giving the 892 remaining genomes in input and using iLP728u as reference. Briefly, iLP728u was expanded with new, strain-specific reactions to form a draft pan-GSMM. Energy-generating cycles (EGCs) [107] were detected and removed from the draft pan-GSMM. 38 universally blocked reactions in iLP728u resulted unblocked in the draft pan-GSMM as a consequence of the reference expansion procedure. The draft pan-GSMM was then used to derive strain-specific GSMMs, one for each strain. Thanks to the hybrid reconstruction method, the strain-specific GSMMs were based on iLP728u but also expanded with new genes, reactions, and metabolites (**Figure 4.4**).

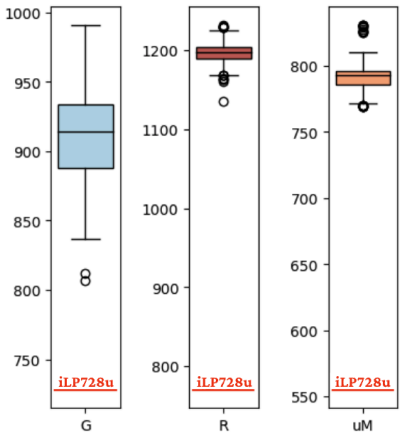

Figure 4.4. Content of strain-specific GSMMs compared with the reference iLP728u (red line). "G": number of genes. "R": number of reactions. "uM": number of unique metabolites (i.e., without considering the compartment).

Literature data for substrate usage in different *L. plantarum* strains were searched in literature. Data from Siezen *et al.* [343] were selected, transformed into binary format and compared with simulations. While the number of modeled substrates and sequenced strains was low, an accuracy of 89.2% was reached without any prior manual gap-filling (**Figure 4.5**).

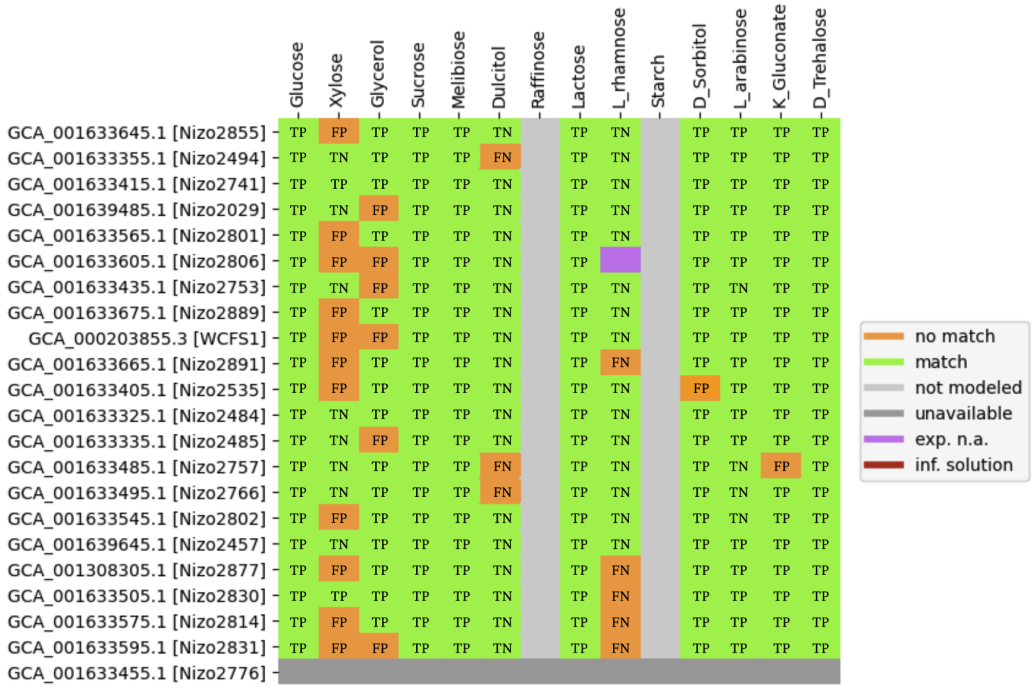

Figure 4.5. Match between growth simulations and experimental substrate utilization data. Accessions in row refer to the sequenced strains in [343]. Substrates in column refer to substrates tested in [343] and disposing of an exchange reaction in the BiGG database [28,56]. "TP": true positive; "TN": true negative; "FP": false positive; "FN": false negative. GCA_001633455.1 (dark gray) did not respect the quality thresholds, meaning that it was discarded based either on N50, number of contigs, or missing BUSCO orthologs [179]. "exp. n.a.": experimental data not available. "inf. solution": the system of equations described by FBA has no solution, not even 0 (solver status is "infeasible") (no occurrences in this comparison).

4.2.5 Multi-strain analysis

For each strain-specific GSMM, the reaction content was analyzed, and eventual auxotrophies for amino acids and vitamins were predicted. Moreover, the potential utilization of alternative C, N, P and S sources was predicted as well. Combining these three layers of binary data (presence / absence of reactions, auxotrophies, and ability to grow on alternative substrates), a "phylometabolic" tree was built, which presents each strain in close proximity with those having similar metabolic potential.

To divide the phylometabolic tree in clusters, a Silhouette analysis [219] was performed, which evaluates the quality of clustering by measuring how well each data point fits within its own cluster compared to neighboring clusters. The Silhouette Coefficient quantifies the cohesion and separation of individual

data points, ranging from -1 (misclassified) to +1 (well-clustered), providing insight into the suitability of cluster assignments. The Average Silhouette Score, calculated across all data points, gives an overall measure of clustering efficacy, helping identify optimal cluster numbers and detect poorly formed clusters. In the range considered, the average Silhouette score (**Figure 4.6A**) steadily increased; however, after manual inspection, it was decided to extract 8 clusters (**Figure 4.6B**), as there was a concrete risk of overfitting at higher values.

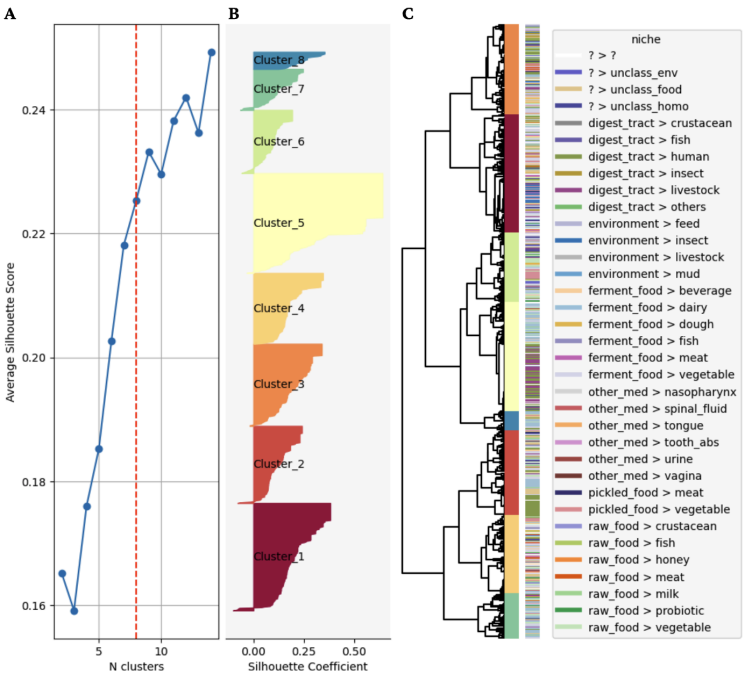

Figure 4.6. Evaluation of different numbers of clusters to be used during the cutting of the phylometabolic tree. **A)** Average Silhouette score for different numbers of clusters. The red line represents the number chosen after manual inspection. **B)** Strain-specific silhouette coefficients when dividing the phylometabolic tree in 8 clusters. **C)** Phylometabolic tree associated with data-driven clusters. The isolation niche metadata is plotted next to the cluster attribute for each strain.

Clusters were represented side-by-side with the isolation niche metadata (**Figure 4.6B**), to inspect for eventual consistencies. Unfortunately, it was not possible to detect consistency between a cluster and any category of isolation niche, despite clusters being evidently characterized by different metabolic features (**Figure 4.7**).

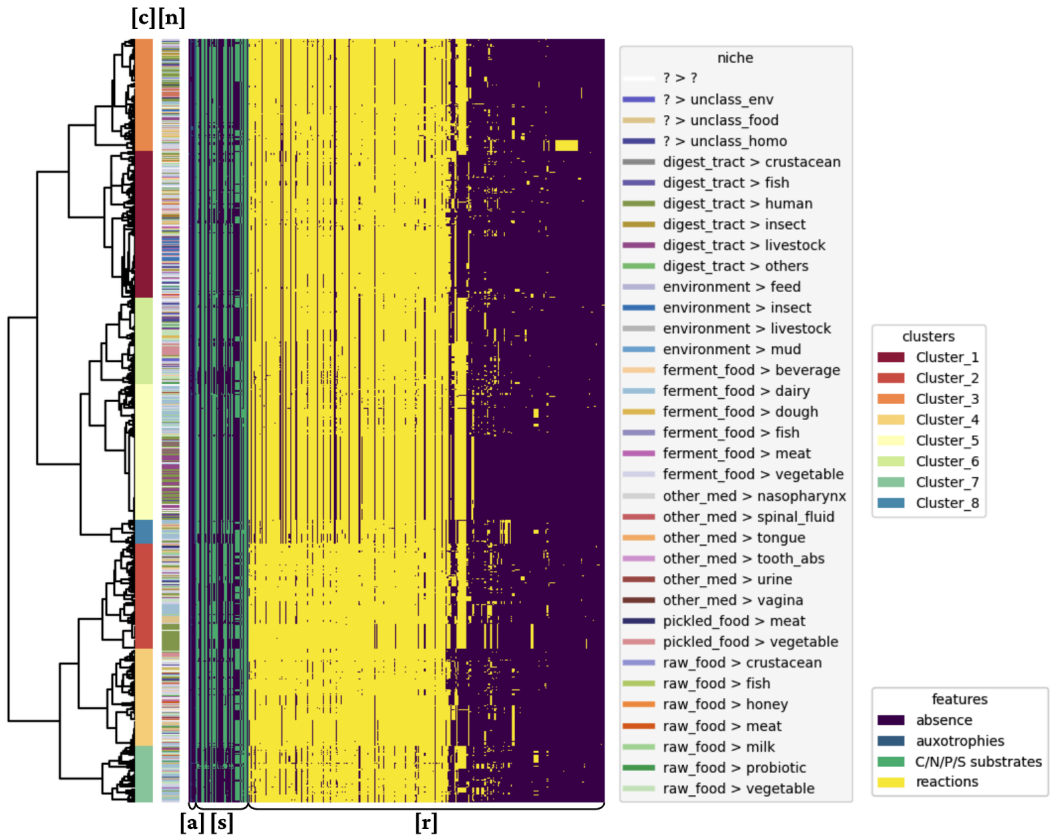

Figure 4.7. Binary metabolic features for each strain in the phylometabolic tree. Three layers of features are represented from left to right: presence of auxotrophies ([a], narrow blue region), ability to grow on alternative substrates ([s], green region), presence metabolic reactions ([r], yellow region). Regardless of the layer, absence of a feature is represented with the same color (violet). Two columns are represented next to the feature matrix: the cluster attribute for each strain ([c], left column); the isolation niche metadata ([n], right column). Only varying features (i.e., not always present or absent in all strains) are shown.

Each cluster was characterized in terms of metabolic features generally present or absent (**Figure 4.8**). To give some examples, *L. plantarum* strains belonging to Cluster_2, Cluster_4, and Cluster_7 are generally able to produce thiamine, while strains in the other clusters are auxotrophs ("[aux]thm"). Moreover, the same strains are also able to produce a functional nitrate reductase ("NITR", "NTRARy"), including the molybdenum cofactor that is needed by the nitrate reductase to operate ("MOCOS_LPL", molybdenum cofactor synthase). Strains in Cluster_5 are not able to grow an arabinose ("[C]arab__L"), as the dedicated transporter ("ARAB_Lt"), the arabinose isomerase ("ARAI") and the ribulokinase ("RBK_L1") are missing from their reaction network, preventing the sugar to be catabolized into ribulose-5-phosphate.

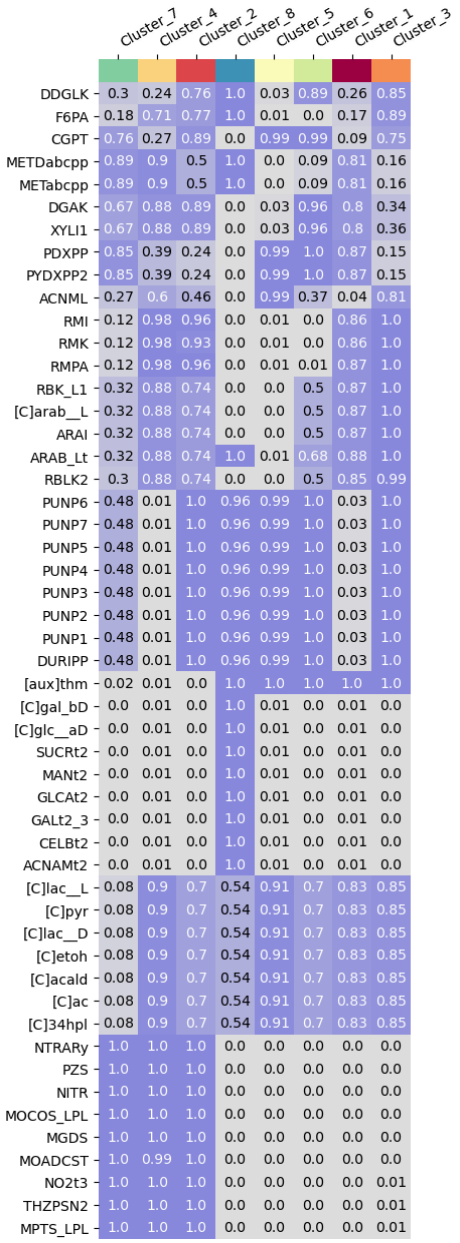

Figure 4.8. Metabolic features characterizing each data-driven cluster. Each cell shows the relative frequency of each feature in each cluster, in terms of presence. Only features with relative frequency ≤ 0.1 and ≥ 0.9 in two different clusters are shown. Features starting with "[aux]" refer to the presence of auxotrophies. Features starting with "[C]", refer to the ability to catabolize C substrates. All the other features refer to the presence of metabolic reactions.

4.3 Methods

Manipulations and simulations of GSMMs were performed with COBRApy v0.29.0 [38] using CPLEX v22.1.1 [39] as backend linear optimization solver, unless otherwise stated.

4.3.1 Reference update

The reference GSMM for *L. plantarum* WCFS1 [8] was downloaded from [34] supplementary materials in an updated, BiGG-compatible version called iLP728. The associated reference proteome was downloaded from the same source. The biomass equation was switched from "biomass_LPL_RETB_t576 _NoATP" (suitable to simulate growth in a retentostat [344]) to "biomass_ LPL60". The demand reaction for 4-hydroxy-5-methyl-3(2H)-furanone ("hmfurn_c") was replaced with a diffusion reaction, as the metabolite is known to be volatile [345]. The maximum bound constant of 999999 was replaced with 1000, as it is more commonly used. The chemical formulas of fatty acids bonded to the acyl-carrier protein (ACP) were redefined to show all the hydrogen bonds; for example, hexanoyl-ACP, originally modeled as "C6H10OX", was updated to "C6H11OX". Moreover, a formula was defined for the ACP ("apoACP_c", "ACP_c") and the molybdopterin synthase ("MCOSH_c", "MCOOH_c"), and a charge was defined for the lipoteichoic acids ("LTA_LPL_c", "LTAglc_LPL_c"), which were originally missing. With these changes, the model resulted without mass- or charge-unbalanced reactions. Finally, the capsular polysaccharides biosynthetic reaction ("CPSS_LPL2") was updated in its GPR using information coming from a study published after the original model [346]. The resulting model (iLP728u) was used as reference in the multi-strain reconstruction.

4.3.2 Reference validation

Growth of iLP728u was evaluated through FBA using two datasets. The first was a set of experimental uptake and secretion rates derived from tables in [8], describing a chemostat culture of WCFS1 on a CDM with 25 mM glucose at dilution rate 0.3 h^{-1}; with this dataset, exchange reactions were set up to accommodate the experimental error (**Table 4.S1**). The second dataset represented solution concentrations of the CDM (**Table 4.S2**), that were directly used to constrain exchange reactions [68]. With the latter dataset, in order to force the FBA optimization to resemble a homolactic fermentation, bounds of the reactions pyruvate formate lyase (PFL) and acetate kinase (ACKr) were arbitrarily set to 0.

4.3.3 Genomes download

A set of 1022 *L. plantarum* genome assemblies was downloaded from NCBI [178] on 2nd November 2023 with the ncbi-genome-download[18] v0.3.3 tool using parameters *--species-taxids 1590 --section genbank*. With Biopython v1.80 [197],

[18] https://github.com/kblin/ncbi-genome-download

the TSC attribute was retrieved from NCBI for each genome, by executing a custom Python script relying on the *Bio.Entrez* module.

4.3.4 Niche metadata definition

When NCBI users upload a new genome assembly on GenBank, they are free to associate to the assembly any number of attributes, defining them manually; with a custom Python script based on *Bio.Entrez*, all the user-defined attributes were downloaded for each genome. 152 attributes were obtained and manually inspected, and 36 of them were selected because they contained information somehow connected to the isolation niche: “env_material”, “sample_type”, “env_feature”, “project_name”, “environment (material)”, “env_medium”, “depth”, “food_product_type”, “ref_biomaterial”, “metagenome_source”, “isolation source”, “host scientific name”, “host”, “isolation-source”, “env_broad_scale”, “ifsac_category”, “environment (feature)”, “miscellaneous environmental package”, “environment (biome)”, “nat_host”, “geo_loc_name”, “nat-host”, “geographic location (country and/or sea)”, “geographic location (region and locality)”, “env_local_scale”, “env_biome”, “project name”, “metagenomic source”, “isolation_source”, “host_tissue_sampled”, “coll_site_geo_feat”, “Isolation Site”, “biome”, “feature”, “misc_param: HMP supersite”, “specific host”.

The selected attributes were concatenated together forming a single attribute string. For each genome assembly, the attribute string was processed by searching manually defined sub-strings using boolean logic, in order to associate each genome to a standardized niche. This niche metadata was arbitrarily divided into a main level (N1) and a sub-level (N2). For example, if the attribute string contained the following sub-strings “(‘ferment’) AND (‘sausage’ OR ‘meat’ OR ‘salami’ OR ‘pork’)”, then N1 was assigned to “raw_food”, and N2 to “meat”.

N1 levels included: digestive tract (“digest_tract”), raw food (“raw_food”), fermented food (“ferment_food”), pickled food (“pickled_food”), medical specimens (“other_med”), environment (“environment”), unclassified (“?”).

N2 sub-levels of “digest_tract” included: “insect”, “crustacean”, “fish”, “human”, “livestock”, “others”; N2 sub-levels of “raw_food” included: “vegetable”, “milk”, “meat”, “fish”, “crustacean”, “honey”, “probiotic”; N2 sub-levels of “ferment_food” included: “vegetable”, “beverage”, “meat”, “fish”, “dairy”, “dough”; N2 sub-levels of “pickled_food” included: “vegetable”, “meat”. N2 sub-levels of “other_med” included: “vagina”, “urine”, “spinal_fluid”, “tooth_abs” (tooth abscess), “tongue”, “nasopharynx”. N2 sub-levels of

"environment" included: "mud", "livestock", "insect", "feed", "etoh_ind" (ethanol industry). Unclassified N2 levels were: "unclass_food" (unclassified food sample), "unclass_env" (unclassified environmental sample), "unclass_homo" (unclassified human sample), "?" (impossible to categorize).

4.3.5 Genomes filtering

Downloaded genomes were filtered both for quality and for taxonomy. For each downloaded genome, N50 and number of contigs were computed with SeqKit v2.2.0 [47] using *stats --tabular --basename --all*. Genes were predicted with Prodigal v2.6.3 [48] run through Prokka v1.14.6 [49] with options *--noanno --norrna --notrna*. Resulting protein sequences were evaluated with BUSCO v5.4.0 [36] using parameters *--mode proteins --lineage_dataset lactobacillales_odb10*. Downloaded genomes were discarded when having either more than 8 missing BUSCO orthologs (out of 402 orthologs contained in the database *lactobacillales_odb10*), more than 200 contigs, or N50 lower than 50000. When two or more good-quality genomes were associated to the same strain, the assembly with less contigs was retained, discarding the others.

Using the remaining genomes, an all-vs-all ANI analysis was performed with FastANI v1.34 [347]. A newick tree was produced using the ANIclustermap[19] v1.3.0 tool based on the ANI matrix produced by FastANI. A custom script relying on matplotlib[20] v3.7.0 was used to plot the newick tree alongside the ANI heatmap and the N1 + N2 niche metadata. Using ETE Toolkit v3.1.3 [348], all the 2nd-level branches in the tree were identified, then the branch with the bigger number of leaves (genomes) was removed. The procedure was depicted using custom scripts relying on matplotlib and the Bio.Phylo module of Biopython. Then, genomes were further filtered by applying an ANI threshold of 97%, taking as reference both the WCFS1 strain (GCA_000203855.3) and the type strain (*L. plantarum* subsp. *plantarum*, GCA_003076435.1).

4.3.6 Multi-strain reconstruction

For the multi-strain reconstruction and analysis, Gempipe v1.37.4 was used. To reconstruct GSMMs, the program "gempipe recon" was used first. Genomes filtered for quality and taxonomy were given in input with *--genomes* (the built-in genome filtering was bypassed). iLP728u and the associated proteome were used as reference and given in input with *--refmodel* and *--refproteome*, respectively. A file of manual corrections (**Table 4.S3**) was also given in input with *--mancor*.

[19] https://github.com/moshi4/ANIclustermap
[20] https://github.com/matplotlib/matplotlib

Once the draft pan-GSMM was produced, EGCs [107] were detected with the Gempipe API and manually removed. Closed metabolic reactions (bounds 0;0) were opened according to their reversibility. Universally blocked reactions (i.e., reactions blocked in any growth medium), were computed using the *find_blocked_reactions* function from COBRApy [290], opening all the exchange reactions. The biomass equation was deprived of strain-specific components, namely molybdopterin guanine dinucleotide ("mocogdp_c") [337,349], capsular polysaccharide unit ("CPS_LPL2_c") [350] and ribitol teichoic acid unit ("RTAglc_c") [351,352].

The draft pan-GSMM was then given in input to "gempipe derive" to produce strain-specific GSMMs. The CDM with glucose (**Table 4.S2**) was indicated as gap-filling medium with *--media*. Minimum flux through the gap-filling objective was set with *--minflux 0.5*. Strain-specific auxotrophies and alternative C, N, P, and S substrates were computed by activating the *--aux* and *--cnps* options, respectively. With *--cnps_minmed 0.5*, alternative C, N, P, and S substrates were computed on a minimal medium assuring the specified minimal objective value.

4.3.7 Comparison with experimental data

The reference [343] contains a phenotypic screening for a large number of *L. plantarum* strains, where growth was evaluated using different substrates. Among the strains phenotypically screened, those with an available genome assembly were considered. In the original screening, each strain-substrate combination was associated to a color based on the percentage of optical density reached after 48 h on the substrate, compared to that on glucose: dark blue < 10%, light blue 10% < x < 30%, dark green 30% < x < 50%, light green 50% < x < 70%, orange 70% < x < 90%, red x > 90%. The original color code was binarized as follows: no growth (0) if < 10%, growth (1) in all the other cases.

Among the substrates reported, those with an exchange reaction available in the BiGG database [28,56] were considered. Growth simulations were performed on the CDM (**Table 4.S2**), each time replacing glucose with an alternative substrate. Similarly to the procedure adopted in [343], the maximal theoretical growth yield on each substrate was reported in percentage with respect to that on glucose. Then, the same conversion to binary format was applied. Accuracy was computed as (TP + TN) / (TP + TN + FP + FN), where TP, TN, FP and FN are the number of true positive, true negative, false positive and false negative matches, respectively.

4.3.8 Multi-strain clustering

Based on the available strain-specific binary features (presence / absence of reactions, auxotrophies, and ability to grow on alternative substrates) a similarity matrix was built using the Jaccard index. Then, a dendrogram was created parsing the matrix with the Ward's agglomerative clustering [114]. The resulting dendrogram can be called a “phylometabolic” tree, as it relates strains according to their metabolic potential.

The phylometabolic tree was then divided in 8 metabolically coherent clusters, evaluating their Silhouette score [219]. The entire process from the raw binary features to the Silhouette analysis was automated using the *silhouette_analysis* function from the Gempipe API. Finally, the presence / absence of binary features associated with the phylometabolic tree was depicted with the *heatmap_multilayer* function from the Gempipe API, selecting for features that were not constant in all the strains (varying features), and visualizing the data-driven clusters side-by-side the isolation niche metadata.

4.3.9 Extraction of cluster-specific features

To identify the binary metabolic features characterizing each metabolic cluster, the *discriminant_feat* function from the Gempipe API was used. Briefly, the presence of each varying feature was computed relatively to each cluster. Then, features having a relative frequency of ≥0.9 and ≤0.1 in two different clusters were selected and represented in a heatmap.

4.4 Discussion

In the present study, the metabolic biodiversity within the *L. plantarum* species was assessed by means of strain-specific GSMMs, investigating whether the isolation niche of strains was reflecting particular metabolic patterns. This investigation was inspired by previous works [114,115,177], where the metabolic potential of *Escherichia* and *Salmonella* strains was correlated to their environmental niche.

One of such works considered only closed genomes, excluding plasmids [115]. Here, plasmids were retained, as they may shape the metabolic potential of an organism [353]; moreover, even draft genome assemblies were used, filtering them for good quality based both on technical and biological metrics. Starting from quality genomes, strain-specific GSMMs were built based on iLP728 [8], a manually curated model here used as reference, which was expanded with new genes, reactions and metabolites thanks to the hybrid reconstruction method

provided by Gempipe. Reaction content and predictions of GSMMs were used to create a phylometabolic tree, which placed in close proximity strains sharing similar metabolic potential. From the phylometabolic tree, metabolically coherent clusters of *L. plantarum* strains were derived, but these clusters were apparently not consistent with the isolation niche.

A large-scale comparative genomics study of the entire *Lactobacillaceae* family was recently published by Rajput *et al.* [205], where all the available public genomes of the family were downloaded and assigned to a genome-based taxonomy provided by GTDB [354,355]. Then, authors divided *L. plantarum* into 9 different phylogroups based on nucleotide similarity [205], with one phylogroup being particularly divergent, as also reported in earlier works [356,357]. However, the taxonomy of strains belonging to this phylogroup was not questioned. The same phylogroup of divergent genomes was identified here, and it was shown to possess an ANI threshold close to 95% with respect to the type strain, which is conventionally used to delimit the taxonomic perimeter of a species [358]. Most of the genomes in this phylogroup were associated to an NCBI TCS attribute equal to "failed", meaning that taxonomy indicated upon the upload may not be correct. Indeed, for example, strain SRCM101187 reported by [356], and strains SF2A35B, Nizo2264, Nizo1839, Nizo1840, CIF17A2, CIF17A4, CIF17AN2, CIF17A5 and CIF17AN8 reported by [357], were all originally submitted as *L. plantarum*, but the closer type strain suggested by NCBI [178] is *L. argentoratensis*. In order to work on *L. plantarum* genomes with confidence, this outlier phylogroup was here discarded from subsequent analysis.

The quality and taxonomy filtering led to 892 remaining genomes, from which an equal number of strain-specific GSMMs was derived. Clearly, it can be too demanding to comprehensively validate all the models produced. For example, Seif *et al.* [115] modeled 410 strains of *Salmonella*, but validated only 12 of them. This should not be a problem as long as the strain selection encompasses the wider possible range of metabolic variation [29], [56]. In the present study, no set of high throughput, strain-specific phenotypic screenings (e.g., Biolog® Phenotype MicroArray) associated with sequenced strains was found in literature. While a comprehensive validation would have benefited from such type of screenings, here a small dataset of traditional growth trials was used. Using this dataset to compare predictions, the resulting accuracy was acceptable, considering that no manual gap-filling was applied. While it was interesting to compare automated predictions against this dataset, it may not be sufficient to fully validate the deck of 892 strain-specific GSMMs. As a follow up of this study, few representative strains for each data-driven cluster could be

selected to perform a high-throughput substrate phenotyping (e.g., Biolog®), providing better data for the strain-specific GSMMs validation.

In this work, strain-specific biomass components were removed from the biomass equation of the pan-GSMM, as indicated in a recent protocol [48]. Specifically, molybdenum cofactor, teichoic acids and capsular polysaccharides were removed. If they were retained, the introduction of incorrect orphan reactions during the strain-specific gap-filling phase would have been verified, as experienced in early reconstruction attempts.

Molybdenum is a metal ion that is usually uptaken and integrated into the molybdenum cofactor, which is required by some enzymes to operate [359]. Molybdenum cofactor is indispensable for the activity of the nitrate reductase, and indeed its biosynthetic genes are placed together with those for the nitrate reductase [360,361]. However, some LABs are known to live even without molybdenum [359]. While some *L. plantarum* strains, other than WCFS1, are known to encode for the molybdenum cofactor [362,363], some others do not have this feature [337,349]. For this reason, molybdopterin guanine dinucleotide ("mocogdp_c") was removed from the pan-GSMM biomass assembly.

Teichoic acids are an important component of the gram-positive cell wall [364], constituting more than half of its dry weight [351,365]. Their composition is highly variable [351] with the species or even the subspecies [365], and includes carbohydrates such as erythritol, mannitol, N-acetylglucosamine and glucose, in addition to ribitol or glycerol [364]. *L. plantarum* strains are known to be variable in the composition of teichoic acids [351,352], therefore the ribitol teichoic acid unit ("RTAglc_c") was removed from the biomass equation.

Exopolysaccharides, including capsular polysaccharides and slime polysaccharides [350,366], are diffusely produced by LABs [367]. They have many biological roles, for example, protection from desiccation, adhesion to surfaces, and communication with other organisms [350,367]. In *L. plantarum*, composition of capsular polysaccharides is known to be strain-dependent and medium-dependent [350]. Indeed, they are biosynthesized by four operons cps1-4, of which only cps4 is conserved in *L. plantarum*, while the others are located in a "lifestyle adaptation region" [350,367]. Standing on this, the capsular polysaccharide unit ("CPS_LPL2_c") was removed from the biomass equation of the draft pan-GSMM.

This study tried to relate data-driven metabolic clusters to the isolation niche of strains, but failed. One of the main concerns was the manual assignment of

metadata: metabolic clusters might not have been consistent with the isolation niche because of the highly arbitrary and possibly too numerous categories. Therefore, the analysis was repeated using various other niche classification schemas. For example, genome assemblies were associated with a simpler niche metadata based on just 6 categories "diary / meat (fermented)", "plant-derived food / feed", "raw material / environment", "insects", "human / vertebrate", "unclassified"; with this alternative niche metadata, as well as with all the others tried, no significant improvements were noticed, neither in the ANI clustering (**Figure 4.S2**) nor in the phylometabolic tree (**Figure 4.S3**). As reliable metadata are crucial in this kind of analysis, the standardized metadata provided by the GOLD database [368] were tried first. However, only 43.8% of the input genomes were annotated with GOLD, so custom metadata had to be defined for working with all the publicly available *L. plantarum* genomes. It should be noted, however, that particular isolation niches were apparently enriched in some clusters, for example "human/vertebrate" in Cluster_5 or "dairy/meat (fermented)" in Cluster_7, but this cannot constitute a rule as many other strains from the same isolation niche were scattered across the entire phylometabolic tree. This situation resembles what previously reported by Li *et al.* [357], where many strains of dairy and animal origin (but not all of them) were reported in two different clades of a *L. planturm* phylogenetic tree. In the future, to further assess other niche classification schemas, the FoodEx2 standard classification system for foods, developed by EFSA, could be tried (https://www.efsa.europa.eu/en/data/data-standardisation).

Another possible explanation for the unseen correlations between isolation niche and metabolic potential could rely on the features evaluated. The phylometabolic tree used here considered just three kinds of binary features: presence / absence of reactions, auxotrophies, and ability to utilize substrates. However, a phylometabolic tree can accommodate any number of feature layers, including those not directly related to GSMMs. Indeed, given a database of interesting genetic features, such as factors for virulence or antibiotic resistance, they could be searched in each genome via blast-like alignments, and their presence reported in the tree [116]. Moreover, additional features stemming from metabolic modeling could be added as well, for example the potential production of flavours or short-chain fatty acids [369], which could also be provided as semi-quantitative data rather than binary [136,168]. In this sense, the approach presented here could be extended for specific application scopes, for example the evaluation of the strain-specific probiotic potential [370]. I this scenario, the ability to release beneficial or detrimental molecules in the gut could be evaluated under different human dietary regimens [369] and, together with the detection of specific gut adhesion factors and the absence of pathogenicity factors, could be condensed in a single "probioticity score" [370].

While hampered by the technical limitations outlined above, a correlation between isolation niches and metabolic potential could simply not be possible for this species. Early works reported a successful clusterization of *L. plantarum* strains according to the food of origin, except for those isolated from feces, scattered among food clusters as might be expected [343]. However, in recent papers exploiting the impressive amount of genomes accumulated over the years, the achievability of this type of clustering was questioned. Choi *et al.* [371] and Evanovich *et al.* [372], for example, were not able to relate the gene content of genomes to the isolation niche. Li *et al.* [357] built a phylogenetic tree in which the niche categories were dispersed across the identified clades, with some exceptions (reported above). This behaviour was well summarized by Fidanza and colleagues [373], stating that the process of specialization of a clade for a niche, consisting in the decay of useless genes and concomitant enrichment of useful traits, did not occur in *L. plantarum*. Indeed, while other LAB species could be clustered according to their environmental niche (e.g., *Lactobacillus gasseri*, *Lactobacillus jensenii* and *Lactobacillus reuteri* [373]), *L. plantarum* retained most of the niche-specific genetic features, leading its strains to potentially colonize many different habitats at the same time [373]. For this reason, *L. plantarum* has been described as a "nomadic" species [338,373,374].

The reconstruction process provided by Gempipe involves the addition of reactions that are not part of the model used as reference, copying them from a reference-free reconstruction which is based on a "universe" derived from genes and reactions stored in BiGG [56] (see **Chapter 2**). While supported by alignment scores, the introduction of these exogenous reactions could not be coherent with the physiology of *L. plantarum*. For example, many strain-specific GSMMs, in particular those associated to Cluster_8, underwent the addition of a transport reaction for sucrose described as a proton symport ("SUCRt2"). While different mechanisms of sucrose transport are known in bacteria [375], in *L. plantarum* (at least in the reference strain WCFS1, included in Cluster_2) this substrate is known to be uptaken with a PTS transporter [342,376]. The verification of such exogenous reactions should be part of the curation of the draft pan-GSMM, a manual process that was not included here, which would definitely have been a better input for the method implemented in this study. However, even the draft pan-GSMM adopted was able to account for several strain-specific phenotypes which seemed to be realistic for *L. plantarum*, including the self-production of thiamine [377], the catabolic activity for arabinose [378], and the presence of molybdenum cofactor [349]. Lastly, the detection of features characterizing each cluster would have benefited from the definition of reaction modules, which collect all the reactions needed to

complete a pathway, similarly to what is implemented in KEGG [30] with its "KEGG Modules" [379]; with this feature, for example, reactions composing the catabolic pathway for rhamnose ("ARAB_Lt", "ARAI" and "RBK_L1", see Results) would have been lumped in single module, whose presence / absence would have been evaluated directly.

In conclusion, the approach presented here looks promising: genes of strains are *structured* and *contextualized* into a GSMMs, enabling one to move from a general description of which functional categories are enriched in a genome (like in [357], where the Clusters of Orthologous Groups (COG) categories were used), to a list of metabolic tasks that a strain is actually capable to carry on. Despite limitations presented, this approach can be considered an initial step towards the rational selection of strains where, given a defined bioprocess and a wide catalog of strains, the optimal strain can be selected based on its predicted phenotype.

4.5 Supplementary Material

4.5.1 Supplementary Figures

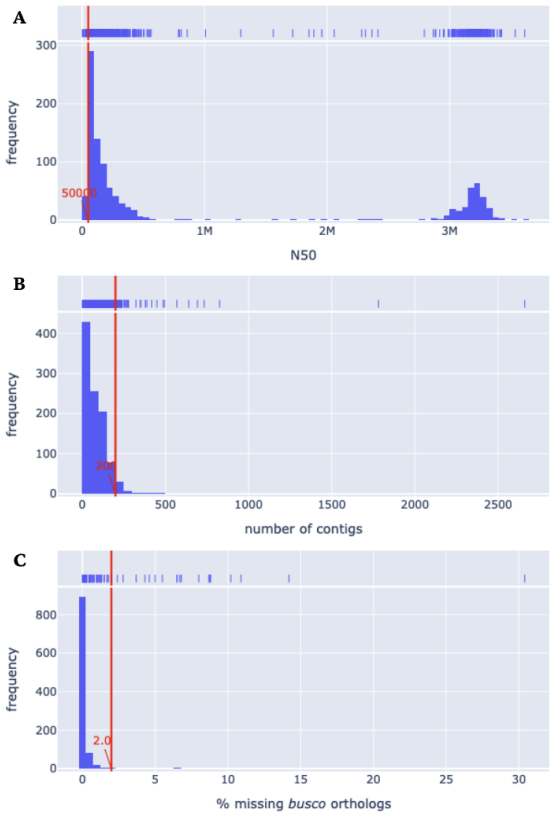

Figure 4.S1. Histograms of the metrics used to filter low-quality genome assemblies. In red are shown the thresholds applied. **A)** N50. **B)** Number of contigs. **C)** Percentage of missing BUSCO orthologs [179] according to the database *lactobacillales_odb10*.

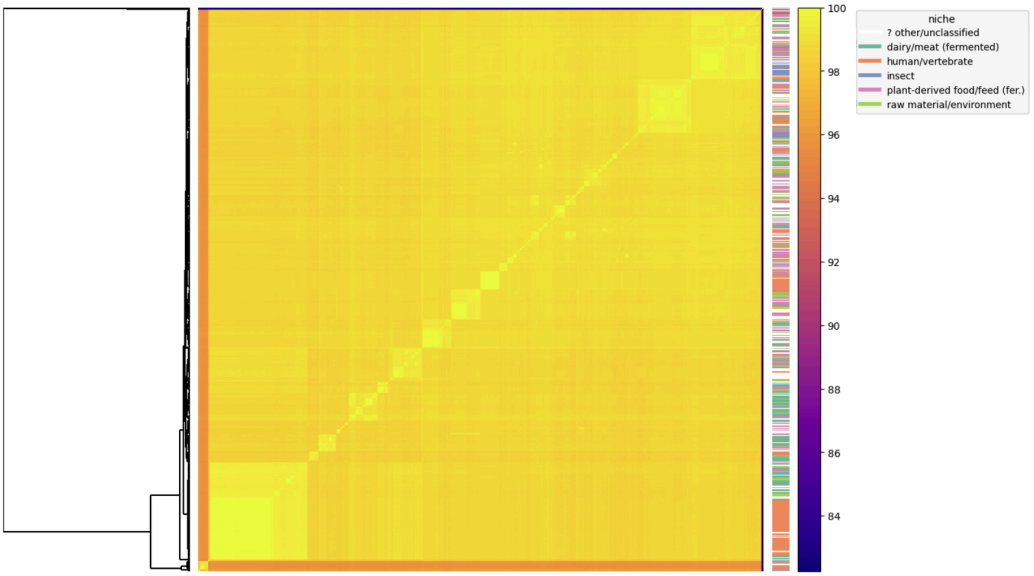

Figure 4.S2. ANI analysis of 912 good-quality assemblies of *L. plantarum*. On the left, a dendrogram based on the ANI matrix is shown. A colormap indicating the niche metadata assigned to each genome assembly is reported on the right. This figure replicates **Figure 4.2**, with the exception of a simplified niche classification schema.

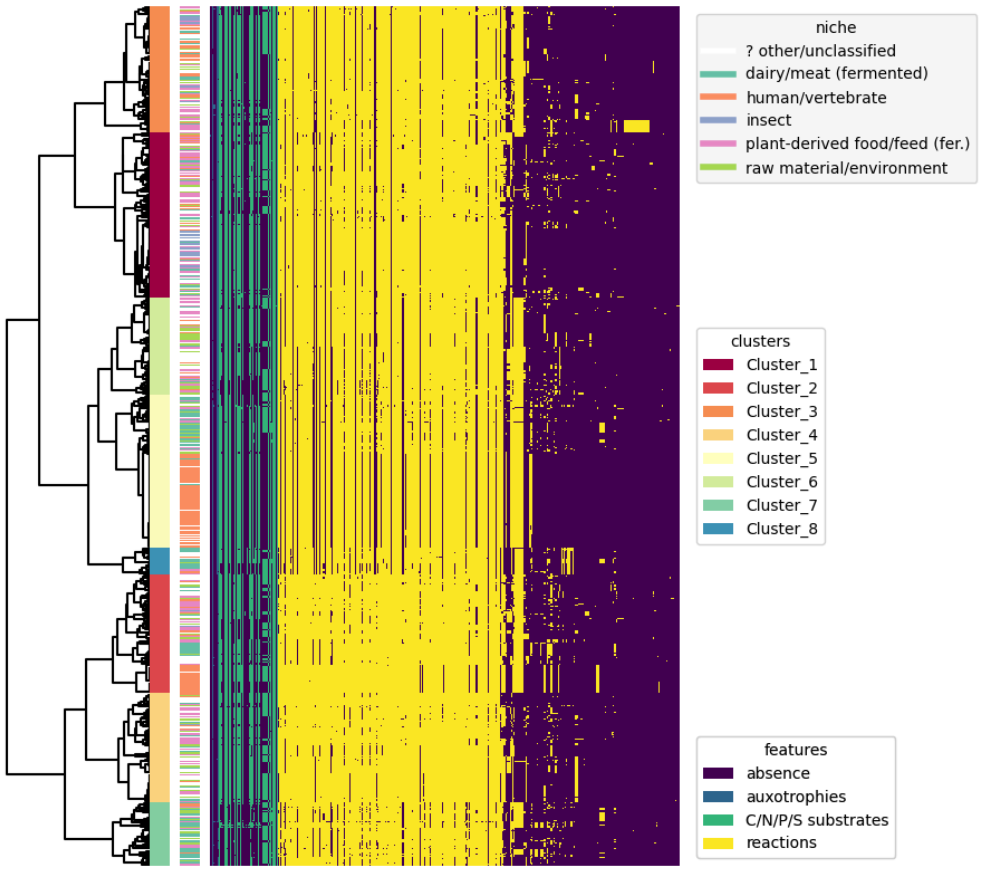

Figure 4.S3. Phylometabolic tree and metabolic features of the associated strains. Cluster attribute and niche metadata are paired for comparison. Only varying features are shown. This figure replicates **Figure 4.7**, with the exception of a simplified niche classification schema.

4.5.2 Supplementary Tables

Table 4.S1. Experimental uptake and secretion rates in a CDM for *L. plantarum* WCFS1 [3,132]. Data taken from [8] for 25 mM glucose at dilution rate 0.3 h^{-1}. "exr": ID of the exchange reaction. "lb" and "ub": lower and upper bound of the exchange reaction, respectively, expressed as flux (mmol/gDW/h). Exchange reactions not listed here but appearing in **Table 4.S2** are set to (-1000, 1000); all the other exchange reactions in the model are set to (0, 1000).

	substrate	exr	rate	error	lb	ub
1	Glucose	EX_glc__D_e	-7.410	0.339	-7.749	-7.071
2	Citric acid	EX_cit_e	-0.600	0.063	-0.663	-0.537
3	Lactate	EX_lac__D_e	12.150	0.495	11.655	12.645
4	Pyruvate	EX_pyr_e	0.060	0.009	0.051	0.069
5	Formate	EX_for_e	1.170	0.513	0.657	1.683
6	Acetate	EX_ac_e	2.040	0.444	1.596	2.484
7	Ethanol	EX_etoh_e	0.420	0.114	0.306	0.534
8	Succinate	EX_succ_e	0.720	0.015	0.705	0.735
9	Alanine	EX_ala__L_e	-0.030	0.006	-0.036	-0.024
10	Arginine	EX_arg__L_e	-0.036	0.006	-0.042	-0.030

11	Aspartate	EX_asp__L_e	-0.147	0.021	-0.168	-0.126
12	Cystine	EX_cys__L_e	-0.021	0.009	-0.030	-0.012
13	Glutamate	EX_glu__L_e	-0.096	0.009	-0.105	-0.087
15	Glycine	EX_gly_e	-0.078	0.021	-0.099	-0.057
16	Isoleucine	EX_ile__L_e	-0.042	0.003	-0.045	-0.039
17	Leucine	EX_leu__L_e	-0.066	0.027	-0.093	-0.039
19	Lysine	EX_lys__L_e	-0.051	0.006	-0.057	-0.045
20	Phenylalanine	EX_phe__L_e	-0.027	0.003	-0.030	-0.024
21	Proline	EX_pro__L_e	-0.054	0.030	-0.084	-0.024
22	Serine	EX_ser__L_e	-0.474	0.318	-0.792	-0.156
24	Threonine	EX_thr__L_e	-0.210	0.054	-0.264	-0.156
25	Tyrosine	EX_tyr__L_e	-0.069	0.033	-0.102	-0.036
26	Valine	EX_val__L_e	-0.096	0.015	-0.111	-0.081

Table 4.S2. Representation of the CDM for *L. plantarum* WCFS1 [3,132]. "exr": ID of the exchange reaction. "lb": lower bound of the exchange reaction, expressed as concentration (mmol/L). Upper bound of all the exchange reactions in the model is set to 1000. Exchange of water ("EX_h2o_e") is left unconstrained. Concentrations of oxygen and carbon dioxide are approximated to those in pure water at 1 atm and 37 °C. Concentration of protons represents pH 6.

	substrate	exr	lb
1	Glucose	EX_glc__D_e	-25.0000
2	Protons	EX_h_e	-0.0010
3	Oxygen	EX_o2_e	-0.2097
5	Carbon dioxide	EX_co2_e	-22.7273
6	Alanine	EX_ala__L_e	-2.6939
8	Arginine	EX_arg__L_e	-0.7176
9	Aspartate	EX_asp__L_e	-3.1555
10	Cysteine	EX_cys__L_e	-1.0730
12	Glutamate	EX_glu__L_e	-3.3984
13	Glycine	EX_gly_e	-2.3312
14	Histidine	EX_his__L_e	-0.9668
15	Isoleucine	EX_ile__L_e	-1.6009
16	Leucine	EX_leu__L_e	-3.6211
17	Lysine	EX_lys__L_e	-3.0098
18	Methionine	EX_met__L_e	-0.8378
19	Phenylalanine	EX_phe__L_e	-1.6647
20	Proline	EX_pro__L_e	-5.8629
21	Serine	EX_ser__L_e	-3.2352
22	Threonine	EX_thr__L_e	-1.8889
23	Tryptophan	EX_trp__L_e	-0.2448
24	Tyrosine	EX_tyr__L_e	-1.3798
25	Valine	EX_val__L_e	-2.7743
26	Pantothenate	EX_pnto__R_e	-0.0042
28	Biotin	EX_btn_e	-0.0102
29	Nicotinate	EX_nac_e	-0.0081
30	4-Aminobenzoate	EX_4abz_e	-0.0729
31	Pyridoxamine	EX_pydam_e	-0.0207
32	Pyridoxine	EX_pydxn_e	-0.0097
33	Riboflavin	EX_ribflv_e	-0.0027
35	Thiamine	EX_thm_e	-0.0030
36	Folic acid	EX_fol_e	-0.0023
37	Adenine	EX_ade_e	-0.0740
38	Guanine	EX_gua_e	-0.0662
39	Inosine	EX_ins_e	-0.0186
40	Orotate	EX_orot_e	-0.0320
41	Thymidine	EX_thymd_e	-0.0206

42	Uracil	EX_ura_e	-0.0892
43	Xanthine	EX_xan_e	-0.0657
44	Phosphate	EX_pi_e	-42.4816
45	Sodium	EX_na1_e	-12.1907
46	Citrate	EX_cit_e	-2.6526
47	Ammonium	EX_nh4_e	-5.3184
48	Manganese	EX_mn2_e	-0.1348

Table 4.S3. Manual correction rules to be applied during the reference expansion phase.

blacklist.CO2FO	charge.gtca3_45_BS:-45	charge.octeACP:0
blacklist.ACPpds	charge.gtca2_45_BS:-45	charge.hdeACP:0
blacklist.IBHH	charge.gtca1_45_BS:-45	charge.tdeACP:0
blacklist.LIPO3S24_BS	charge.nadphx__R:-4	charge.3hcmrs7eACP:0
blacklist.PHET	charge.nadhx__R:-2	charge.t3c7mrseACP:0
blacklist.SMIA1	charge.nadphx__S:-4	charge.dgal6p:-2
blacklist.SMIA1abc	charge.nadhx__S:-2	charge.lac6p:-2
blacklist.SMIA2abc	charge.metglcur:-1	charge.2mb_p:-2
blacklist.SMIB1	charge.2agpg120:-1	charge.iv_p:-2
blacklist.FEENTERES	charge.pg120:-1	charge.ib_p:-2
blacklist.ENTERES2	charge.pg180:-1	charge.ib_p:-2
charge.3hasp__L:-1	charge.pgp120:-3	charge.hethmpp:-2
charge.adsel:-2	charge.pgp140:-3	charge.acmum6p:-3
charge.pphos:-2	charge.pgp141:-3	charge.fadh2:-2
charge.4hoxoh:-1	charge.pgp160:-3	charge.oc2coa:-4
charge.5cmhmsa:-2	charge.pgp161:-3	charge.dccoa:-4
charge.ggaptn:-1	charge.pgp180:-3	charge.td2coa:-4
charge.ggala_B:-1	charge.pgp181:-3	charge.gly_asp__L:-1
charge.hkmpp:-2	charge.malt1p:-2	charge.prohisglu:-1
charge.4h2kpi:-2	charge.fa11:-1	charge.serglugly:-1
charge.murein3px3p:-4	charge.fa12:-1	charge.ala_L_glu__L:-1
charge.murein5px4px4p:-6	charge.fa3:-1	charge.gly_glu__L:-1
charge.vacc:-1	charge.fa4:-1	charge.26dapime:-1
charge.fe3dcit:-3	charge.fa6:-1	charge.dag_MRSA:2
charge.pheme:-2	charge.fa1:-1	charge.23ddhb:-1
charge.ah6p__D:-2	charge.gcvHL_ADPr:-3	charge.sbzcoa:-5
charge.fecpp3:-4	charge.gcvHL_nhLA:-2	charge.2dhguln:-1
charge.cpp3:-4	charge.gcvhlipl:-5	charge.dr5p:-2
charge.4hadnt:-1	charge.gcvHalip:-4	charge.s17bp:-4
charge.5cohe:-3	charge.gcvHdhlip:-5	charge.fmnh2:-2
charge.2ohed:-2	charge.alpro:1	charge.tag6p__D:-2
charge.24dhhed:-2	charge.g3pg:-1	charge.galctr__D:-2
charge.2hh24dd:-2	charge.fc1p:-2	charge.ppgpp:-6
charge.op4en:-1	charge.strcoa:-4	charge.g6p_A:-2
charge.3mgcoa:-5	charge.tdecoa:-4	charge.h2isocapcoa:-4
charge.salc:-1	charge.cmcbtt:1	charge.mmet:1
charge.hmpscoa:-5	charge.fcmcbtt:4	charge.isocapcoa:-4
charge.2maacoa:-4	charge.cdigmp:-2	charge.isocaprecoa:-4
charge.23dhbzs3:-1	charge.S2hglut:-2	charge.lipoamp:-1
charge.23dhbzs:-1	charge.1hdec9eg3p:-2	charge.lipopb:1
charge.bz12diol:-1	charge.pa141:-2	charge.man6pglyc:-3
charge.appl:1	charge.pa140:-2	charge.2me4p:-2
charge.3padsel:-4	charge.pa120:-2	charge.pa161:-2
charge.4hoxpac:-1	charge.anhgm3p:-2	charge.pa180:-2
charge.2agpg180:-1	charge.ocdceap:-1	charge.pa181:-2
charge.1agpg161:-1	charge.ocdcap:-1	charge.uagmda:-4
charge.2agpg161:-1	charge.hdceap:-1	charge.tagdp__D:-4

charge.murein4px4p:-4 charge.murein5px4p:-4 charge.murein5p4p:-4	charge.hdcap:-1 charge.ttdceap:-1 charge.ddcap:-1	charge.pgp160:-3 charge.pg160:-1 charge.ppi50:-51

charge.hepdp:-3 charge.tsul:-2 charge.scys__L:-1 charge.2shchc:-2 charge.udpacgal:-2 charge.uaagmda:-4 charge.frulysp:-1 charge.frulys:1 charge.pa_LLA:-200 formula.fcmcbtt:C33H48N5O13Fe formula.octeACP:C18H33OX formula.isocaprecoa:C27H40N7O17P3S formula.uagmda:C87H139N7O23P2 formula.dmlz:C13H18N4O6 formula.uaagmda:C95H152N8O28P2 formula.ibtol:C4H10O formula.selhcys:C4H9NO2Se	formula.t3c7mrseACP:C14H23OX formula.fadh2:C27H33N9O15P2 formula.gly_pro__L:C7H12N2O3 formula.met_L_ala__L:C8H16N2O3S formula.fmnh2:C17H21N4O9P formula.bglycogen:C6H10O5 formula.gly_cys:C5H10N2O3S formula.gly_leu:C8H16N2O3 formula.gly_phe:C11H14N2O3 formula.gly_tyr:C11H14N2O4 formula.isocapcoa:C27H42N7O17P3S formula.val__D:C5H11NO2 formula.fecpp3:C36FeH32N4O8 formula.cpp3:C36H34N4O8 formula.hdeACP:C16H29OX formula.tdeACP:C14H25OX formula.3hcmrs7eACP:C14H25O2X

reaction.CMCBTFU:fcmcbtt_c --> cmcbtt_c + fe3_c
reaction.3HAD141:3hcmrs7eACP_c --> h2o_c + t3c7mrseACP_c
reaction.AMPEP11:gly_pro__L_c + h2o_c --> gly_c + pro__L_c
reaction.AMPEP14:h2o_c + met_L_ala__L_c <=> ala__L_c + met__L_c
reaction.FMNRy:fmn_c + h_c + nadph_c --> fmnh2_c + nadp_c
reaction.GLYCYSAP:gly_cys_c + h2o_c <=> cys__L_c + gly_c
reaction.GLYCYSabc:atp_c + gly_cys_e + h2o_c --> adp_c + gly_cys_c + pi_c + h_c
reaction.GLYLEUAP:gly_leu_c + h2o_c <=> gly_c + leu__L_c
reaction.GLYLEUtr:atp_c + gly_leu_e + h2o_c --> adp_c + gly_leu_c + pi_c + h_c
reaction.GLYPHEAP:gly_phe_c + h2o_c <=> gly_c + phe__L_c
reaction.GLYPHEtr:atp_c + gly_phe_e + h2o_c --> adp_c + gly_phe_c + pi_c + h_c
reaction.GLYTYRAP:gly_tyr_c + h2o_c <=> gly_c + tyr__L_c
reaction.GLYTYRabc:atp_c + gly_tyr_e + h2o_c --> adp_c + gly_tyr_c + pi_c + h_c
reaction.ICCT:h2o_c + isocapcoa_c --> coa_c + isocap_c
reaction.NTPP10:ditp_c + h2o_c --> dimp_c + ppi_c
reaction.NTD12:dimp_c + h2o_c --> din_c + pi_c + h_c
reaction.SEAHCYSHYD_1:h2o_c + seahcys_c <=> adn_c + selhcys_c
reaction.DAAD12:fad_c + h2o_c + val__D_c --> 3mob_c + fadh2_c + nh4_c
reaction.FECPP3O:2.0 amet_c + fecpp3_c --> 2.0 co2_c + 2.0 dad_5_c + 2.0 met__L_c + pheme_c
reaction.CPP3O1:cpppg3_c + 3.0 fad_c --> cpp3_c + 3.0 fadh2_c

Chapter 5 | Topology curation of a genome-scale metabolic model for *Thauera* sp. *sel9*

5.1 Introduction

Polyhydroxyalkanoates (PHAs) are promising bioplastic polymers, stored by prokaryotes in granules as energy reserve [380]. Isolated bacteria are known to produce PHAs from noble sugars like glucose or saccharose [381]. Alternatively, microbial mixed cultures (MMCs) are known to produce PHAs from volatile fatty acids (VFAs) [382,383] thanks to the activity of organisms mainly belonging to the genera *Azoarcus*, *Paracoccus* and *Thauera* [384,385]. Recently, Andreolli and colleagues [12] isolated, for the first time, a PHAs-producing *Thauera* strain from MMCs. This strain, named *T.* sp. *Sel9*, was isolated from a sequencing batch reactor (SBR, a type of activated sludge plant for wastewater treatment), and it is able to produce PHAs while growing on a mineral medium supplemented with one or more VFAs (e.g., valerate or butyrate) [12].

Because of the potential production of biopolymers from waste, *T.* sp. *Sel9* is considered of biotechnological interest. However, to fully appreciate and exploit its potential, its metabolism must be first characterized. To pursue this aim, genome-scale metabolic models (GSMMs) represent ideal tools. Here, a preliminary draft model for *T.* sp. *Sel9* was generated, and its ration network topology was curated by means of a high-throughput substrate screening system. Specifically, the Biolog® Phenotype MicroArray (PM) system was used, where growth tests for 190 alternative carbon sources were run simultaneously. While the resulting draft is still not ready for prospective use, this preliminary work represents a necessary starting point for characterizing the *T.* sp. *Sel9* metabolism. In the future, a finalized GSMM could be used as hypothesis generation platform aimed at improving the PHAs yield.

5.2 Results

5.2.1 Reconstruction of *T.* sp. *Sel9* draft GSMM

All available genome assemblies for the genus *Thauera* were downloaded from NCBI (**Table 5.S1**). Starting from 119 genomes, 32 were retained based on quality filters for both technical and biological metrics. An all-vs-all average

nucleotide identity (ANI) was computed for the remaining genomes (**Figure 5.1**), and the species most similar to *T.* sp. *Sel9* resulted in *T. butanivorns* (ANI 93.7%), followed by *T. linaloolentis* (ANI 87.8%).

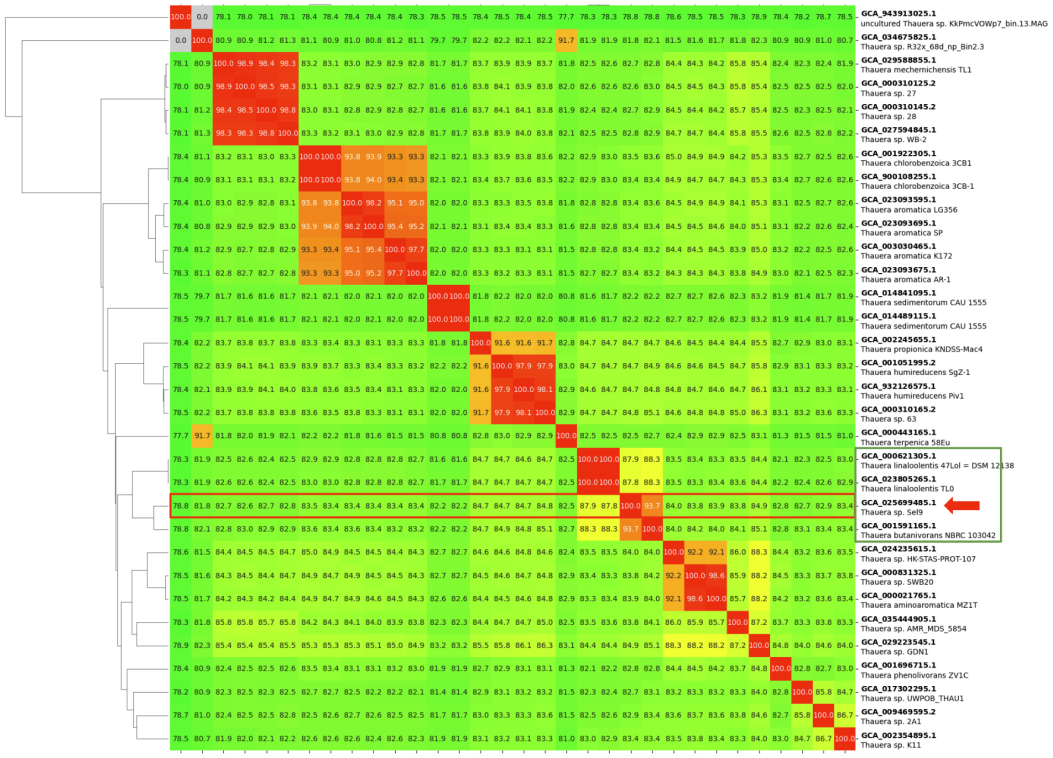

Figure 5.1. ANI plot produced by ANIclustermap[21] for 32 quality-filtered genome assemblies associated on NCBI with the genus *Thauera* (taxid 33057). On the right, organisms and accessions are reported. The row belonging to *T.* sp. *Sel9* is highlighted by a red rectangle and arrow. The green rectangle contains the organisms most similar to *T.* sp. *Sel9*: *T. linaloolentis* 47Lol, *T. linaloolentis* TL0 and *T. butanivorns* NBRC 103042.

To reconstruct the draft GSMM for *T.* sp. *Sel9*, Gempipe (see **Chapter 2**) was used without providing a reference model. Even if the reconstruction target was exclusively *T.* sp. *Sel9*, in order to fully exploit the gene recovery feature provided by Gempipe, all the quality-filtered genomes were selected as Gempipe inputs.

The first phase of gene recovery, evaluating protein broken in two pieces by premature stop codons (assumed as sequencing or assembling errors), resulted in 0 recovered sequences. The second phase, dealing with genomic regions not considered by the gene caller, resulted in 8 recovered sequences, 5 of which functionally annotated with KEGG Orthology codes [216]: (i) K02231, an

[21] https://github.com/moshi4/ANIclustermap

adenosylcobinamide kinase, part of the cobalamin coenzyme biosynthetic pathway; (ii) K02914, a large subunit ribosomal protein, not modeled in this kind of GSMMs; (iii) K11688, a C4-dicarboxylate-binding protein, part of a transporter for malate, succinate, and fumarate; (iv) K18365, an aldolase involved in benzoate and xylene degradation; (v) K01477, an allantoicase involved in xanthine degradation to urea. The third phase of gene recovery, dealing with eventual overlapping genes in the same genomic regions, resulted in 1 recovered sequence: K03574, classified as "DNA repair and recombination proteins", not modeled in this kind of GSMMs.

5.2.2 Biolog®-based manual curation

Biolog® phenotypic screenings were performed using PM1 and PM2A plates, incubating *T.* sp. *sel9* for 10 days and recording metabolic activity each 24 h. The data points of each well were converted to binary format (substrate consumed / not consumed) to enable comparison with simulations. Binary Biolog® data were used not only to evaluate accuracy in predicting substrate utilization, but also to drive manual curation of the reaction network topology.

Before starting the manual curation, the unedited draft model built with Gempipe (sel9_draft_v1) was subjected to an early Biolog® simulation together with a draft model built with CarveMe v1.6.1 [44] for comparison. These early simulations, based on the same chemically defined medium (CDM) [12], showed 18 false positive mismatches in the CarveMe-based draft that were not present in sel9_draft_v1 (**Figure 5.S1**).

Then, based on the binary Biolog® data, the Gempipe API was extensively used to manually curate sel9_draft_v1, leading to an improved version named sel9_draft_v2. This version presented 25 new metabolic reactions and 21 new transport reactions, 10 and 4 of which, respectively, were not associated with a gene-to-reaction association (GPR), remaining classified as "orphan" (**Figure 5.2A**). Moreover, Biolog® data revealed the presence of 5 false-positive transport reactions due to misannotated transport genes: they were removed from sel9_draft_v2. After these edits, 19 mismatches were corrected, and sel9_draft_v2 globally accounted for 1137 genes, 1807 reactions, and 1648 metabolites. Repeating Biolog® simulations, the accuracy increased from 80% to 93% (**Figure 5.2B** and **C**).

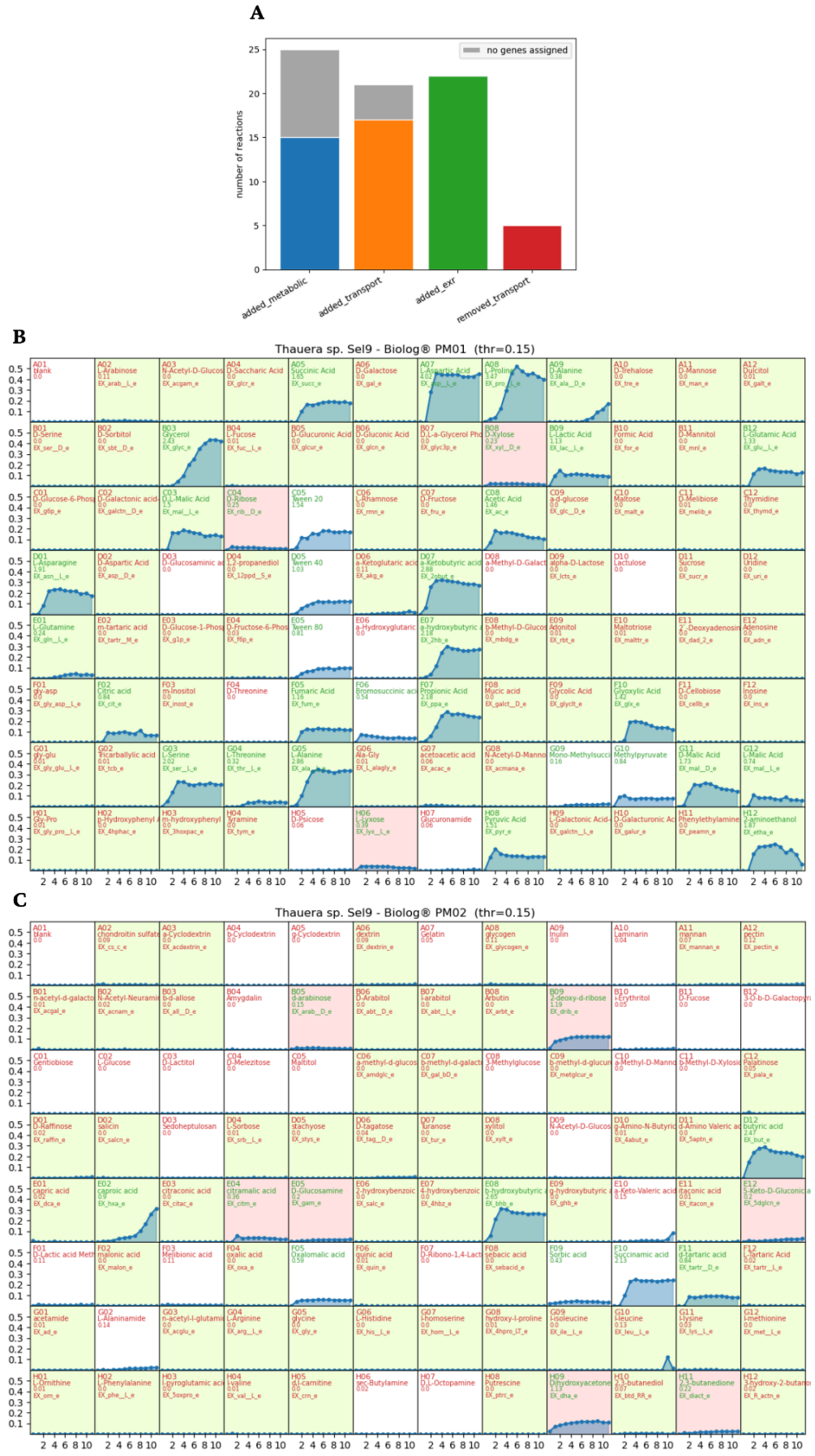
A
no genes assigned
number of reactions
added_metabolic
added_transport
added_exr
removed_transport
B
Thauera sp. Sel9 - Biolog® PM01 (thr=0.15)
C
Thauera sp. Sel9 - Biolog® PM02 (thr=0.15)

Figure 5.2. Biolog®-based network topology manual curation. **A)** Synthesis of the manual curation. "added_metabolic": number of metabolic reactions added; "added_transport": number of transport reactions added; "added_exr": number of exchange reactions added; "removed_transport": number of false-positive transport reactions removed. The proportion of orphan reactions is indicated in grey. **B)** Comparison between experimental and sel9_draft_v2-simulated binary Biolog® data for PM1 plate and **C)** PM2A plate. In each box (i.e., each well) the normalized Biolog® signal is represented. Red boxes represent mismatches with the experimental data (either FP or FN); green boxes represent matches (either TP or TN). White boxes represent substrates not found in the BiGG database [28]. Each box reports the following information (from top to bottom): well ID, substrate name, area under the curve (blue area), corresponding exchange reaction in BiGG [28]. Information is displayed in green when experimental growth was verified (according to thresholds, see 5.3 Methods), otherwise in red.

5.3 Methods

5.3.1 Genome download, filtering and ANI

A set of 119 genome assemblies belonging to the genus *Thauera* (**Table 5.S1**) was downloaded from NCBI [178] on 4st February 2024 with the NCBI Datasets command-line interface v16.1.1 [386] using the subprogram *datasets download genome taxon* with parameters *33057 --include genome --assembly-source GenBank --assembly-version latest*.

For each downloaded genome, N50 and number of contigs were computed with SeqKit v2.2.0 [202] using *stats --tabular --basename --all*. Genes were predicted with Prodigal v2.6.3 [151] run through Prokka v1.14.6 [196] with options *--noanno --norrna --notrna*. Resulting protein sequences were evaluated with BUSCO v5.4.0 [179] using parameters *--mode proteins --lineage_dataset betaproteobacteria_odb10*. Downloaded genomes were discarded when having either more than 2% missing BUSCO orthologs (out of 569 orthologs contained in the database *betaproteobacteria_odb10*), more than 200 contigs, or N50 lower than 50000.

Using the remaining genomes, an all-vs-all ANI analysis was performed with FastANI v1.34 [347]. A dendrogram and the associated heatmap were automatically produced by the ANIclustermap[22] v1.3.0 tool based on the ANI matrix produced by FastANI.

[22] https://github.com/moshi4/ANIclustermap

5.3.2 Biolog® screenings

Thauera Sel9^{T} cells were re-cultured in TSB medium (Oxoid) for 48 h (27 ℃, 200 rpm), harvested (4800 rpm, 10 min), and washed twice in physiological solution (0.9% NaCl). Cells were resuspended in a sterile capped tube containing 2 mL sterile water. Cell concentration was adjusted to 0.2 absorption at 590 nm in a 1 cm cuvette on a Cary 60 UV-Vis spectrophotometer (Agilent Technologies).

Inoculating fluid was prepared as follow: 20 mL of CDM (modified Brunner medium described in [12]), 0.24 mL Dye mix D (100x, Biolog), 3.26 mL sterile water, and 0.5 mL cell suspension (0.2 absorption at 590 nm in a 1 cm cuvette). Using a multichannel pipette, 100 µL per well of inoculating fluid was added to each PM1 and PM2A plate followed by incubation for 10 days (27 ℃, 200 rpm). The colour changes were visually inspected every 24 h and optical density (OD) was measured using a Biotek synergy™ Neo2 Hybrid Multimode Reader (BioTek) at 590 nm and 750 nm with Corning™ Costar Half-Area well plate with flat bottom wells (Product number 3696) as plate type.

5.3.3 Biolog® data analysis

While OD_{590} provides the maximum absorption for colour changes in redox dye, OD_{750} measures turbidity and not colour. Therefore, a corrected optical density, defined as $OD_{corr} = OD_{590} - OD_{750}$, was computed for each well to correct for any interference from turbidity eventually originating from cell growth or chemical precipitation. Then, for each well, a normalized optical density was computed, defined as $OD_{norm} = OD_{corr} - OD_{corr}^{A01}$, where the latter term is the corrected optical density in the blank well (A01) of each plate ($OD_{norm} < 0$ where set 0).

The 11 data points for each plate were used to compute their area under the curve (AUC) using the *trapz* function from the numpy[23] v1.26.2 library, which computes the definite integral of the data points using the trapezoidal rule. When AUC ≥ *t*, where *t* was an arbitrary threshold equal to 0.15, the substrate was considered catabolized by the organism. In this case, the binary outcome corresponded to 1 (the organism can grow on that substrate), otherwise 0. Biolog® simulations were performed using the *biolog_preview* function from the Gempipe API (see **Chapter 2**), which provides a table of binary growth simulations with one well per row. For each well, the comparison felt into four alternative categories: true positive match, when experimental and simulated binary outcomes agreed on growth; true negative match, when they agreed on growth absence; false positive mismatch, when growth was predicted but not

[23] https://numpy.org

experimentally observed; false negative mismatch, when no growth was predicted while it was experimentally observed. Accuracy was defined as (TP + TN) / (TP + TN + FP + FN), where TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives, respectively.

5.3.4 Draft model reconstruction and curation

Gempipe was used to reconstruct a draft GSMM for *T.* sp. *Sel9*. First, "gempipe recon" was used with all the *Thauera* quality-filtered genomes in input. With this program, strain-specific genes were annotated and grouped in clusters. Without providing any reference model, clusters were used to perform a reference-free reconstruction using the gram negative universe (*-s neg*). Importantly, gene clusters were also functionally annotated using EggNOG-mapper [182]. At the end of this procedure (see **Chapter 2** for more details) a draft pan-GSMM was produced, encoding gene clusters instead of genes. This was subsequently manually curated using the Gempipe API. The updated pan-GSMM was finally used as input for "gempipe derive" with default parameters to generate strain-specific GSMMs, including the one for *T.* sp. *Sel9*.

The procedure for manually correcting the topology was the following. For each growth-positive C source (having an associated exchange reaction in the BiGG database [56]), flux-balance analysis (FBA) was performed simulating growth of the universe, pan-GSMM, or *T.* sp. *Sel9* draft model. Growth was simulated on the CDM used in [12] supplemented with the C source, representing the medium by setting the exchange reactions as substrate concentrations in *mmol/L* [68]. Lower bounds were constrained as follow: PO_4^{3-} ("EX_pi_e") -28.35, K^+ ("EX_k_e") -11.16, SO_4^{2-} ("EX_so4_e") -4.59, NH_4^+ ("EX_nh4_e") -7.57, Ca^{2+} ("EX_ca2_e") -0.34, Mg^{2+} ("EX_mg2_e") -0.81, Cl^- ("EX_cl_e") -0.68. Other trace elements entering directly into the biomass equation (Fe^{2+}, Zn^{2+}, Mn^{2+}, Cu^{2+}, Co^{2+}) were provided without constraints. No exchange reaction or biomass requirement for Na^+, MoO_4^{2-} and Ni^{2+} were present in the gram negative universe. No vitamins were provided, as no auxotrophies were detected. Aerobic conditions were represented by allowing an unconstrained input of water, protons, and oxygen. Lower bound of the variable C source was set to -1000. Upper bound of exchange reactions was never constrained (1000).

First, growth was simulated with the universe, from which some catabolic routes were missing. For example, to catabolize 2-hydroxybutyrate ("2hb"), an exchange reaction, a proton symporter, and a dehydrogenase were inserted to allow the conversion to 2-oxobutanoate ("2obut"). Similarly, new reactions were added to the universe to represent catabolic routes for D-tartrate and citramalate. New reactions and metabolites were added using the *add_reaction*

and *add_metabolite* functions from the Gempipe API, respectively. Missing catabolic routes were retrieved from the KEGG Reaction [30] database, applying mass and charge balances if needed.

Second, FBA for a specific growth-positive C source was performed with the *T.* sp. *Sel9* draft model and, if no growth was predicted, gap-filling reactions were proposed using the *perform_gapfilling* function from the Gempipe API, setting the universe as the reaction source when the draft pan-GSMM was not sufficient. Each reaction proposed was contextualized in a hand-drawn Escher map [35] representing the gram negative universe content. Using the function *import_from_universe* from the Gempipe API, gap-filling reactions were introduced into the pan-GSMM and not into the *T.* sp. *Sel9* draft model, providing a gene-reaction-rule (GPR) composed of gene clusters ("Cluster_*"). Gap-filled reactions were propagated to the *T.* sp. *Sel9* draft model by using the *ss_preview* function from the Gempipe API, which updates the strain-specific model based on the pan-GSMM and gene presence / absence matrix (PAM, see **Chapter 2**).

In order to compose GPRs with correct gene clusters ("Cluster_*"), the *query_pam* function from the Gempipe API was used. This function enabled searches among the functional annotations provided by EggNOG v6.0 [218], (assigned to gene clusters as part of the Gempipe reconstruction process) presenting results as a PAM, with cluster IDs in row and strains in column. Searches involved different attributes of the functional annotation, including: KEGG reaction and KEGG orthology codes [30,216], PFAM domain codes [217], common gene symbol, textual description of the function. At times, reactions were already included in the draft pan-GSMM, but with an incomplete GPR: in this case, the GPR was updated using the *update_genes_from_gpr* from COBRApy [81].

During the reconstruction process, some transporter-encoding genes were associated with more than one transport reaction. According to the experimental Biolog® data, false-positive transport reactions were removed.

5.4 Discussion

In this preliminary study, a GSMM for *T.* sp. *Sel9* was drafted and its network topology manually curated. While the resulting model cannot be considered ready to be used in prospective applications, a first, necessary step of manual curation was completed.

In order to fully exploit the gene recovery functions of Gempipe (see **Chapter 2**), the *T.* sp. *Sel9* genome could not be given in input alone. Instead, 31 other genomes were processed together to feed the 3-steps gene recovery, which yielded important metabolic genes. An ANI analysis revealed the similarity between *T.* sp. *Sel9* and the other *Thauera* species, of which *T. butanivorans* was confirmed to be the closest, as previously found by analyzing the 16S gene sequence [12].

Here, the manual correction was exclusively driven by the Biolog® data, which evaluates substrate usage capabilities. For this reason, most of the added reactions are transporters, or metabolic reactions that connect the uptaken substrate to the main carbon metabolism: in other words, the manual curation was focused on the most "peripheral" part of the reaction network. This means that other missing reactions are expected, involved in some other parts of the metabolism that the Biolog® screening is not able to probe. In particular, among the gene clusters associated with a KEGG Reaction code [30] during the draft pan-GSMM reconstruction, 253 contains genes from *T.* sp. *Sel9* and are still unmodeled in sel9_draft_v2. This data not only suggests the amount of curation work still missing, but also gives an idea of the poor coverage of metabolic genes contained in the BiGG database [28,56], on which both Gempipe and CarveMe [44] are based.

Interestingly, when a draft model for *T.* sp. *Sel9* was reconstructed with CarveMe [44], it showed many more false positive growths compared to the Gempipe reconstruction, despite the same proteome and the same gene database being used. Once again (see also **Chapter 2**), this highlights the side effects of the particular gap-filling algorithm implemented in CarveMe, designed to *"enforce network connectivity"* [44]. This gap-filling could make a positive impact when fragmented metagenome-assembled genomes (MAGs) are used to model communities; however, in the context of this study, better performances are provided by the more conservative introduction of reactions implemented in Gempipe.

Biolog® data also exposed the problem of transporters annotation [183]: substrate specificity is difficult to deduce via traditional alignments because of the high sequence similarity among transporters [387], which are often part of large families of paralog genes [388], where a single point mutation can be sufficient to change substrate specificity [389]. In this regard, it was recently demonstrated [183] that even highly studied *E. coli* strains, which are overrepresented in BiGG [56], when modeled with CarveMe [44] receive ~30% of incorrect transport reactions (percentage expected to be much higher in non-model species [183]).

Following the Biolog®-based manual curation, the *T.* sp. *Sel9* draft model reached a 93% accuracy evaluated on 148 substrates available in the BiGG database [56]. Accuracy was not full: for 10 substrates surpassing the AUC threshold, it was not possible to find transporters, nor any of the internal metabolic reactions connecting the substrate to the central carbon metabolism. These substrates were, specifically, D-xylose (B08), D-ribose (C04) and D-lyxose (H06) in the PM1 plate; D-arabinose (B05), 2-deoxy-D-ribose (B09), citramalate (E04), D-glucosamine (E05), 5-keto-D-gluconate (E12), dihydroxyacetone (H09) and 2,3-butanedione (H11) in the PM2A plate. All these substrates, except for 2,3-butanedione, were suggested by three independent groups as false positive, since they cause abiotic Dye mix D reduction [126–128]. Moreover, they are recognized as "potential false positives" even in the Biolog® official documentation, which also states that it is not possible to provide a definitive list of potential false positives, as they are variable with pH, temperature and duration of the incubation. Considering 7 of these 9 substrates as actual false positives (2-deoxy-D-ribose and dihydroxyacetone should be further investigated given their higher signal), the accuracy would improve reaching a 98%.

This preliminary work is the first part of a larger project aimed at characterizing the metabolism of *T.* sp. *Sel9*, possibly in the perspective of improving its yield in biopolymers. Apart from a complete reaction network (currently still missing), the quantitative description of the *T.* sp. *Sel9* metabolism requires additional experimental data (see **Chapter 1**) including, for example, the experimental biomass composition, an estimate of the maintenance energy, and uptake / secretion rates to constrain the model. For organizational reasons, it was not possible to complete the experiments, which have been postponed to the next year.

However, a challenge is expected in modeling and improving the PHAs *in silico* yield: FBA, given its nature (see **Chapter 1**), can only be used to improve the yield of biochemicals whose production is coupled with biomass formation. Indeed, FBA predicts fluxes equal to zero for each reaction that do not directly contribute to improving the biomass yield, when the biomass equation is set as objective [74]. This is the case of PHAs production, as some of their precursors are also used in cell replication, meaning that the higher is the growth rate, the lower will be the PHAs formation rate, and vice versa [74].

5.5 Supplementary Materials

5.5.1 Supplementary Figures

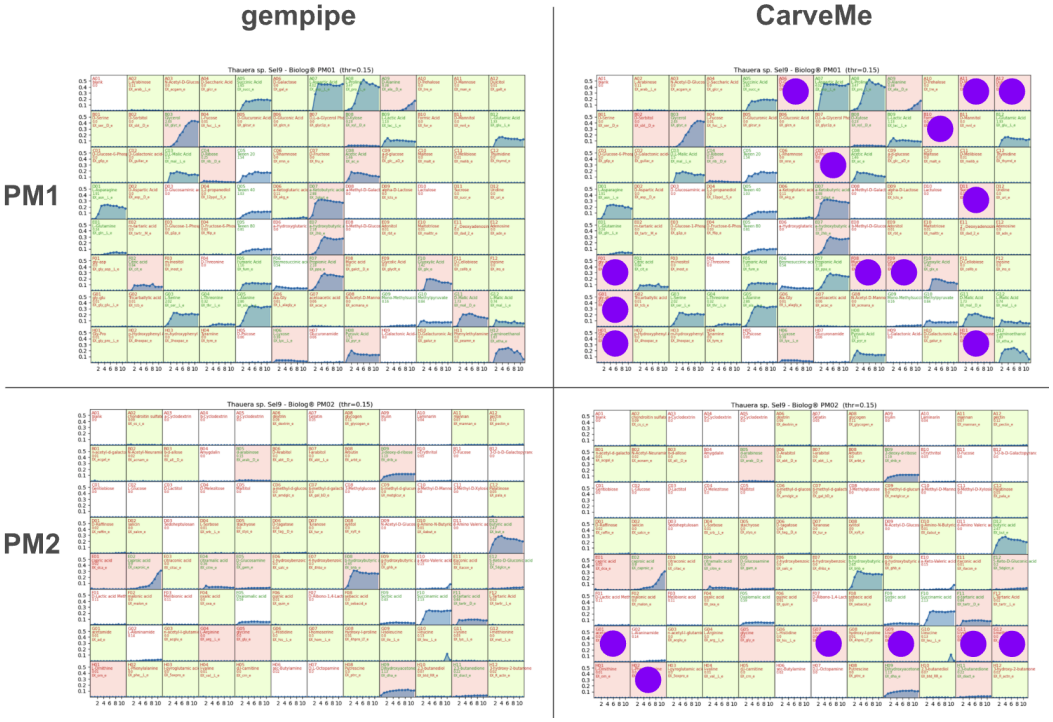

Figure 5.S1. Comparison of Biolog® PM1 and PM2A simulations between a raw (non-curated) *T.* sp. *Sel9* model produced with Gempipe, and a model produced by CarveMe v1.6.1 (default parameters) starting from the same proteome. Simulations were performed for both models on the same CDM medium and using the same *biolog_preview* function from the Gempipe API. In each box (i.e., each well) the normalized Biolog® signal is represented. Red boxes represent mismatches with the experimental data (either FP or FN); green boxes represent matches (either TP or TN). White boxes represent substrates outside the scope of BiGG [28]. Violet circles indicate 18 FP mismatches obtained with the draft GSMM built by CarveMe, and not with the one built by Gempipe.

5.5.2 Supplementary Tables

Table 5.S1. Genome assemblies used in this study. "*": good-quality assembly (based on N50, number of contigs, and missing BUSCO orthologs [179]).

	accession		organisms
1	GCA_000021765.1	*	*Thauera aminoaromatica* MZ1T
2	GCA_008015255.1		*Thauera aminoaromatica* Bin_27_1
3	GCA_023150165.1		*Thauera aminoaromatica* MAG.280
4	GCA_026646855.1		*Thauera aminoaromatica* PDA_C_MAG.20
5	GCA_035328525.1		*Thauera aminoaromatica* AMR_MDS_0285
6	GCA_035330645.1		*Thauera aminoaromatica* AMR_MDS_0373

7	GCA_035330945.1		*Thauera aminoaromatica* AMR_MDS_0357
8	GCA_035332105.1		*Thauera aminoaromatica* AMR_MDS_0495
9	GCA_035333225.1		*Thauera aminoaromatica* AMR_MDS_0440
10	GCA_035333565.1		*Thauera aminoaromatica* AMR_MDS_0423
11	GCA_035334805.1		*Thauera aminoaromatica* AMR_MDS_0564
12	GCA_035335305.1		*Thauera aminoaromatica* AMR_MDS_0539
13	GCA_035337835.1		*Thauera aminoaromatica* AMR_MDS_0711
14	GCA_035338655.1		*Thauera aminoaromatica* AMR_MDS_0670
15	GCA_035339135.1		*Thauera aminoaromatica* AMR_MDS_0647
16	GCA_035339555.1		*Thauera aminoaromatica* AMR_MDS_0627
17	GCA_035344935.1		*Thauera aminoaromatica* AMR_MDS_0932
18	GCA_035346115.1		*Thauera aminoaromatica* AMR_MDS_0876
19	GCA_035355185.1		*Thauera aminoaromatica* AMR_MDS_1230
20	GCA_035400865.1		*Thauera aminoaromatica* AMR_MDS_3980
21	GCA_035439025.1		*Thauera aminoaromatica* AMR_MDS_5552
22	GCA_000310185.1		*Thauera aminoaromatica* S2
23	GCA_023093595.1	*	*Thauera aromatica* LG356
24	GCA_023093675.1	*	*Thauera aromatica* AR-1
25	GCA_023093695.1	*	*Thauera aromatica* SP
26	GCA_003030465.1	*	*Thauera aromatica* K172
27	GCA_001591165.1	*	*Thauera butanivorans* NBRC 103042
28	GCA_001922305.1	*	*Thauera chlorobenzoica* 3CB1
29	GCA_900108255.1	*	*Thauera chlorobenzoica* 3CB-1
30	GCA_001051995.2	*	*Thauera humireducens* SgZ-1
31	GCA_932126575.1	*	*Thauera humireducens* Piv1
32	GCA_023805265.1	*	*Thauera linaloolentis* TL0
33	GCA_000310205.1		*Thauera linaloolentis* 47Lol = DSM 12138
34	GCA_000621305.1	*	*Thauera linaloolentis* 47Lol = DSM 12138
35	GCA_029588855.1	*	*Thauera mechernichensis* TL1
36	GCA_001696715.1	*	*Thauera phenolivorans* ZV1C
37	GCA_012515175.1		*Thauera phenolivorans* AS06rmzACSIP_256
38	GCA_035435145.1		*Thauera phenylacetica* AMR_MDS_5750
39	GCA_000310225.1		*Thauera phenylacetica* B4P
40	GCA_002245655.1	*	*Thauera propionica* KNDSS-Mac4
41	GCA_028697245.1		*Thauera propionica* STE_157
42	GCA_034004985.1		*Thauera propionica* BF4_124
43	GCA_014489115.1	*	*Thauera sedimentorum* CAU 1555
44	GCA_014841095.1	*	*Thauera sedimentorum* CAU 1555
45	GCA_000284915.1		*Thauera selenatis* AX ATCC 55363
46	GCA_003248915.1		*Thauera* sp. S2_005_003_R2_39
47	GCA_003446655.1		*Thauera* sp. UBA9855
48	GCA_003450855.1		*Thauera* sp. UBA12224
49	GCA_003486975.1		*Thauera* sp. UBA8772
50	GCA_015665165.1		*Thauera* sp. NTO_MAG2
51	GCA_016789945.1		*Thauera* sp. new MAG-178
52	GCA_016790365.1		*Thauera* sp. new MAG-177
53	GCA_017302295.1	*	*Thauera* sp. UWPOB_THAU1
54	GCA_017985115.1		*Thauera* sp. Go_Eff_bin_267
55	GCA_017988795.1		*Thauera* sp. Gw_SlPrim_bin_4
56	GCA_017988825.1		*Thauera* sp. Gw_SlPrim_bin_139
57	GCA_017990655.1		*Thauera* sp. Gw_UH_bin_42
58	GCA_017990925.1		*Thauera* sp. Go_Inlet_bin_83
59	GCA_017993915.1		*Thauera* sp. Go_SlAct_bin_483
60	GCA_017997665.1		*Thauera* sp. Go_SlPrim_bin_232
61	GCA_017997795.1		*Thauera* sp. Go_UH_bin_156
62	GCA_017998995.1		*Thauera* sp. Go_SlPrim_bin_184
63	GCA_018006565.1		*Thauera* sp. Go_Inlet_bin_108
64	GCA_018056015.1		*Thauera* sp. Gw_UH_bin_189
65	GCA_019105005.1		*Thauera* sp. DR_8_1.67
66	GCA_019635105.1		*Thauera* sp. ACE_PRO56

67	GCA_020445825.1		*Thauera* sp. HKST-UBA160
68	GCA_023150215.1		*Thauera* sp. MAG.443
69	GCA_024235615.1	*	*Thauera* sp. HK-STAS-PROT-107
70	GCA_024709385.1		*Thauera* sp. RS262_metabat.bin.19
71	GCA_024709915.1		*Thauera* sp. RC341_metabat.bin.36
72	GCA_024710225.1		*Thauera* sp. co_assembly_metabat.bin.1
73	GCA_024710625.1		*Thauera* sp. RXS1_metabat.bin.27
74	GCA_024710965.1		*Thauera* sp. RX2_metabat.bin.53
75	GCA_026646795.1		*Thauera* sp. PDA_C_MAG.59
76	GCA_026646815.1		*Thauera* sp. PDA_C_MAG.51
77	GCA_029286945.1		*Thauera* sp. THA1
78	GCA_029982735.1		*Thauera* sp. DJKA107_ANT-F21_bin.41
79	GCA_030603795.1		*Thauera* sp. CSMAG_1205
80	GCA_030603925.1		*Thauera* sp. CSMAG_1199
81	GCA_030604685.1		*Thauera* sp. CSMAG_1162
82	GCA_033838845.1		*Thauera* sp. PE-1-142
83	GCA_034002425.1		*Thauera* sp. BF4_92
84	GCA_034003485.1		*Thauera* sp. BF4_29
85	GCA_034005275.1		*Thauera* sp. BF4_105
86	GCA_034675825.1	*	*Thauera* sp. R32x_68d_np_Bin2.3
87	GCA_034676565.1		*Thauera* sp. R4x_68d_np_Bin1.13
88	GCA_035341255.1		*Thauera* sp. AMR_MDS_1103
89	GCA_035342315.1		*Thauera* sp. AMR_MDS_1053
90	GCA_035391985.1		*Thauera* sp. AMR_MDS_3211
91	GCA_035426965.1		*Thauera* sp. AMR_MDS_5126
92	GCA_035433105.1		*Thauera* sp. AMR_MDS_5268
93	GCA_035437105.1		*Thauera* sp. AMR_MDS_5652
94	GCA_035437605.1		*Thauera* sp. AMR_MDS_5625
95	GCA_035443205.1		*Thauera* sp. AMR_MDS_5332
96	GCA_035444445.1		*Thauera* sp. AMR_MDS_5881
97	GCA_035444905.1	*	*Thauera* sp. AMR_MDS_5854
98	GCA_035445705.1		*Thauera* sp. AMR_MDS_5814
99	GCA_035448205.1		*Thauera* sp. AMR_MDS_3024
100	GCA_035904165.1		*Thauera* sp. ST_AS_maxbin.018
101	GCA_035913055.1		*Thauera* sp. RA_AS_maxbin.121
102	GCA_000310125.2	*	*Thauera* sp. 27
103	GCA_000310145.2	*	*Thauera* sp. 28
104	GCA_009469595.2	*	*Thauera* sp. 2A1
105	GCA_000310165.2	*	*Thauera* sp. 63
106	GCA_001310835.1		*Thauera* sp. DNT-1 JCM 12309
107	GCA_029223545.1	*	*Thauera* sp. GDN1
108	GCA_002354895.1	*	*Thauera* sp. K11
109	GCA_025699485.1	*	*Thauera* sp. Sel9
110	GCA_000831325.1	*	*Thauera* sp. SWB20
111	GCA_002382285.1		*Thauera* sp. UBA4044
112	GCA_002422685.1		*Thauera* sp. UBA6194
113	GCA_006007845.1		*Thauera* sp. UPWRP
114	GCA_027594845.1	*	*Thauera* sp. WB-2
115	GCA_000443165.1	*	*Thauera terpenica* 58Eu
116	GCA_934272915.1		uncultured *Thauera* sp. (...) Nepal_MoBio (...) RAJ1013YZ.45
117	GCA_937863675.1		uncultured *Thauera* sp. SRR5676468 (...) MAG
118	GCA_937983465.1		uncultured *Thauera* sp. SRR1506951 (...) MAG
119	GCA_943913025.1	*	uncultured *Thauera* sp. KkPmcVOWp7_bin.13.MAG

Conclusions

In this PhD thesis, I presented my research on genome-scale metabolic models (GSMMs), with a particular focus on multi-strain reconstructions and analyses. These models, based on detailed genome annotations, enable a direct connection between genotype and phenotype, serving as platforms for prediction and hypothesis generation [1].

Here, I introduced Gempipe (**Chapter 2**), a versatile tool that embodies my approach to the reconstruction, curation, and analysis of GSMMs. It was instrumental during this entire thesis, where it was used for suggesting growth factors for an endosymbiont (**Chapter 3**), investigating the metabolic diversity of a species (**Chapter 4**), and quickly improving the topology of a draft model (**Chapter 5**). While already useful, I do not consider Gempipe a "closed chapter": it will be used in other projects and its development will continue, fixing bugs and adding new features.

My journey through reconstruction tools, as an absolute beginner, started with CarveMe [44], which I still admire for many reasons: it was the first to introduce the concept of "universe" model on which to base all the reconstructions; its code was elegantly written; its models were always ready to perform flux-balance analysis (FBA); it was incredibly fast and simple to use. I remember the first time I tried it: I thought *"is the reconstruction of GSMMs really so easy?"*. The skepticism started when I tried to model *Candidatus* Erwinia dacicola for the first time. With CarveMe, the *in silico* biomass production resulted strictly dependent on the uptake of benzoate and petroselinate, so I spent quite some time searching in literature for possible explanations supporting these substrates. Months later, I realized how many orphan reactions (i.e., with no associated genes) were dividing benzoate and petroselinate from the rest of the metabolism, so they quickly lost their appeal as growth factors. This was how I began to appreciate the power, but also the hidden danger, behind CarveMe's gap-filling algorithm, which should be applied in different contexts. Together with the identification of the other drawbacks discussed throughout this thesis, the idea of creating an alternative reconstruction tool based on different principles gained traction and, gradually, Gempipe was conceived.

At times, however, independently from the reconstruction tool, performance of GSMMs can be unsatisfactory, especially with non-model organisms. From my observations, aside from specific phenotypes like overflow metabolism and transcriptional repressions, GSMMs and flux-balance analysis (FBA) are, in their simplicity, remarkably robust. Eventual poor performances are mostly

attributable to an insufficient refinement of the model, where for "refinement" I mean the level of completeness and accuracy in representing the biochemical network of the organism (including the associated genes, see for example subsection 2.5.1). This aspect is relevant especially for biotechnological applications, where non-model organisms are frequently involved, and the time required to obtain a quality GSMM may be too long.

Aside from the quality of GSMMs, however, I believe that multi-strain analysis methods – pioneered in 2013 by Monk and colleagues [114] – still have significant untapped potential. In taxonomy, knowing the reactome of each strain allows for the determination of the core reactome of a species: this could delineate not only the metabolic boundaries that distinguish a species from the others, but also the metabolic strategy that a clade is using to proliferate in a specific environment, which could ultimately lead to speciation. In biotechnology, the evolution of these methods will enable the selection of the best strain for a specific bioprocess, based on strain-specific predictions.

In this regard, the field of GSMM analysis methods is vast, and I feel I have just scratched the surface. There are numerous methods I would like to master in the future, some of which could be relevant in the context of biodiversity exploration and rational strain selection. For example, I would like to go beyond steady-state and be able to describe the temporal variations characterizing a batch culture. This is achievable through the dynamic-FBA, an extension of FBA that approximates changes in substrates, products, and biomass concentrations over time [75,390]. Then, I would like to tackle the next-generation models which, contrary to the "first-generation" addressed in this thesis, consider the limited capacity of a cell in synthesizing its proteins, and allow simulations of overflow metabolisms [391]. Among the next-generation models, "fine-grained" models explicitly account for the mass and energy spent by the cell in producing each enzyme (both in terms of transcription and translation), but they are hindered by the availability of specific experimental data [392], which could limit eventual multi-strain applications. Conversely, "coarse-grained" next-generation models require a minimal data input: an estimate of the cell's total enzyme concentration, and the molecular weight and enzyme turnover number (k_{cat}) for each enzyme [392]. Although the latter data may appear as not accessible for non-model species, a breakthrough came in 2022 with the work of Li and colleagues [393], who used deep-learning techniques to enable the estimation of k_{cat} values for each enzyme in a genome, using just the substrate structure and protein sequence as inputs [393].

However, it is crucial to emphasize that the success of the aforementioned methods, and of any model analysis technique in general, depends on the

refinement of the model itself (see above). In this perspective, the reconstruction of first-generation quality GSMMs still remains a bottleneck. In my opinion, a big challenge in the field of reconstructions is the development of a single, generalized tool capable of generating the highest quality ready-to-use models, based on a standard universe developed and maintained by the community; this universe would represent the entire known genetic and metabolic diversity of bacteria, with every reaction mapped to a consensus, public metabolic map. In such perhaps utopian conditions, models would be truly compatible between each other and, importantly, the paradigm of manual curation would shift: corrections would not be applied on a specific model (causing fragmentations and incompatibility), but directly on the universe, propagating them instantly to every model. In these conditions, every sequenced genome would have its always up-to-date, FBA-ready model; moreover, the concept of model versioning would become obsolete, as the only versioning would be applied to the universe.

Beyond the methods implemented and the individual case studies, this thesis highlights how the adoption of GSMMs enables genomics, a traditionally descriptive discipline, to progress toward a more mechanistic understanding of biological problems. Indeed, while genomics essentially suggests gene functions and compares their presence across organisms, GSMMs weave these genes into a data structure that reveals the interdependence of components in a complex cellular system.

When I began this PhD journey, I believed it was possible, given only a genome and a growth medium recipe, to predict secretion and biomass concentrations over time for any bacterium. I now understand that, in addition to the genome, various experimental data are required to properly parameterize a model. Nevertheless, models remain undeniably valuable. Perhaps one day, genome interpretation will reach a level of sophistication where every possible cellular behavior can be quantitatively predicted without requiring other experimental input. We are still far from that goal. The road ahead is long and winding but… extremely fascinating.

References

1. O'Brien, E.J.; Monk, J.M.; Palsson, B.O. Using Genome-Scale Models to Predict Biological Capabilities. *Cell* **2015**, *161*, 971–987, doi:10.1016/j.cell.2015.05.019.
2. Francke, C.; Siezen, R.J.; Teusink, B. Reconstructing the Metabolic Network of a Bacterium from Its Genome. *Trends in Microbiology* **2005**, *13*, 550–558, doi:10.1016/j.tim.2005.09.001.
3. Teusink, B.; van Enckevort, F.H.J.; Francke, C.; Wiersma, A.; Wegkamp, A.; Smid, E.J.; Siezen, R.J. In Silico Reconstruction of the Metabolic Pathways of *Lactobacillus Plantarum* : Comparing Predictions of Nutrient Requirements with Those from Growth Experiments. *Appl Environ Microbiol* **2005**, *71*, 7253–7262, doi:10.1128/AEM.71.11.7253-7262.2005.
4. Pavlidi, N.; Gioti, A.; Wybouw, N.; Dermauw, W.; Ben-Yosef, M.; Yuval, B.; Jurkevich, E.; Kampouraki, A.; Van Leeuwen, T.; Vontas, J. Transcriptomic Responses of the Olive Fruit Fly Bactrocera Oleae and Its Symbiont Candidatus Erwinia Dacicola to Olive Feeding. *Sci Rep* **2017**, *7*, 42633, doi:10.1038/srep42633.
5. Ben-Yosef, M.; Pasternak, Z.; Jurkevitch, E.; Yuval, B. Symbiotic Bacteria Enable Olive Fly Larvae to Overcome Host Defences. *R. Soc. open sci.* **2015**, *2*, 150170, doi:10.1098/rsos.150170.
6. Arora, A.K.; Douglas, A.E. Hype or Opportunity? Using Microbial Symbionts in Novel Strategies for Insect Pest Control. *Journal of Insect Physiology* **2017**, *103*, 10–17, doi:10.1016/j.jinsphys.2017.09.011.
7. Rosselló-Móra, R.; Amann, R. Past and Future Species Definitions for Bacteria and Archaea. *Systematic and Applied Microbiology* **2015**, *38*, 209–216, doi:10.1016/j.syapm.2015.02.001.
8. Teusink, B.; Wiersma, A.; Molenaar, D.; Francke, C.; de Vos, W.M.; Siezen, R.J.; Smid, E.J. Analysis of Growth of Lactobacillus Plantarum WCFS1 on a Complex Medium Using a Genome-Scale Metabolic Model. *Journal of Biological Chemistry* **2006**, *281*, 40041–40048, doi:10.1074/jbc.M606263200.
9. Vinay-Lara, E.; Hamilton, J.J.; Stahl, B.; Broadbent, J.R.; Reed, J.L.; Steele, J.L. Genome –Scale Reconstruction of Metabolic Networks of Lactobacillus Casei ATCC 334 and 12A. *PLoS ONE* **2014**, *9*, e110785, doi:10.1371/journal.pone.0110785.
10. Li, W.; Wu, Q.; Kwok, L.; Zhang, H.; Gan, R.; Sun, Z. Population and Functional Genomics of Lactic Acid Bacteria, an Important Group of Food Microorganism: Current Knowledge, Challenges, and Perspectives. *Food Frontiers* **2024**, *5*, 3–23, doi:10.1002/fft2.321.
11. Siezen, R.J.; van Hylckama Vlieg, J.E. Genomic Diversity and Versatility of Lactobacillus Plantarum, a Natural Metabolic Engineer. *Microb Cell Fact* **2011**, *10*, S3, doi:10.1186/1475-2859-10-S1-S3.
12. Andreolli, M.; Scerbacov, V.; Frison, N.; Zaccone, C.; Lampis, S. Thauera Sp. Sel9, a New Bacterial Strain for Polyhydroxyalkanoates Production from Volatile Fatty Acids. *New Biotechnology* **2022**, *72*, 71–79, doi:10.1016/j.nbt.2022.09.004.

13. Giani, A.M.; Gallo, G.R.; Gianfranceschi, L.; Formenti, G. Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly. *Computational and Structural Biotechnology Journal* **2020**, *18*, 9–19, doi:10.1016/j.csbj.2019.11.002.
14. Panikov, N.S. Genome-Scale Reconstruction of Microbial Dynamic Phenotype: Successes and Challenges. *Microorganisms* **2021**, *9*, doi:10.3390/microorganisms9112352.
15. Sauer, U.; Heinemann, M.; Zamboni, N. Getting Closer to the Whole Picture. *Science* **2007**, *316*, 550–551, doi:10.1126/science.1142502.
16. Bruggeman, F.J.; Westerhoff, H.V. The Nature of Systems Biology. *Trends in Microbiology* **2007**, *15*, 45–50, doi:10.1016/j.tim.2006.11.003.
17. Oulas, A.; Minadakis, G.; Zachariou, M.; Sokratous, K.; Bourdakou, M.M.; Spyrou, G.M. Systems Bioinformatics: Increasing Precision of Computational Diagnostics and Therapeutics through Network-Based Approaches. *Briefings in Bioinformatics* **2019**, *20*, 806–824, doi:10.1093/bib/bbx151.
18. Santos, F.; Boele, J.; Teusink, B. A Practical Guide to Genome-Scale Metabolic Models and Their Analysis. In *Methods in Enzymology*; Elsevier, 2011; Vol. 500, pp. 509–532 ISBN 978-0-12-385118-5.
19. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.-F.; Dougherty, B.A.; Merrick, J.M.; et al. Whole-Genome Random Sequencing and Assembly of *Haemophilus Influenzae* Rd. *Science* **1995**, *269*, 496–512, doi:10.1126/science.7542800.
20. Fang, X.; Lloyd, C.J.; Palsson, B.O. Reconstructing Organisms in Silico: Genome-Scale Models and Their Emerging Applications. *Nat Rev Microbiol* **2020**, *18*, 731–743, doi:10.1038/s41579-020-00440-4.
21. Thiele, I.; Palsson, B.Ø. A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction. *Nat Protoc* **2010**, *5*, 93–121, doi:10.1038/nprot.2009.203.
22. Varma, A.; Palsson, B.O. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nat Biotechnol* **1994**, *12*, 994–998, doi:10.1038/nbt1094-994.
23. Lee, S.Y.; Nielsen, J.; Stephanopoulos, G. *Metabolic Engineering: Concepts and Applications*; Advanced biotechnology; Wiley, 2021; ISBN 978-3-527-82345-1.
24. Branco dos Santos, F.; de Vos, W.M.; Teusink, B. Towards Metagenome-Scale Models for Industrial Applications—the Case of Lactic Acid Bacteria. *Current Opinion in Biotechnology* **2013**, *24*, 200–206, doi:10.1016/j.copbio.2012.11.003.
25. García Sánchez, C.E.; Torres Sáez, R.G. Comparison and Analysis of Objective Functions in Flux Balance Analysis. *Biotechnol Progress* **2014**, *30*, 985–991, doi:10.1002/btpr.1949.
26. Keating, S.M.; Waltemath, D.; König, M.; Zhang, F.; Dräger, A.; Chaouiya, C.; Bergmann, F.T.; Finney, A.; Gillespie, C.S.; Helikar, T.; et al. SBML Level 3: An Extensible Format for the Exchange and Reuse of Biological Models. *Mol Syst Biol* **2020**, *16*, doi:10.15252/msb.20199110.
27. Orth, J.D.; Thiele, I.; Palsson, B.Ø. What Is Flux Balance Analysis? *Nat Biotechnol* **2010**, *28*, 245–248, doi:10.1038/nbt.1614.

28. King, Z.A.; Lu, J.; Dräger, A.; Miller, P.; Federowicz, S.; Lerman, J.A.; Ebrahim, A.; Palsson, B.O.; Lewis, N.E. BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models. *Nucleic Acids Res* **2016**, *44*, D515–D522, doi:10.1093/nar/gkv1049.
29. Seaver, S.M.D.; Liu, F.; Zhang, Q.; Jeffryes, J.; Faria, J.P.; Edirisinghe, J.N.; Mundy, M.; Chia, N.; Noor, E.; Beber, M.E.; et al. The ModelSEED Biochemistry Database for the Integration of Metabolic Annotations and the Reconstruction, Comparison and Analysis of Metabolic Models for Plants, Fungi and Microbes. *Nucleic Acids Research* **2021**, *49*, D575–D588, doi:10.1093/nar/gkaa746.
30. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000**, *28*, 27–30, doi:10.1093/nar/28.1.27.
31. Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C.A.; Holland, T.A.; Keseler, I.M.; Kothari, A.; Kubo, A.; et al. The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases. *Nucl. Acids Res.* **2014**, *42*, D459–D471, doi:10.1093/nar/gkt1103.
32. Moretti, S.; Martin, O.; Van Du Tran, T.; Bridge, A.; Morgat, A.; Pagni, M. MetaNetX/MNXref – Reconciliation of Metabolites and Biochemical Reactions to Bring Together Genome-Scale Metabolic Networks. *Nucleic Acids Res* **2016**, *44*, D523–D526, doi:10.1093/nar/gkv1117.
33. Pham, N.; Van Heck, R.G.A.; Van Dam, J.C.J.; Schaap, P.J.; Saccenti, E.; Suarez-Diez, M. Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling. *Metabolites* **2019**, *9*, 28, doi:10.3390/metabo9020028.
34. Mendoza, S.N.; Olivier, B.G.; Molenaar, D.; Teusink, B. A Systematic Assessment of Current Genome-Scale Metabolic Reconstruction Tools. *Genome Biol* **2019**, *20*, 158, doi:10.1186/s13059-019-1769-1.
35. King, Z.A.; Dräger, A.; Ebrahim, A.; Sonnenschein, N.; Lewis, N.E.; Palsson, B.O. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput Biol* **2015**, *11*, e1004321, doi:10.1371/journal.pcbi.1004321.
36. Rowe, E.; Palsson, B.O.; King, Z.A. Escher-FBA: A Web Application for Interactive Flux Balance Analysis. *BMC Systems Biology* **2018**, *12*, 84, doi:10.1186/s12918-018-0607-5.
37. Norsigian, C.J.; Kavvas, E.; Seif, Y.; Palsson, B.O.; Monk, J.M. iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of Acinetobacter Baumannii AYE. *Front. Genet.* **2018**, *9*, 121, doi:10.3389/fgene.2018.00121.
38. Norsigian, C.J.; Danhof, H.A.; Brand, C.K.; Oezguen, N.; Midani, F.S.; Palsson, B.O.; Savidge, T.C.; Britton, R.A.; Spinler, J.K.; Monk, J.M. Systems Biology Analysis of the Clostridioides Difficile Core-Genome Contextualizes Microenvironmental Evolutionary Pressures Leading to Genotypic and Phenotypic Divergence. *npj Syst Biol Appl* **2020**, *6*, 31, doi:10.1038/s41540-020-00151-9.
39. Orth, J.D.; Conrad, T.M.; Na, J.; Lerman, J.A.; Nam, H.; Feist, A.M.; Palsson, B.Ø. A Comprehensive Genome-scale Reconstruction of *Escherichia Coli* Metabolism—2011. *Molecular Systems Biology* **2011**, *7*, 535,

doi:10.1038/msb.2011.65.
40. Joshi, C.J.; Peebles, C.A.M.; Prasad, A. Modeling and Analysis of Flux Distribution and Bioproduct Formation in Synechocystis Sp. PCC 6803 Using a New Genome-Scale Metabolic Reconstruction. *Algal Research* **2017**, *27*, 295–310, doi:10.1016/j.algal.2017.09.013.
41. Gevorgyan, A.; Poolman, M.G.; Fell, D.A. Detection of Stoichiometric Inconsistencies in Biomolecular Models. *Bioinformatics* **2008**, *24*, 2245–2251, doi:10.1093/bioinformatics/btn425.
42. Shin, W.; Hellerstein, J.L. Isolating Structural Errors in Reaction Networks in Systems Biology. *Bioinformatics* **2021**, *37*, 388–395, doi:10.1093/bioinformatics/btaa720.
43. Rawls, K.D.; Dougherty, B.V.; Blais, E.M.; Stancliffe, E.; Kolling, G.L.; Vinnakota, K.; Pannala, V.R.; Wallqvist, A.; Papin, J.A. A Simplified Metabolic Network Reconstruction to Promote Understanding and Development of Flux Balance Analysis Tools. *Computers in Biology and Medicine* **2019**, *105*, 64–71, doi:10.1016/j.compbiomed.2018.12.010.
44. Machado, D.; Andrejev, S.; Tramontano, M.; Patil, K.R. Fast Automated Reconstruction of Genome-Scale Metabolic Models for Microbial Species and Communities. *Nucleic Acids Research* **2018**, *46*, 7542–7553, doi:10.1093/nar/gky537.
45. Ponce-de-León, M.; Montero, F.; Peretó, J. Solving Gap Metabolites and Blocked Reactions in Genome-Scale Models: Application to the Metabolic Network of Blattabacterium Cuenoti. *BMC Syst Biol* **2013**, *7*, 114, doi:10.1186/1752-0509-7-114.
46. Linard, B.; Ebersberger, I.; McGlynn, S.E.; Glover, N.; Mochizuki, T.; Patricio, M.; Lecompte, O.; Nevers, Y.; Thomas, P.D.; Gabaldón, T.; et al. Ten Years of Collaborative Progress in the Quest for Orthologs. *Molecular Biology and Evolution* **2021**, *38*, 3033–3045, doi:10.1093/molbev/msab098.
47. Miller, J.B.; Pickett, B.D.; Ridge, P.G. JustOrthologs: A Fast, Accurate and User-Friendly Ortholog Identification Algorithm. *Bioinformatics* **2019**, *35*, 546–552, doi:10.1093/bioinformatics/bty669.
48. Norsigian, C.J.; Fang, X.; Seif, Y.; Monk, J.M.; Palsson, B.O. A Workflow for Generating Multi-Strain Genome-Scale Metabolic Models of Prokaryotes. *Nat Protoc* **2020**, *15*, 1–14, doi:10.1038/s41596-019-0254-3.
49. Hu, X.; Friedberg, I. SwiftOrtho: A Fast, Memory-Efficient, Multiple Genome Orthology Classifier. *GigaScience* **2019**, *8*, giz118, doi:10.1093/gigascience/giz118.
50. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and Applications. *BMC Bioinformatics* **2009**, *10*, 421, doi:10.1186/1471-2105-10-421.
51. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol* **2019**, *20*, 238, doi:10.1186/s13059-019-1832-y.
52. Cosentino, S.; Iwasaki, W. SonicParanoid: Fast, Accurate and Easy Orthology Inference. *Bioinformatics* **2019**, *35*, 149–151, doi:10.1093/bioinformatics/bty631.
53. Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P.F.; Prohaska, S.J. Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. *BMC*

Bioinformatics **2011**, *12*, 124, doi:10.1186/1471-2105-12-124.
54. Derelle, R.; Philippe, H.; Colbourne, J.K. Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. *Molecular Biology and Evolution* **2020**, *37*, 3389–3396, doi:10.1093/molbev/msaa159.
55. Cuevas, D.A.; Edirisinghe, J.; Henry, C.S.; Overbeek, R.; O'Connell, T.G.; Edwards, R.A. From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model. *Front. Microbiol.* **2016**, *7*, doi:10.3389/fmicb.2016.00907.
56. Norsigian, C.J.; Pusarla, N.; McConn, J.L.; Yurkovich, J.T.; Dräger, A.; Palsson, B.O.; King, Z. BiGG Models 2020: Multi-Strain Genome-Scale Models and Expansion across the Phylogenetic Tree. *Nucleic Acids Research* **2019**, gkz1054, doi:10.1093/nar/gkz1054.
57. Iffland-Stettner, A.; Okano, H.; Gralka, M.; Guessous, G.; Amarnath, K.; Cordero, O.X.; Hwa, T.; Bonhoeffer, S. A Genome-Scale Metabolic Model of Marine Heterotroph Vibrio Splendidus Strain 1A01. *mSystems* **2023**, e00377-22, doi:10.1128/msystems.00377-22.
58. Chan, S.H.J.; Cai, J.; Wang, L.; Simons-Senftle, M.N.; Maranas, C.D. Standardizing Biomass Reactions and Ensuring Complete Mass Balance in Genome-Scale Metabolic Models. *Bioinformatics* **2017**, *33*, 3603–3609, doi:10.1093/bioinformatics/btx453.
59. Feist, A.M.; Palsson, B.O. The Biomass Objective Function. *Current Opinion in Microbiology* **2010**, *13*, 344–349, doi:10.1016/j.mib.2010.03.003.
60. Monk, J.; Nogales, J.; Palsson, B.O. Optimizing Genome-Scale Network Reconstructions. *Nat Biotechnol* **2014**, *32*, 447–452, doi:10.1038/nbt.2870.
61. Lachance, J.-C.; Lloyd, C.J.; Monk, J.M.; Yang, L.; Sastry, A.V.; Seif, Y.; Palsson, B.O.; Rodrigue, S.; Feist, A.M.; King, Z.A.; et al. BOFdat: Generating Biomass Objective Functions for Genome-Scale Metabolic Models from Experimental Data. *PLOS Computational Biology* **2019**, *15*, e1006971, doi:10.1371/journal.pcbi.1006971.
62. Beck, A.; Hunt, K.; Carlson, R. Measuring Cellular Biomass Composition for Computational Biology Applications. *Processes* **2018**, *6*, 38, doi:10.3390/pr6050038.
63. Zuñiga, C.; Li, C.-T.; Huelsman, T.; Levering, J.; Zielinski, D.C.; McConnell, B.O.; Long, C.P.; Knoshaug, E.P.; Guarnieri, M.T.; Antoniewicz, M.R.; et al. Genome-Scale Metabolic Model for the Green Alga *Chlorella Vulgaris* UTEX 395 Accurately Predicts Phenotypes under Autotrophic, Heterotrophic, and Mixotrophic Growth Conditions. *Plant Physiol.* **2016**, *172*, 589–602, doi:10.1104/pp.16.00593.
64. Fondi, M.; Liò, P. Genome-Scale Metabolic Network Reconstruction. In *Bacterial Pangenomics*; Mengoni, A., Galardini, M., Fondi, M., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2015; Vol. 1231, pp. 233–256 ISBN 978-1-4939-1719-8.
65. Zimmermann, J.; Kaleta, C.; Waschina, S. Gapseq: Informed Prediction of Bacterial Metabolic Pathways and Reconstruction of Accurate Metabolic Models. *Genome Biol* **2021**, *22*, 81, doi:10.1186/s13059-021-02295-1.
66. Henry, C.S.; DeJongh, M.; Best, A.A.; Frybarger, P.M.; Linsay, B.; Stevens, R.L. High-Throughput Generation, Optimization and Analysis of Genome-Scale Metabolic Models. *Nat Biotechnol* **2010**, *28*, 977–982,

doi:10.1038/nbt.1672.
67. Saier, M.H.; Reddy, V.S.; Moreno-Hagelsieb, G.; Hendargo, K.J.; Zhang, Y.; Iddamsetty, V.; Lam, K.J.K.; Tian, N.; Russum, S.; Wang, J.; et al. The Transporter Classification Database (TCDB): 2021 Update. *Nucleic Acids Research* **2021**, *49*, D461–D467, doi:10.1093/nar/gkaa1004.
68. Marinos, G.; Kaleta, C.; Waschina, S. Defining the Nutritional Input for Genome-Scale Metabolic Models: A Roadmap. *PLoS ONE* **2020**, *15*, e0236890, doi:10.1371/journal.pone.0236890.
69. Chang, R.L.; Ghamsari, L.; Manichaikul, A.; Hom, E.F.Y.; Balaji, S.; Fu, W.; Shen, Y.; Hao, T.; Palsson, B.Ø.; Salehi-Ashtiani, K.; et al. Metabolic Network Reconstruction of *Chlamydomonas* Offers Insight into Light-driven Algal Metabolism. *Molecular Systems Biology* **2011**, *7*, 518, doi:10.1038/msb.2011.52.
70. Lieven, C.; Beber, M.E.; Olivier, B.G.; Bergmann, F.T.; Ataman, M.; Babaei, P.; Bartell, J.A.; Blank, L.M.; Chauhan, S.; Correia, K.; et al. MEMOTE for Standardized Genome-Scale Metabolic Model Testing. *Nat Biotechnol* **2020**, *38*, 272–276, doi:10.1038/s41587-020-0446-y.
71. Garcia, S.; Thompson, R.A.; Giannone, R.J.; Dash, S.; Maranas, C.D.; Trinh, C.T. Development of a Genome-Scale Metabolic Model of Clostridium Thermocellum and Its Applications for Integration of Multi-Omics Datasets and Computational Strain Design. *Front. Bioeng. Biotechnol.* **2020**, *8*, 772, doi:10.3389/fbioe.2020.00772.
72. Diogo de Lucena, A. Comparative Systems Biology Analyses of Lactococcus Lactis Subsp. Lactis Strain LMG 19460, Universidade de Lisboa, 2020.
73. Teusink, B.; Smid, E.J. Modelling Strategies for the Industrial Exploitation of Lactic Acid Bacteria. *Nat Rev Microbiol* **2006**, *4*, 46–56, doi:10.1038/nrmicro1319.
74. Puchałka, J.; Oberhardt, M.A.; Godinho, M.; Bielecka, A.; Regenhardt, D.; Timmis, K.N.; Papin, J.A.; Martins dos Santos, V.A.P. Genome-Scale Reconstruction and Analysis of the Pseudomonas Putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. *PLoS Comput Biol* **2008**, *4*, e1000210, doi:10.1371/journal.pcbi.1000210.
75. Becker, S.A.; Feist, A.M.; Mo, M.L.; Hannum, G.; Palsson, B.Ø.; Herrgard, M.J. Quantitative Prediction of Cellular Metabolism with Constraint-Based Models: The COBRA Toolbox. *Nat Protoc* **2007**, *2*, 727–738, doi:10.1038/nprot.2007.99.
76. Mendoza, S.N.; Saa, P.A.; Teusink, B.; Agosin, E. Metabolic Modeling of Wine Fermentation at Genome Scale. In *Computational Systems Biology in Medicine and Biotechnology*; Cortassa, S., Aon, M.A., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2022; Vol. 2399, pp. 395–454 ISBN 978-1-0716-1830-1.
77. Teusink, B.; Wiersma, A.; Jacobs, L.; Notebaart, R.A.; Smid, E.J. Understanding the Adaptive Growth Strategy of Lactobacillus Plantarum by In Silico Optimisation. *PLoS Comput Biol* **2009**, *5*, e1000410, doi:10.1371/journal.pcbi.1000410.
78. Heirendt, L.; Arreckx, S.; Pfau, T.; Mendoza, S.N.; Richelle, A.; Heinken, A.; Haraldsdóttir, H.S.; Wachowiak, J.; Keating, S.M.; Vlasov, V.; et al.

Creation and Analysis of Biochemical Constraint-Based Models Using the COBRA Toolbox v.3.0. *Nat Protoc* **2019**, *14*, 639–702, doi:10.1038/s41596-018-0098-2.
79. Ng, R.H.; Lee, J.W.; Baloni, P.; Diener, C.; Heath, J.R.; Su, Y. Constraint-Based Reconstruction and Analyses of Metabolic Models: Open-Source Python Tools and Applications to Cancer. *Front. Oncol.* **2022**, *12*, 914594, doi:10.3389/fonc.2022.914594.
80. Machado, D. A Benchmark of Optimization Solvers for Genome-Scale Metabolic Modeling of Organisms and Communities. *mSystems* **2024**, e00833-23, doi:10.1128/msystems.00833-23.
81. Ebrahim, A.; Lerman, J.A.; Palsson, B.O.; Hyduke, D.R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **2013**, *7*, 74, doi:10.1186/1752-0509-7-74.
82. Schellenberger, J.; Que, R.; Fleming, R.M.T.; Thiele, I.; Orth, J.D.; Feist, A.M.; Zielinski, D.C.; Bordbar, A.; Lewis, N.E.; Rahmanian, S.; et al. Quantitative Prediction of Cellular Metabolism with Constraint-Based Models: The COBRA Toolbox v2.0. *Nat Protoc* **2011**, *6*, 1290–1307, doi:10.1038/nprot.2011.308.
83. Lewis, N.E.; Hixson, K.K.; Conrad, T.M.; Lerman, J.A.; Charusanti, P.; Polpitiya, A.D.; Adkins, J.N.; Schramm, G.; Purvine, S.O.; Lopez-Ferrer, D.; et al. Omic Data from Evolved *E. Coli* Are Consistent with Computed Optimal Growth from Genome-scale Models. *Molecular Systems Biology* **2010**, *6*, 390, doi:10.1038/msb.2010.47.
84. Herrmann, H.A.; Dyson, B.C.; Vass, L.; Johnson, G.N.; Schwartz, J.-M. Flux Sampling Is a Powerful Tool to Study Metabolism under Changing Environmental Conditions. *npj Syst Biol Appl* **2019**, *5*, 32, doi:10.1038/s41540-019-0109-0.
85. Herrmann, H.A.; Schwartz, J.-M.; Johnson, G.N. Metabolic Acclimation—a Key to Enhancing Photosynthesis in Changing Environments? *Journal of Experimental Botany* **2019**, *70*, 3043–3056, doi:10.1093/jxb/erz157.
86. Van Rosmalen, R.P.; Moreno-Paz, S.; Duman-Özdamar, Z.E.; Suarez-Diez, M. CFSA: Comparative Flux Sampling Analysis as a Guide for Strain Design. *Metabolic Engineering Communications* **2024**, *19*, e00244, doi:10.1016/j.mec.2024.e00244.
87. Sertbas, M.; Ulgen, K.O. Genome-Scale Metabolic Modeling for Unraveling Molecular Mechanisms of High Threat Pathogens. *Front. Cell Dev. Biol.* **2020**, *8*, 566702, doi:10.3389/fcell.2020.566702.
88. Gelbach, P.E.; Cetin, H.; Finley, S.D. Flux Sampling in Genome-Scale Metabolic Modeling of Microbial Communities. *BMC Bioinformatics* **2024**, *25*, 45, doi:10.1186/s12859-024-05655-3.
89. Ankrah, N.Y.D.; Barker, B.E.; Song, J.; Wu, C.; McMullen, J.G.; Douglas, A.E. Predicted Metabolic Function of the Gut Microbiota of Drosophila Melanogaster. *mSystems* **2021**, *6*, doi:10.1128/mSystems.01369-20.
90. Hyduke, D.R.; Lewis, N.E.; Palsson, B.Ø. Analysis of Omics Data with Genome-Scale Models of Metabolism. *Mol. BioSyst.* **2013**, *9*, 167–174, doi:10.1039/C2MB25453K.
91. Nogales, J.; Agudo, L. A Practical Protocol for Integration of Transcriptomics Data into Genome-Scale Metabolic Reconstructions. In

Hydrocarbon and Lipid Microbiology Protocols; McGenity, T.J., Timmis, K.N., Nogales, B., Eds.; Springer Protocols Handbooks; Springer Berlin Heidelberg: Berlin, Heidelberg, 2015; pp. 135–152 ISBN 978-3-662-50430-7.
92. Machado, D.; Herrgård, M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Comput Biol* **2014**, *10*, e1003580, doi:10.1371/journal.pcbi.1003580.
93. Robaina EstÃ©vez, S.; Nikoloski, Z. Generalized Framework for Context-Specific Metabolic Model Extraction Methods. *Front. Plant Sci.* **2014**, *5*, doi:10.3389/fpls.2014.00491.
94. Schuster, S.; Pfeiffer, T.; Fell, D.A. Is Maximization of Molar Yield in Metabolic Networks Favoured by Evolution? *Journal of Theoretical Biology* **2008**, *252*, 497–504, doi:10.1016/j.jtbi.2007.12.008.
95. Adadi, R.; Volkmer, B.; Milo, R.; Heinemann, M.; Shlomi, T. Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters. *PLoS Comput Biol* **2012**, *8*, e1002575, doi:10.1371/journal.pcbi.1002575.
96. Hintermayer, S.B.; Weuster-Botz, D. Experimental Validation of in Silico Estimated Biomass Yields of *Pseudomonas Putida* KT2440. *Biotechnol. J.* **2017**, *12*, 1600720, doi:10.1002/biot.201600720.
97. Karp, P.D.; Weaver, D.; Latendresse, M. How Accurate Is Automated Gap Filling of Metabolic Models? *BMC Syst Biol* **2018**, *12*, 73, doi:10.1186/s12918-018-0593-7.
98. Pan, S.; Reed, J.L. Advances in Gap-Filling Genome-Scale Metabolic Models and Model-Driven Experiments Lead to Novel Metabolic Discoveries. *Current Opinion in Biotechnology* **2018**, *51*, 103–108, doi:10.1016/j.copbio.2017.12.012.
99. Reed, J.L.; Patel, T.R.; Chen, K.H.; Joyce, A.R.; Applebee, M.K.; Herring, C.D.; Bui, O.T.; Knight, E.M.; Fong, S.S.; Palsson, B.O. Systems Approach to Refining Genome Annotation. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17480–17484, doi:10.1073/pnas.0603364103.
100. Kumar, V.S.; Dasika, M.S.; Maranas, C.D. Optimization Based Automated Curation of Metabolic Reconstructions. *BMC Bioinformatics* **2007**, *8*, 212, doi:10.1186/1471-2105-8-212.
101. Kumar, V.S.; Maranas, C.D. GrowMatch: An Automated Method for Reconciling In Silico/In Vivo Growth Predictions. *PLoS Comput Biol* **2009**, *5*, e1000308, doi:10.1371/journal.pcbi.1000308.
102. Thiele, I.; Vlassis, N.; Fleming, R.M.T. FASTGAPFILL: Efficient Gap Filling in Metabolic Networks. *Bioinformatics* **2014**, *30*, 2529–2531, doi:10.1093/bioinformatics/btu321.
103. Bernstein, D.B.; Sulheim, S.; Almaas, E.; Segrè, D. Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis. *Genome Biol* **2021**, *22*, 64, doi:10.1186/s13059-021-02289-z.
104. Prigent, S.; Frioux, C.; Dittami, S.M.; Thiele, S.; Larhlimi, A.; Collet, G.; Gutknecht, F.; Got, J.; Eveillard, D.; Bourdon, J.; et al. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. *PLoS Comput Biol* **2017**, *13*, e1005276,

doi:10.1371/journal.pcbi.1005276.
105. Suthers, P.F.; Dinh, H.V.; Fatma, Z.; Shen, Y.; Chan, S.H.J.; Rabinowitz, J.D.; Zhao, H.; Maranas, C.D. Genome-Scale Metabolic Reconstruction of the Non-Model Yeast Issatchenkia Orientalis SD108 and Its Application to Organic Acids Production. *Metabolic Engineering Communications* **2020**, *11*, e00148, doi:10.1016/j.mec.2020.e00148.
106. Vitkin, E.; Shlomi, T. MIRAGE: A Functional Genomics-Based Approach for Metabolic Network Model Reconstruction and Its Application to Cyanobacteria Networks. *Genome Biol* **2012**, *13*, R111, doi:10.1186/gb-2012-13-11-r111.
107. Fritzemeier, C.J.; Hartleb, D.; Szappanos, B.; Papp, B.; Lercher, M.J. Erroneous Energy-Generating Cycles in Published Genome Scale Metabolic Networks: Identification and Removal. *PLoS Comput Biol* **2017**, *13*, e1005494, doi:10.1371/journal.pcbi.1005494.
108. Herbert, D.; Elsworth, R.; Telling, R.C. The Continuous Culture of Bacteria; a Theoretical and Experimental Study. *Journal of General Microbiology* **1956**, *14*, 601–622, doi:10.1099/00221287-14-3-601.
109. Fonseca, G.G.; Gombert, A.K.; Heinzle, E.; Wittmann, C. Physiology of the Yeast *Kluyveromyces Marxianus* during Batch and Chemostat Cultures with Glucose as the Sole Carbon Source. *FEMS Yeast Research* **2007**, *7*, 422–435, doi:10.1111/j.1567-1364.2006.00192.x.
110. Sauer, U.; Lasko, D.R.; Fiaux, J.; Hochuli, M.; Glaser, R.; Szyperski, T.; Wüthrich, K.; Bailey, J.E. Metabolic Flux Ratio Analysis of Genetic and Environmental Modulations of *Escherichia Coli* Central Carbon Metabolism. *J Bacteriol* **1999**, *181*, 6679–6688, doi:10.1128/JB.181.21.6679-6688.1999.
111. Ventorim, R.Z.; Ferreira, M.A. de M.; de Almeida, E.L.M.; Kerkhoven, E.J.; da Silveira, W.B. Genome-Scale Metabolic Model of Oleaginous Yeast Papiliotrema Laurentii. *Biochemical Engineering Journal* **2022**, *180*, 108353, doi:10.1016/j.bej.2022.108353.
112. Viana, R.; Couceiro, D.; Carreiro, T.; Dias, O.; Rocha, I.; Teixeira, M.C. A Genome-Scale Metabolic Model for the Human Pathogen Candida Parapsilosis and Early Identification of Putative Novel Antifungal Drug Targets. *Genes* **2022**, *13*, 303, doi:10.3390/genes13020303.
113. Maarleveld, T.R.; Khandelwal, R.A.; Olivier, B.G.; Teusink, B.; Bruggeman, F.J. Basic Concepts and Principles of Stoichiometric Modeling of Metabolic Networks. *Biotechnology Journal* **2013**, *8*, 997–1008, doi:10.1002/biot.201200291.
114. Monk, J.M.; Charusanti, P.; Aziz, R.K.; Lerman, J.A.; Premyodhin, N.; Orth, J.D.; Feist, A.M.; Palsson, B.Ø. Genome-Scale Metabolic Reconstructions of Multiple *Escherichia Coli* Strains Highlight Strain-Specific Adaptations to Nutritional Environments. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 20338–20343, doi:10.1073/pnas.1307797110.
115. Seif, Y.; Kavvas, E.; Lachance, J.-C.; Yurkovich, J.T.; Nuccio, S.-P.; Fang, X.; Catoiu, E.; Raffatellu, M.; Palsson, B.O.; Monk, J.M. Genome-Scale Metabolic Reconstructions of Multiple Salmonella Strains Reveal Serovar-Specific Metabolic Traits. *Nat Commun* **2018**, *9*, 3771, doi:10.1038/s41467-018-06112-5.

116. Bosi, E.; Monk, J.M.; Aziz, R.K.; Fondi, M.; Nizet, V.; Palsson, B.Ø. Comparative Genome-Scale Modelling of *Staphylococcus Aureus* Strains Identifies Strain-Specific Metabolic Capabilities Linked to Pathogenicity. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, doi:10.1073/pnas.1523199113.
117. Viti, C.; Decorosi, F.; Marchi, E.; Galardini, M.; Giovannetti, L. High-Throughput Phenomics. In *Bacterial Pangenomics*; Mengoni, A., Galardini, M., Fondi, M., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2015; Vol. 1231, pp. 99–123 ISBN 978-1-4939-1719-8.
118. DeNittis, M.; Querol, A.; Zanoni, B.; Minati, J.L.; Ambrosoli, R. Possible Use of Biolog Methodology for Monitoring Yeast Presence in Alcoholic Fermentation for Wine-Making. *Journal of Applied Microbiology* **2010**, *108*, 1199–1206, doi:10.1111/j.1365-2672.2009.04547.x.
119. Sofo, A.; Ricciuti, P. A Standardized Method for Estimating the Functional Diversity of Soil Bacterial Community by Biolog® EcoPlatesTM Assay—The Case Study of a Sustainable Olive Orchard. *Applied Sciences* **2019**, *9*, 4035, doi:10.3390/app9194035.
120. Pylak, M.; Oszust, K.; Frąc, M. Searching for New Beneficial Bacterial Isolates of Wild Raspberries for Biocontrol of Phytopathogens-Antagonistic Properties and Functional Characterization. *IJMS* **2020**, *21*, 9361, doi:10.3390/ijms21249361.
121. Gao, C.-H.; Cao, H.; Cai, P.; Sørensen, S.J. The Initial Inoculation Ratio Regulates Bacterial Coculture Interactions and Metabolic Capacity. *ISME J* **2021**, *15*, 29–40, doi:10.1038/s41396-020-00751-7.
122. Oh, Y.-K.; Palsson, B.O.; Park, S.M.; Schilling, C.H.; Mahadevan, R. Genome-Scale Reconstruction of Metabolic Network in Bacillus Subtilis Based on High-Throughput Phenotyping and Gene Essentiality Data. *Journal of Biological Chemistry* **2007**, *282*, 28791–28799, doi:10.1074/jbc.M703759200.
123. Zhu, Y.; Czauderna, T.; Zhao, J.; Klapperstueck, M.; Maifiah, M.H.M.; Han, M.-L.; Lu, J.; Sommer, B.; Velkov, T.; Lithgow, T.; et al. Genome-Scale Metabolic Modeling of Responses to Polymyxins in *Pseudomonas Aeruginosa*. *GigaScience* **2018**, *7*, doi:10.1093/gigascience/giy021.
124. Biondi, E.G.; Tatti, E.; Comparini, D.; Giuntini, E.; Mocali, S.; Giovannetti, L.; Bazzicalupo, M.; Mengoni, A.; Viti, C. Metabolic Capacity of *Sinorhizobium* (*Ensifer*) *Meliloti* Strains as Determined by Phenotype MicroArray Analysis. *Appl Environ Microbiol* **2009**, *75*, 5396–5404, doi:10.1128/AEM.00196-09.
125. diCenzo, G.C.; Checcucci, A.; Bazzicalupo, M.; Mengoni, A.; Viti, C.; Dziewit, L.; Finan, T.M.; Galardini, M.; Fondi, M. Metabolic Modelling Reveals the Specialization of Secondary Replicons for Niche Adaptation in Sinorhizobium Meliloti. *Nat Commun* **2016**, *7*, 12219, doi:10.1038/ncomms12219.
126. Line, J.E.; Hiett, K.L.; Guard-Bouldin, J.; Seal, B.S. Differential Carbon Source Utilization by Campylobacter Jejuni 11168 in Response to Growth Temperature Variation. *Journal of Microbiological Methods* **2010**, *80*, 198–202, doi:10.1016/j.mimet.2009.12.011.
127. Wagley, S.; Newcombe, J.; Laing, E.; Yusuf, E.; Sambles, C.M.; Studholme,

D.J.; La Ragione, R.M.; Titball, R.W.; Champion, O.L. Differences in Carbon Source Utilisation Distinguish Campylobacter Jejuni from Campylobacter Coli. *BMC Microbiol* **2014**, *14*, 262, doi:10.1186/s12866-014-0262-y.
128. Buckner, M.M.C.; Blair, J.M.A.; La Ragione, R.M.; Newcombe, J.; Dwyer, D.J.; Ivens, A.; Piddock, L.J.V. Beyond Antimicrobial Resistance: Evidence for a Distinct Role of the AcrD Efflux Pump in *Salmonella* Biology. *mBio* **2016**, *7*, e01916-16, doi:10.1128/mBio.01916-16.
129. Vezina, B.; Watts, S.C.; Hawkey, J.; Cooper, H.B.; Judd, L.M.; Jenney, A.W.J.; Monk, J.M.; Holt, K.E.; Wyres, K.L. *Bactabolize: A Tool for High-Throughput Generation of Bacterial Strain-Specific Metabolic Models*; elife, 2023;
130. Henry, C.S.; Rotman, E.; Lathem, W.W.; Tyo, K.E.J.; Hauser, A.R.; Mandel, M.J. Generation and Validation of the iKp1289 Metabolic Model for Klebsiella Pneumoniae KPPR1. *The Journal of Infectious Diseases* **2017**, *215*, S37–S43, doi:10.1093/infdis/jiw465.
131. Renz, A.; Widerspick, L.; Dräger, A. First Genome-Scale Metabolic Model of Dolosigranulum Pigrum Confirms Multiple Auxotrophies. *Metabolites* **2021**, *11*, 232, doi:10.3390/metabo11040232.
132. Wegkamp, A.; Teusink, B.; De Vos, W.M.; Smid, E.J. Development of a Minimal Growth Medium for Lactobacillus Plantarum: Minimal Medium for Lactobacillus Plantarum. *Letters in Applied Microbiology* **2010**, *50*, 57–64, doi:10.1111/j.1472-765X.2009.02752.x.
133. Pearcy, N.; Garavaglia, M.; Millat, T.; Gilbert, J.P.; Song, Y.; Hartman, H.; Woods, C.; Tomi-Andrino, C.; Reddy Bommareddy, R.; Cho, B.-K.; et al. A Genome-Scale Metabolic Model of Cupriavidus Necator H16 Integrated with TraDIS and Transcriptomic Data Reveals Metabolic Insights for Biotechnological Applications. *PLoS Comput Biol* **2022**, *18*, e1010106, doi:10.1371/journal.pcbi.1010106.
134. Hawkey, J.; Vezina, B.; Monk, J.M.; Judd, L.M.; Harshegyi, T.; López-Fernández, S.; Rodrigues, C.; Brisse, S.; Holt, K.E.; Wyres, K.L. A Curated Collection of *Klebsiella* Metabolic Models Reveals Variable Substrate Usage and Gene Essentiality. *Genome Res.* **2022**, genome;gr.276289.121v2, doi:10.1101/gr.276289.121.
135. Fang, X.; Monk, J.M.; Nurk, S.; Akseshina, M.; Zhu, Q.; Gemmell, C.; Gianetto-Hill, C.; Leung, N.; Szubin, R.; Sanders, J.; et al. Metagenomics-Based, Strain-Level Analysis of Escherichia Coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. *Front. Microbiol.* **2018**, *9*, 2559, doi:10.3389/fmicb.2018.02559.
136. Norsigian, C.J.; Attia, H.; Szubin, R.; Yassin, A.S.; Palsson, B.Ø.; Aziz, R.K.; Monk, J.M. Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase–Producing Multidrug-Resistant Klebsiella Pneumoniae Clinical Isolates. *Front. Cell. Infect. Microbiol.* **2019**, *9*, 161, doi:10.3389/fcimb.2019.00161.
137. Prigent, S.; Nielsen, J.C.; Frisvad, J.C.; Nielsen, J. Reconstruction of 24 *Penicillium* Genome-scale Metabolic Models Shows Diversity Based on Their Secondary Metabolism. *Biotechnology and Bioengineering* **2018**, *115*, 2604–2612, doi:10.1002/bit.26739.
138. Nielsen, J.C.; Grijseels, S.; Prigent, S.; Ji, B.; Dainat, J.; Nielsen, K.F.; Frisvad,

J.C.; Workman, M.; Nielsen, J. Global Analysis of Biosynthetic Gene Clusters Reveals Vast Potential of Secondary Metabolite Production in Penicillium Species. *Nat Microbiol* **2017**, *2*, 17044, doi:10.1038/nmicrobiol.2017.44.
139. Cruz, F.; Capela, J.; Ferreira, E.C.; Rocha, M.; Dias, O. BioISO *: An Objective-Oriented Application for Assisting the Curation of Genome-Scale Metabolic Models*; Systems Biology, 2021;
140. Juty, N.; Le Novere, N.; Laibe, C. Identifiers.Org and MIRIAM Registry: Community Resources to Provide Persistent Identification. *Nucleic Acids Research* **2012**, *40*, D580–D586, doi:10.1093/nar/gkr1097.
141. Renz, A.; Dräger, A. Curating and Comparing 114 Strain-Specific Genome-Scale Metabolic Models of Staphylococcus Aureus. *npj Syst Biol Appl* **2021**, *7*, 30, doi:10.1038/s41540-021-00188-4.
142. Gautam, J.; Xu, Z. Construction and Validation of a Genome-Scale Metabolic Network of Thermotoga Sp. Strain RQ7. *Appl Biochem Biotechnol* **2021**, *193*, 896–911, doi:10.1007/s12010-020-03470-z.
143. Klanchui, A.; Dulsawat, S.; Chaloemngam, K.; Cheevadhanarak, S.; Prommeenate, P.; Meechai, A. An Improved Genome-Scale Metabolic Model of Arthrospira Platensis C1 (iAK888) and Its Application in Glycogen Overproduction. *Metabolites* **2018**, *8*, 84, doi:10.3390/metabo8040084.
144. Agren, R.; Liu, L.; Shoaie, S.; Vongsangnak, W.; Nookaew, I.; Nielsen, J. The RAVEN Toolbox and Its Use for Generating a Genome-Scale Metabolic Model for Penicillium Chrysogenum. *PLoS Comput Biol* **2013**, *9*, e1002980, doi:10.1371/journal.pcbi.1002980.
145. Dias, O.; Rocha, M.; Ferreira, E.C.; Rocha, I. Reconstructing Genome-Scale Metabolic Models with Merlin. *Nucleic Acids Research* **2015**, *43*, 3899–3910, doi:10.1093/nar/gkv294.
146. Capela, J.; Lagoa, D.; Rodrigues, R.; Cunha, E.; Cruz, F.; Barbosa, A.; Bastos, J.; Lima, D.; Ferreira, E.C.; Rocha, M.; et al. *Merlin* , an Improved Framework for the Reconstruction of High-Quality Genome-Scale Metabolic Models. *Nucleic Acids Research* **2022**, *50*, 6052–6066, doi:10.1093/nar/gkac459.
147. Wang, H.; Marcišauskas, S.; Sánchez, B.J.; Domenzain, I.; Hermansson, D.; Agren, R.; Nielsen, J.; Kerkhoven, E.J. RAVEN 2.0: A Versatile Toolbox for Metabolic Network Reconstruction and a Case Study on Streptomyces Coelicolor. *PLoS Comput Biol* **2018**, *14*, e1006541, doi:10.1371/journal.pcbi.1006541.
148. Notebaart, R.A.; van Enckevort, F.H.; Francke, C.; Siezen, R.J.; Teusink, B. Accelerating the Reconstruction of Genome-Scale Metabolic Networks. *BMC Bioinformatics* **2006**, *7*, 296, doi:10.1186/1471-2105-7-296.
149. Baroukh, C.; Cottret, L.; Pires, E.; Peyraud, R.; Guidot, A.; Genin, S. Insights into the Metabolic Specificities of Pathogenic Strains from the *Ralstonia Solanacearum* Species Complex. *mSystems* **2023**, e00083-23, doi:10.1128/msystems.00083-23.
150. Battjes, J.; Melkonian, C.; Mendoza, S.N.; Haver, A.; Al-Nakeeb, K.; Koza, A.; Schrubbers, L.; Wagner, M.; Zeidan, A.A.; Molenaar, D.; et al. Ethanol-Lactate Transition of Lachancea Thermotolerans Is Linked to

Nitrogen Metabolism. *Food Microbiology* **2023**, *110*, 104167, doi:10.1016/j.fm.2022.104167.
151. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **2010**, *11*, 119, doi:10.1186/1471-2105-11-119.
152. Blázquez, B.; San León, D.; Rojas, A.; Tortajada, M.; Nogales, J. New Insights on Metabolic Features of Bacillus Subtilis Based on Multistrain Genome-Scale Metabolic Modeling. *IJMS* **2023**, *24*, 7091, doi:10.3390/ijms24087091.
153. Nursimulu, N.; Moses, A.M.; Parkinson, J. Architect: A Tool for Aiding the Reconstruction of High-Quality Metabolic Models through Improved Enzyme Annotation. *PLoS Comput Biol* **2022**, *18*, e1010452, doi:10.1371/journal.pcbi.1010452.
154. Jenior, M.L.; Glass, E.M.; Papin, J.A. Reconstructor: A COBRApy Compatible Tool for Automated Genome-Scale Metabolic Network Reconstruction with Parsimonious Flux-Based Gap-Filling. *Bioinformatics* **2023**, *39*, btad367, doi:10.1093/bioinformatics/btad367.
155. Tarzi, C.; Zampieri, G.; Sullivan, N.; Angione, C. Emerging Methods for Genome-Scale Metabolic Modeling of Microbial Communities. *Trends in Endocrinology & Metabolism* **2024**, S1043276024000626, doi:10.1016/j.tem.2024.02.018.
156. Gu, C.; Kim, G.B.; Kim, W.J.; Kim, H.U.; Lee, S.Y. Current Status and Applications of Genome-Scale Metabolic Models. *Genome Biol* **2019**, *20*, 121, doi:10.1186/s13059-019-1730-3.
157. Machado, D.; Maistrenko, O.M.; Andrejev, S.; Kim, Y.; Bork, P.; Patil, K.R.; Patil, K.R. Polarization of Microbial Communities between Competitive and Cooperative Metabolism. *Nat Ecol Evol* **2021**, *5*, 195–203, doi:10.1038/s41559-020-01353-4.
158. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat Methods* **2015**, *12*, 59–60, doi:10.1038/nmeth.3176.
159. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nat Methods* **2021**, *18*, 366–368, doi:10.1038/s41592-021-01101-x.
160. Fahimipour, A.K.; Gross, T. Mapping the Bacterial Metabolic Niche Space. *Nat Commun* **2020**, *11*, 4887, doi:10.1038/s41467-020-18695-z.
161. Blasche, S.; Kim, Y.; Mars, R.A.T.; Machado, D.; Maansson, M.; Kafkia, E.; Milanese, A.; Zeller, G.; Teusink, B.; Nielsen, J.; et al. Metabolic Cooperation and Spatiotemporal Niche Partitioning in a Kefir Microbial Community. *Nat Microbiol* **2021**, *6*, 196–208, doi:10.1038/s41564-020-00816-5.
162. Schäfer, M.; Pacheco, A.R.; Künzler, R.; Bortfeld-Miller, M.; Field, C.M.; Vayena, E.; Hatzimanikatis, V.; Vorholt, J.A. Metabolic Interaction Models Recapitulate Leaf Microbiota Ecology. *Science* **2023**, *381*, eadf5121, doi:10.1126/science.adf5121.
163. Melkonian, C.; Zorrilla, F.; Kjærbølling, I.; Blasche, S.; Machado, D.; Junge, M.; Sørensen, K.I.; Andersen, L.T.; Patil, K.R.; Zeidan, A.A. Microbial Interactions Shape Cheese Flavour Formation. *Nat Commun* **2023**, *14*,

8348, doi:10.1038/s41467-023-41059-2.
164. Burcham, Z.M.; Belk, A.D.; McGivern, B.B.; Bouslimani, A.; Ghadermazi, P.; Martino, C.; Shenhav, L.; Zhang, A.R.; Shi, P.; Emmons, A.; et al. A Conserved Interdomain Microbial Network Underpins Cadaver Decomposition despite Environmental Variables. *Nat Microbiol* **2024**, *9*, 595–613, doi:10.1038/s41564-023-01580-y.
165. Peng, X.; Wang, S.; Wang, M.; Feng, K.; He, Q.; Yang, X.; Hou, W.; Li, F.; Zhao, Y.; Hu, B.; et al. Metabolic Interdependencies in Thermophilic Communities Are Revealed Using Co-Occurrence and Complementarity Networks. *Nat Commun* **2024**, *15*, 8166, doi:10.1038/s41467-024-52532-x.
166. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E.H.; Britto, R.; Bye-A-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **2023**, *51*, D523–D531, doi:10.1093/nar/gkac1052.
167. Nogales, J.; Mueller, J.; Gudmundsson, S.; Canalejo, F.J.; Duque, E.; Monk, J.; Feist, A.M.; Ramos, J.L.; Niu, W.; Palsson, B.O. High-quality Genome-scale Metabolic Modelling of *Pseudomonas Putida* Highlights Its Broad Metabolic Capabilities. *Environ Microbiol* **2020**, *22*, 255–269, doi:10.1111/1462-2920.14843.
168. Neal, M.; Brakewood, W.; Betenbaugh, M.; Zengler, K. Pan-Genome-Scale Metabolic Modeling of *Bacillus Subtilis* Reveals Functionally Distinct Groups. *mSystems* **2024**, e00923-24, doi:10.1128/msystems.00923-24.
169. De Bernardini, N.; Zampieri, G.; Campanaro, S.; Zimmermann, J.; Waschina, S.; Treu, L. Pan-Draft: Automated Reconstruction of Species-Representative Metabolic Models from Multiple Genomes. *Genome Biol* **2024**, *25*, 280, doi:10.1186/s13059-024-03425-1.
170. Domingo-Sananes, M.R.; McInerney, J.O. Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology* **2021**, *29*, 493–503, doi:10.1016/j.tim.2020.12.004.
171. Nysten, J.; Sofras, D.; Van Dijck, P. One Species, Many Faces: The Underappreciated Importance of Strain Diversity. *PLoS Pathog* **2024**, *20*, e1011931, doi:10.1371/journal.ppat.1011931.
172. Norsigian, C.J.; Fang, X.; Palsson, B.O.; Monk, J.M. Pangenome Flux Balance Analysis Toward Panphenomes. In *The Pangenome*; Tettelin, H., Medini, D., Eds.; Springer International Publishing: Cham, 2020; pp. 219–232 ISBN 978-3-030-38280-3.
173. Somerville, V.; Grigaitis, P.; Battjes, J.; Moro, F.; Teusink, B. Use and Limitations of Genome-Scale Metabolic Models in Food Microbiology. *Current Opinion in Food Science* **2022**, *43*, 225–231, doi:10.1016/j.cofs.2021.12.010.
174. Capozzi, V.; Russo, P.; Dueñas, M.T.; López, P.; Spano, G. Lactic Acid Bacteria Producing B-Group Vitamins: A Great Potential for Functional Cereals Products. *Appl Microbiol Biotechnol* **2012**, *96*, 1383–1394, doi:10.1007/s00253-012-4440-2.
175. Cooper, H.B.; Vezina, B.; Hawkey, J.; Passet, V.; López-Fernández, S.; Monk, J.M.; Brisse, S.; Holt, K.E.; Wyres, K.L. A Validated Pangenome-Scale Metabolic Model for the Klebsiella Pneumoniae Species

Complex. *Microbial Genomics* **2024**, *10*, doi:10.1099/mgen.0.001206.
176. Liao, Y.-C.; Huang, T.-W.; Chen, F.-C.; Charusanti, P.; Hong, J.S.J.; Chang, H.-Y.; Tsai, S.-F.; Palsson, B.O.; Hsiung, C.A. An Experimentally Validated Genome-Scale Metabolic Reconstruction of *Klebsiella Pneumoniae* MGH 78578, *i* YL1228. *J Bacteriol* **2011**, *193*, 1710–1717, doi:10.1128/JB.01218-10.
177. Monk, J.M. Genome-Scale Metabolic Network Reconstructions of Diverse *Escherichia* Strains Reveal Strain-Specific Adaptations. *Phil. Trans. R. Soc. B* **2022**, *377*, 20210236, doi:10.1098/rstb.2021.0236.
178. Sayers, E.W.; Bolton, E.E.; Brister, J.R.; Canese, K.; Chan, J.; Comeau, D.C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2022**, *50*, D20–D26, doi:10.1093/nar/gkab1112.
179. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.
180. Dimonaco, N.J.; Aubrey, W.; Kenobi, K.; Clare, A.; Creevey, C.J. No One Tool to Rule Them All: Prokaryotic Gene Prediction Tool Annotations Are Highly Dependent on the Organism of Study. *Bioinformatics* **2022**, *38*, 1198–1207, doi:10.1093/bioinformatics/btab827.
181. Flahaut, N.A.L.; Wiersma, A.; Van De Bunt, B.; Martens, D.E.; Schaap, P.J.; Sijtsma, L.; Dos Santos, V.A.M.; De Vos, W.M. Genome-Scale Metabolic Model for Lactococcus Lactis MG1363 and Its Application to the Analysis of Flavor Formation. *Appl Microbiol Biotechnol* **2013**, *97*, 8729–8739, doi:10.1007/s00253-013-5140-2.
182. Cantalapiedra, C.P.; Hernández-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **2021**, *38*, 5825–5829, doi:10.1093/molbev/msab293.
183. Casey, J.; Bennion, B.; D'haeseleer, P.; Kimbrel, J.; Marschmann, G.; Navid, A. Transporter Annotations Are Holding up Progress in Metabolic Modeling. *Front. Syst. Biol.* **2024**, *4*, 1394084, doi:10.3389/fsysb.2024.1394084.
184. Cunha, E.; Lagoa, D.; Faria, J.P.; Liu, F.; Henry, C.S.; Dias, O. *TranSyT* , an Innovative Framework for Identifying Transport Systems. *Bioinformatics* **2023**, *39*, btad466, doi:10.1093/bioinformatics/btad466.
185. Elbourne, L.D.H.; Wilson-Mortier, B.; Ren, Q.; Hassan, K.A.; Tetu, S.G.; Paulsen, I.T. TransAAP: An Automated Annotation Pipeline for Membrane Transporter Prediction in Bacterial Genomes. *Microbial Genomics* **2023**, *9*, doi:10.1099/mgen.0.000927.
186. Moretti, S.; Tran, V.D.T.; Mehl, F.; Ibberson, M.; Pagni, M. MetaNetX/MNXref: Unified Namespace for Metabolites and Biochemical Reactions in the Context of Metabolic Models. *Nucleic Acids Research* **2021**, *49*, D570–D574, doi:10.1093/nar/gkaa992.
187. Wishart, D.S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research* **2022**, *50*, D622–D631,

doi:10.1093/nar/gkab1062.
188. Bansal, P.; Morgat, A.; Axelsen, K.B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T.B.; et al. Rhea, the Reaction Knowledgebase in 2022. *Nucleic Acids Research* **2022**, *50*, D693–D700, doi:10.1093/nar/gkab1016.
189. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res* **2016**, *44*, D1214–D1219, doi:10.1093/nar/gkv1031.
190. Granger, B.E.; Perez, F. Jupyter: Thinking and Storytelling With Code and Data. *Comput. Sci. Eng.* **2021**, *23*, 7–14, doi:10.1109/MCSE.2021.3059263.
191. Zboralski, A.; Biessy, A.; Savoie, M.-C.; Novinscak, A.; Filion, M. Metabolic and Genomic Traits of Phytobeneficial Phenazine-Producing *Pseudomonas* Spp. Are Linked to Rhizosphere Colonization in *Arabidopsis Thaliana* and *Solanum Tuberosum*. *Appl Environ Microbiol* **2020**, *86*, e02443-19, doi:10.1128/AEM.02443-19.
192. Chung, C.H.; Lin, D.-W.; Eames, A.; Chandrasekaran, S. Next-Generation Genome-Scale Metabolic Modeling through Integration of Regulatory Mechanisms. *Metabolites* **2021**, *11*, 606, doi:10.3390/metabo11090606.
193. Courtot, M.; Juty, N.; Knüpfer, C.; Waltemath, D.; Zhukova, A.; Dräger, A.; Dumontier, M.; Finney, A.; Golebiewski, M.; Hastings, J.; et al. Controlled Vocabularies and Semantics in Systems Biology. *Molecular Systems Biology* **2011**, *7*, 543, doi:10.1038/msb.2011.77.
194. Biessy, A.; Novinscak, A.; Blom, J.; Léger, G.; Thomashow, L.S.; Cazorla, F.M.; Josic, D.; Filion, M. Diversity of Phytobeneficial Traits Revealed by Whole-genome Analysis of Worldwide-isolated Phenazine-producing *Pseudomonas* Spp. *Environmental Microbiology* **2019**, *21*, 437–455, doi:10.1111/1462-2920.14476.
195. Nickel, S.; Steinhardt, C.; Schlenker, H.; Burkart, W. IBM ILOG CPLEX Optimization Studio—A Primer. In *Decision Optimization with IBM ILOG CPLEX Optimization Studio*; Graduate Texts in Operations Research; Springer Berlin Heidelberg: Berlin, Heidelberg, 2022; pp. 9–21 ISBN 978-3-662-65480-4.
196. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069, doi:10.1093/bioinformatics/btu153.
197. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423, doi:10.1093/bioinformatics/btp163.
198. Peyraud, R.; Cottret, L.; Marmiesse, L.; Gouzy, J.; Genin, S. A Resource Allocation Trade-Off between Virulence and Proliferation Drives Metabolic Versatility in the Plant Pathogen Ralstonia Solanacearum. *PLoS Pathog* **2016**, *12*, e1005939, doi:10.1371/journal.ppat.1005939.
199. Lu, H.; Kerkhoven, E.J.; Nielsen, J. A Pan-Draft Metabolic Model Reflects Evolutionary Diversity across 332 Yeast Species. *Biomolecules* **2022**, *12*, 1632, doi:10.3390/biom12111632.
200. Brünker, P.; Altenbuchner, J.; Mattes, R. Structure and Function of the

Genes Involved in Mannitol, Arabitol and Glucitol Utilization from Pseudomonas Fluorescens DSM50106. *Gene* **1998**, *206*, 117–126, doi:10.1016/S0378-1119(97)00574-X.
201. Hoffmann, J.; Altenbuchner, J. Functional Characterization of the Mannitol Promoter of Pseudomonas Fluorescens DSM 50106 and Its Application for a Mannitol-Inducible Expression System for Pseudomonas Putida KT2440. *PLoS ONE* **2015**, *10*, e0133248, doi:10.1371/journal.pone.0133248.
202. Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **2016**, *11*, e0163962, doi:10.1371/journal.pone.0163962.
203. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150–3152, doi:10.1093/bioinformatics/bts565.
204. Manni, M.; Berkeley, M.R.; Seppey, M.; Simão, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **2021**, *38*, 4647–4654, doi:10.1093/molbev/msab199.
205. Rajput, A.; Chauhan, S.M.; Mohite, O.S.; Hyun, J.C.; Ardalani, O.; Jahn, L.J.; Sommer, M.Oa.; Palsson, B.O. Pangenome Analysis Reveals the Genetic Basis for Taxonomic Classification of the Lactobacillaceae Family. *Food Microbiology* **2023**, *115*, 104334, doi:10.1016/j.fm.2023.104334.
206. Wittig, U.; Kania, R.; Golebiewski, M.; Rey, M.; Shi, L.; Jong, L.; Algaa, E.; Weidemann, A.; Sauer-Danzwith, H.; Mir, S.; et al. SABIO-RK--Database for Biochemical Reaction Kinetics. *Nucleic Acids Research* **2012**, *40*, D790–D796, doi:10.1093/nar/gkr1046.
207. Conroy, M.J.; Andrews, R.M.; Andrews, S.; Cockayne, L.; Dennis, E.A.; Fahy, E.; Gaud, C.; Griffiths, W.J.; Jukes, G.; Kolchin, M.; et al. LIPID MAPS: Update to Databases and Tools for the Lipidomics Community. *Nucleic Acids Research* **2024**, *52*, D1677–D1682, doi:10.1093/nar/gkad896.
208. Wicker, J.; Lorsbach, T.; Gütlein, M.; Schmid, E.; Latino, D.; Kramer, S.; Fenner, K. enviPath – The Environmental Contaminant Biotransformation Pathway Resource. *Nucleic Acids Res* **2016**, *44*, D502–D508, doi:10.1093/nar/gkv1229.
209. Milacic, M.; Beavers, D.; Conley, P.; Gong, C.; Gillespie, M.; Griss, J.; Haw, R.; Jassal, B.; Matthews, L.; May, B.; et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research* **2024**, *52*, D672–D678, doi:10.1093/nar/gkad1025.
210. Aimo, L.; Liechti, R.; Hyka-Nouspikel, N.; Niknejad, A.; Gleizes, A.; Götz, L.; Kuznetsov, D.; David, F.P.A.; Van Der Goot, F.G.; Riezman, H.; et al. The SwissLipids Knowledgebase for Lipid Biology. *Bioinformatics* **2015**, *31*, 2860–2866, doi:10.1093/bioinformatics/btv285.
211. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **2015**, *7*, 23, doi:10.1186/s13321-015-0068-4.
212. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput.*

Sci. **1988**, *28*, 31–36, doi:10.1021/ci00057a005.
213. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2023 Update. *Nucleic Acids Research* **2023**, *51*, D1373–D1380, doi:10.1093/nar/gkac956.
214. Gasteiger, E. ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis. *Nucleic Acids Research* **2003**, *31*, 3784–3788, doi:10.1093/nar/gkg563.
215. Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Research* **2021**, *49*, D498–D508, doi:10.1093/nar/gkaa1025.
216. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res* **2016**, *44*, D457–D462, doi:10.1093/nar/gkv1070.
217. Bateman, A. The Pfam Protein Families Database. *Nucleic Acids Research* **2004**, *32*, 138D – 141, doi:10.1093/nar/gkh121.
218. Hernández-Plaza, A.; Szklarczyk, D.; Botas, J.; Cantalapiedra, C.P.; Giner-Lamia, J.; Mende, D.R.; Kirsch, R.; Rattei, T.; Letunic, I.; Jensen, L.J.; et al. eggNOG 6.0: Enabling Comparative Genomics across 12 535 Organisms. *Nucleic Acids Research* **2023**, *51*, D389–D394, doi:10.1093/nar/gkac1022.
219. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65, doi:10.1016/0377-0427(87)90125-7.
220. Douglas, A.E. Lessons from Studying Insect Symbioses. *Cell Host & Microbe* **2011**, *10*, 359–367, doi:10.1016/j.chom.2011.09.001.
221. Noman, M.S.; Liu, L.; Bai, Z.; Li, Z. Tephritidae Bacterial Symbionts: Potentials for Pest Management. *Bull. Entomol. Res.* **2020**, *110*, 1–14, doi:10.1017/S0007485319000403.
222. Rupawate, P.S.; Roylawar, P.; Khandagale, K.; Gawande, S.; Ade, A.B.; Jaiswal, D.K.; Borgave, S. Role of Gut Symbionts of Insect Pests: A Novel Target for Insect-Pest Control. *Front. Microbiol.* **2023**, *14*, 1146390, doi:10.3389/fmicb.2023.1146390.
223. Shamjana, U.; Vasu, D.A.; Hembrom, P.S.; Nayak, K.; Grace, T. The Role of Insect Gut Microbiota in Host Fitness, Detoxification and Nutrient Supplementation. *Antonie van Leeuwenhoek* **2024**, *117*, 71, doi:10.1007/s10482-024-01970-0.
224. Wernegreen, J.J. Genome Evolution in Bacterial Endosymbionts of Insects. *Nat Rev Genet* **2002**, *3*, 850–861, doi:10.1038/nrg931.
225. Moran, N.A. Microbial Minimalism. *Cell* **2002**, *108*, 583–586, doi:10.1016/S0092-8674(02)00665-7.
226. Moran, N.A.; Plague, G.R. Genomic Changes Following Host Restriction in Bacteria. *Current Opinion in Genetics & Development* **2004**, *14*, 627–633, doi:10.1016/j.gde.2004.09.003.
227. Moran, N.A.; McCutcheon, J.P.; Nakabachi, A. Genomics and Evolution of Heritable Bacterial Symbionts. *Annu. Rev. Genet.* **2008**, *42*, 165–190, doi:10.1146/annurev.genet.41.110306.130119.
228. Moya, A.; Peretó, J.; Gil, R.; Latorre, A. Learning How to Live Together:

Genomic Insights into Prokaryote–Animal Symbioses. *Nat Rev Genet* **2008**, *9*, 218–229, doi:10.1038/nrg2319.
229. Kikuchi, Y. Endosymbiotic Bacteria in Insects: Their Diversity and Culturability. *Microb. Environ.* **2009**, *24*, 195–204, doi:10.1264/jsme2.ME09140S.
230. Latorre, A.; Manzano-Marín, A. Dissecting Genome Reduction and Trait Loss in Insect Endosymbionts: Trait Loss in Insect Endosymbionts. *Ann. N.Y. Acad. Sci.* **2017**, *1389*, 52–75, doi:10.1111/nyas.13222.
231. McCutcheon, J.P.; McDonald, B.R.; Moran, N.A. Convergent Evolution of Metabolic Roles in Bacterial Co-Symbionts of Insects. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 15394–15399, doi:10.1073/pnas.0906424106.
232. Sabree, Z.L.; Huang, C.Y.; Okusu, A.; Moran, N.A.; Normark, B.B. The Nutrient Supplying Capabilities of *Uzinura* , an Endosymbiont of Armoured Scale Insects: Beneficial Endosymbiont of Armoured Scale Insects. *Environ Microbiol* **2013**, *15*, 1988–1999, doi:10.1111/1462-2920.12058.
233. Manzano-Marín, A.; Coeur D'Acier, A.; Clamens, A.-L.; Orvain, C.; Cruaud, C.; Barbe, V.; Jousselin, E. Serial Horizontal Transfer of Vitamin-Biosynthetic Genes Enables the Establishment of New Nutritional Symbionts in Aphids' Di-Symbiotic Systems. *ISME J* **2020**, *14*, 259–273, doi:10.1038/s41396-019-0533-6.
234. Masson, F.; Lemaitre, B. Growing Ungrowable Bacteria: Overview and Perspectives on Insect Symbiont Culturability. *Microbiol Mol Biol Rev* **2020**, *84*, e00089-20, doi:10.1128/MMBR.00089-20.
235. Hosokawa, T.; Kikuchi, Y.; Nikoh, N.; Shimada, M.; Fukatsu, T. Strict Host-Symbiont Cospeciation and Reductive Genome Evolution in Insect Gut Bacteria. *PLoS Biol* **2006**, *4*, e337, doi:10.1371/journal.pbio.0040337.
236. Capuzzo, C.; Firrao, G.; Mazzon, L.; Squartini, A.; Girolami, V. 'Candidatus Erwinia Dacicola', a Coevolved Symbiotic Bacterium of the Olive Fly Bactrocera Oleae (Gmelin). *International Journal of Systematic and Evolutionary Microbiology* **2005**, *55*, 1641–1647, doi:10.1099/ijs.0.63653-0.
237. Bigiotti, G.; Sacchetti, P.; Pastorelli, R.; Lauzon, C.R.; Belcari, A. Bacterial Symbiosis in *Bactrocera Oleae* , an Achilles' Heel for Its Pest Control. *Insect Science* **2021**, *28*, 874–884, doi:10.1111/1744-7917.12835.
238. Pavlidi, N.; Gioti, A.; Wybouw, N.; Dermauw, W.; Ben-Yosef, M.; Yuval, B.; Jurkevich, E.; Kampouraki, A.; Van Leeuwen, T.; Vontas, J. Transcriptomic Responses of the Olive Fruit Fly Bactrocera Oleae and Its Symbiont Candidatus Erwinia Dacicola to Olive Feeding. *Sci Rep* **2017**, *7*, 42633, doi:10.1038/srep42633.
239. Ben-Yosef, M.; Pasternak, Z.; Jurkevitch, E.; Yuval, B. Symbiotic Bacteria Enable Olive Fly Larvae to Overcome Host Defences. *R. Soc. open sci.* **2015**, *2*, 150170, doi:10.1098/rsos.150170.
240. Siden-Kiamos, I.; Koidou, V.; Livadaras, I.; Skoufa, E.; Papadogiorgaki, S.; Papadakis, S.; Chalepakis, G.; Ioannidis, P.; Vontas, J. Dynamic Interactions between the Symbiont Candidatus Erwinia Dacicola and Its Olive Fruit Fly Host Bactrocera Oleae. *Insect Biochemistry and Molecular Biology* **2022**, *146*, 103793, doi:10.1016/j.ibmb.2022.103793.
241. Ben-Yosef, M.; Aharon, Y.; Jurkevitch, E.; Yuval, B. Give Us the Tools and

We Will Do the Job: Symbiotic Bacteria Affect Olive Fly Fitness in a Diet-Dependent Fashion. *Proc. R. Soc. B.* **2010**, *277*, 1545–1552, doi:10.1098/rspb.2009.2102.
242. Ben-Yosef, M.; Pasternak, Z.; Jurkevitch, E.; Yuval, B. Symbiotic Bacteria Enable Olive Flies (*Bactrocera Oleae*) to Exploit Intractable Sources of Nitrogen. *J. Evol. Biol.* **2014**, *27*, 2695–2705, doi:10.1111/jeb.12527.
243. Arora, A.K.; Douglas, A.E. Hype or Opportunity? Using Microbial Symbionts in Novel Strategies for Insect Pest Control. *Journal of Insect Physiology* **2017**, *103*, 10–17, doi:10.1016/j.jinsphys.2017.09.011.
244. Blow, F.; Gioti, A.; Goodhead, I.B.; Kalyva, M.; Kampouraki, A.; Vontas, J.; Darby, A.C. Functional Genomics of a Symbiotic Community: Shared Traits in the Olive Fruit Fly Gut Microbiota. *Genome Biology and Evolution* **2020**, *12*, 3778–3791, doi:10.1093/gbe/evz258.
245. Estes, A.M.; Hearn, D.J.; Bronstein, J.L.; Pierson, E.A. The Olive Fly Endosymbiont, " *Candidatus* Erwinia Dacicola," Switches from an Intracellular Existence to an Extracellular Existence during Host Insect Development. *Appl Environ Microbiol* **2009**, *75*, 7097–7106, doi:10.1128/AEM.00778-09.
246. Soutar, C.D.; Stavrinides, J. Phylogenomic Analysis of the Erwiniaceae Supports Reclassification of Kalamiella Piersonii to Pantoea Piersonii Comb. Nov. and Erwinia Gerundensis to the New Genus Duffyella Gen. Nov. as Duffyella Gerundensis Comb. Nov. *Mol Genet Genomics* **2022**, *297*, 213–225, doi:10.1007/s00438-021-01829-3.
247. Boscaro, V.; Kolisko, M.; Felletti, M.; Vannini, C.; Lynn, D.H.; Keeling, P.J. Parallel Genome Reduction in Symbionts Descended from Closely Related Free-Living Bacteria. *Nat Ecol Evol* **2017**, *1*, 1160–1167, doi:10.1038/s41559-017-0237-0.
248. Gutleben, J.; Chaib De Mares, M.; Van Elsas, J.D.; Smidt, H.; Overmann, J.; Sipkema, D. The Multi-Omics Promise in Context: From Sequence to Microbial Isolate. *Critical Reviews in Microbiology* **2018**, *44*, 212–229, doi:10.1080/1040841X.2017.1332003.
249. Fang, X.; Lloyd, C.J.; Palsson, B.O. Reconstructing Organisms in Silico: Genome-Scale Models and Their Emerging Applications. *Nat Rev Microbiol* **2020**, *18*, 731–743, doi:10.1038/s41579-020-00440-4.
250. Feist, A.M.; Palsson, B.O. The Biomass Objective Function. *Current Opinion in Microbiology* **2010**, *13*, 344–349, doi:10.1016/j.mib.2010.03.003.
251. O'Brien, E.J.; Monk, J.M.; Palsson, B.O. Using Genome-Scale Models to Predict Biological Capabilities. *Cell* **2015**, *161*, 971–987, doi:10.1016/j.cell.2015.05.019.
252. Monk, J.M.; Charusanti, P.; Aziz, R.K.; Lerman, J.A.; Premyodhin, N.; Orth, J.D.; Feist, A.M.; Palsson, B.Ø. Genome-Scale Metabolic Reconstructions of Multiple *Escherichia Coli* Strains Highlight Strain-Specific Adaptations to Nutritional Environments. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 20338–20343, doi:10.1073/pnas.1307797110.
253. Gu, C.; Kim, G.B.; Kim, W.J.; Kim, H.U.; Lee, S.Y. Current Status and Applications of Genome-Scale Metabolic Models. *Genome Biology* **2019**, *20*, 121, doi:10.1186/s13059-019-1730-3.
254. Thomas, G.H.; Zucker, J.; Macdonald, S.J.; Sorokin, A.; Goryanin, I.;

Douglas, A.E. A Fragile Metabolic Network Adapted for Cooperation in the Symbiotic Bacterium Buchnera Aphidicola. *BMC Syst Biol* **2009**, *3*, 24, doi:10.1186/1752-0509-3-24.
255. González-Domenech, C.M.; Belda, E.; Patiño-Navarrete, R.; Moya, A.; Peretó, J.; Latorre, A. Metabolic Stasis in an Ancient Symbiosis: Genome-Scale Metabolic Networks from Two Blattabacterium Cuenoti Strains, Primary Endosymbionts of Cockroaches. *BMC Microbiol* **2012**, *12*, S5, doi:10.1186/1471-2180-12-S1-S5.
256. Hall, R.J.; Flanagan, L.A.; Bottery, M.J.; Springthorpe, V.; Thorpe, S.; Darby, A.C.; Wood, A.J.; Thomas, G.H. A Tale of Three Species: Adaptation of Sodalis Glossinidius to Tsetse Biology, *Wigglesworthia* Metabolism, and Host Diet. *mBio* **2019**, *10*, e02106-18, doi:10.1128/mBio.02106-18.
257. Driscoll, T.P.; Verhoeve, V.I.; Guillotte, M.L.; Lehman, S.S.; Rennoll, S.A.; Beier-Sexton, M.; Rahman, M.S.; Azad, A.F.; Gillespie, J.J. Wholly *Rickettsia* ! Reconstructed Metabolic Profile of the Quintessential Bacterial Parasite of Eukaryotic Cells. *mBio* **2017**, *8*, e00859-17, doi:10.1128/mBio.00859-17.
258. Jiménez, N.E.; Gerdtzen, Z.P.; Olivera-Nappa, Á.; Salgado, J.C.; Conca, C. A Systems Biology Approach for Studying Wolbachia Metabolism Reveals Points of Interaction with Its Host in the Context of Arboviral Infection. *PLoS Negl Trop Dis* **2019**, *13*, e0007678, doi:10.1371/journal.pntd.0007678.
259. Ankrah, N.Y.D.; Chouaia, B.; Douglas, A.E. The Cost of Metabolic Interactions in Symbioses between Insects and Bacteria with Reduced Genomes. *mBio* **2018**, *9*, e01433-18, doi:10.1128/mBio.01433-18.
260. Ankrah, N.Y.D.; Wilkes, R.A.; Zhang, F.Q.; Zhu, D.; Kaweesi, T.; Aristilde, L.; Douglas, A.E. Syntrophic Splitting of Central Carbon Metabolism in Host Cells Bearing Functionally Different Symbiotic Bacteria. *The ISME Journal* **2020**, *14*, 1982–1993, doi:10.1038/s41396-020-0661-z.
261. Ponce-de-Leon, M.; Tamarit, D.; Calle-Espinosa, J.; Mori, M.; Latorre, A.; Montero, F.; Pereto, J. Determinism and Contingency Shape Metabolic Complementation in an Endosymbiotic Consortium. *Front. Microbiol.* **2017**, *8*, 2290, doi:10.3389/fmicb.2017.02290.
262. Calle-Espinosa, J.; Ponce-de-Leon, M.; Santos-Garcia, D.; Silva, F.J.; Montero, F.; Peretó, J. Nature Lessons: The Whitefly Bacterial Endosymbiont Is a Minimal Amino Acid Factory with Unusual Energetics. *Journal of Theoretical Biology* **2016**, *407*, 303–317, doi:10.1016/j.jtbi.2016.07.024.
263. Ankrah, N.Y.D.; Luan, J.; Douglas, A.E. Cooperative Metabolism in a Three-Partner Insect-Bacterial Symbiosis Revealed by Metabolic Modeling. *J Bacteriol* **2017**, *199*, doi:10.1128/JB.00872-16.
264. Blow, F.; Ankrah, N.Y.D.; Clark, N.; Koo, I.; Allman, E.L.; Liu, Q.; Anitha, M.; Patterson, A.D.; Douglas, A.E. Impact of Facultative Bacteria on the Metabolic Function of an Obligate Insect-Bacterial Symbiosis. *mBio* **2020**, *11*, e00402-20, doi:10.1128/mBio.00402-20.
265. Ponce-de-León, M.; Montero, F.; Peretó, J. Solving Gap Metabolites and Blocked Reactions in Genome-Scale Models: Application to the Metabolic Network of Blattabacterium Cuenoti. *BMC Syst Biol* **2013**, *7*, 114, doi:10.1186/1752-0509-7-114.

266. Pan, S.; Reed, J.L. Advances in Gap-Filling Genome-Scale Metabolic Models and Model-Driven Experiments Lead to Novel Metabolic Discoveries. *Current Opinion in Biotechnology* **2018**, *51*, 103–108, doi:10.1016/j.copbio.2017.12.012.
267. Grula, E.A. CELL DIVISION IN A SPECIES OF *ERWINIA* II: Inhibition of Division by d-Amino Acids. *J Bacteriol* **1960**, *80*, 375–385, doi:10.1128/jb.80.3.375-385.1960.
268. Machado, D.; Andrejev, S.; Tramontano, M.; Patil, K.R. Fast Automated Reconstruction of Genome-Scale Metabolic Models for Microbial Species and Communities. *Nucleic Acids Research* **2018**, *46*, 7542–7553, doi:10.1093/nar/gky537.
269. Henry, C.S.; DeJongh, M.; Best, A.A.; Frybarger, P.M.; Linsay, B.; Stevens, R.L. High-Throughput Generation, Optimization and Analysis of Genome-Scale Metabolic Models. *Nat Biotechnol* **2010**, *28*, 977–982, doi:10.1038/nbt.1672.
270. Zimmermann, J.; Kaleta, C.; Waschina, S. Gapseq: Informed Prediction of Bacterial Metabolic Pathways and Reconstruction of Accurate Metabolic Models. *Genome Biology* **2021**, *22*, 81, doi:10.1186/s13059-021-02295-1.
271. Webb, E.; Claas, K.; Downs, D. thiBPQ Encodes an ABC Transporter Required for Transport of Thiamine and Thiamine Pyrophosphate inSalmonella Typhimurium. *Journal of Biological Chemistry* **1998**, *273*, 8946–8950, doi:10.1074/jbc.273.15.8946.
272. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res* **2016**, *44*, D457–D462, doi:10.1093/nar/gkv1070.
273. Wang, X.; Quinn, P.J.; Yan, A. Kdo $_2$ -lipid A: Structural Diversity and Impact on Immunopharmacology. *Biological Reviews* **2015**, *90*, 408–427, doi:10.1111/brv.12114.
274. Jiang, M.; Cao, Y.; Guo, Z.-F.; Chen, M.; Chen, X.; Guo, Z. Menaquinone Biosynthesis in *Escherichia Coli* : Identification of 2-Succinyl-5-Enolpyruvyl-6-Hydroxy-3-Cyclohexene-1-Carboxylate as a Novel Intermediate and Re-Evaluation of MenD Activity. *Biochemistry* **2007**, *46*, 10979–10989, doi:10.1021/bi700810x.
275. Ishino, F.; Park, W.; Tomioka, S.; Tamaki, S.; Takase, I.; Kunugita, K.; Matsuzawa, H.; Asoh, S.; Ohta, T.; Spratt, B.G. Peptidoglycan Synthetic Activities in Membranes of Escherichia Coli Caused by Overproduction of Penicillin-Binding Protein 2 and rodA Protein. *Journal of Biological Chemistry* **1986**, *261*, 7024–7031, doi:10.1016/S0021-9258(19)62717-1.
276. Shiri, Y.; Khodavirdipour, A.; Kalkali, N. Re-Construction of Co-Expression Network of Genes Involved in Bacterial Cell Wall Synthesis and Their Role in Penicillin Resistance. *Avicenna J Clin Microbiol Infect* **2020**, *7*, 65–71, doi:10.34172/ajcmi.2020.15.
277. Zapun, A.; Contreras-Martel, C.; Vernet, T. Penicillin-Binding Proteins and β-Lactam Resistance. *FEMS Microbiol Rev* **2008**, *32*, 361–385, doi:10.1111/j.1574-6976.2007.00095.x.
278. Ranjitkar, S.; Reck, F.; Ke, X.; Zhu, Q.; McEnroe, G.; Lopez, S.L.; Dean, C.R. Identification of Mutations in the *mrdA* Gene Encoding PBP2 That Reduce Carbapenem and Diazabicyclooctane Susceptibility of Escherichia Coli

Clinical Isolates with Mutations in *ftsI* (PBP3) and Which Carry *Bla* $_{NDM-1}$. *mSphere* **2019**, *4*, e00074-19, doi:10.1128/mSphere.00074-19.
279. Sinno, M.; Bézier, A.; Vinale, F.; Giron, D.; Laudonia, S.; Garonna, A.P.; Pennacchio, F. Symbiosis Disruption in the Olive Fruit Fly, *Bactrocera oleae* (Rossi), as a Potential Tool for Sustainable Control. *Pest Manag Sci* **2020**, *76*, 3199–3207, doi:10.1002/ps.5875.
280. Grula, E.A. CELL DIVISION IN A SPECIES OF *ERWINIA* II: Inhibition of Division by d-Amino Acids. *J Bacteriol* **1960**, *80*, 375–385, doi:10.1128/jb.80.3.375-385.1960.
281. Sayers, E.W.; Bolton, E.E.; Brister, J.R.; Canese, K.; Chan, J.; Comeau, D.C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2022**, *50*, D20–D26, doi:10.1093/nar/gkab1112.
282. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069, doi:10.1093/bioinformatics/btu153.
283. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.
284. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.G.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* **2015**, *31*, 3691–3693, doi:10.1093/bioinformatics/btv421.
285. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics* **2019**, *35*, 4453–4455, doi:10.1093/bioinformatics/btz305.
286. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423, doi:10.1093/bioinformatics/btp163.
287. King, Z.A.; Lu, J.; Dräger, A.; Miller, P.; Federowicz, S.; Lerman, J.A.; Ebrahim, A.; Palsson, B.O.; Lewis, N.E. BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models. *Nucleic Acids Res* **2016**, *44*, D515–D522, doi:10.1093/nar/gkv1049.
288. Marinos, G.; Kaleta, C.; Waschina, S. Defining the Nutritional Input for Genome-Scale Metabolic Models: A Roadmap. *PLoS ONE* **2020**, *15*, e0236890, doi:10.1371/journal.pone.0236890.
289. Fritzemeier, C.J.; Hartleb, D.; Szappanos, B.; Papp, B.; Lercher, M.J. Erroneous Energy-Generating Cycles in Published Genome Scale Metabolic Networks: Identification and Removal. *PLoS Comput Biol* **2017**, *13*, e1005494, doi:10.1371/journal.pcbi.1005494.
290. Ebrahim, A.; Lerman, J.A.; Palsson, B.O.; Hyduke, D.R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **2013**, *7*, 74, doi:10.1186/1752-0509-7-74.
291. Nickel, S.; Steinhardt, C.; Schlenker, H.; Burkart, W. IBM ILOG CPLEX Optimization Studio—A Primer. In *Decision Optimization with IBM ILOG*

CPLEX Optimization Studio; Graduate Texts in Operations Research; Springer Berlin Heidelberg: Berlin, Heidelberg, 2022; pp. 9–21 ISBN 978-3-662-65480-4.
292. Norsigian, C.J.; Fang, X.; Seif, Y.; Monk, J.M.; Palsson, B.O. A Workflow for Generating Multi-Strain Genome-Scale Metabolic Models of Prokaryotes. *Nat Protoc* **2020**, *15*, 1–14, doi:10.1038/s41596-019-0254-3.
293. King, Z.A.; Dräger, A.; Ebrahim, A.; Sonnenschein, N.; Lewis, N.E.; Palsson, B.O. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput Biol* **2015**, *11*, e1004321, doi:10.1371/journal.pcbi.1004321.
294. Kanehisa, M.; Sato, Y.; Kawashima, M. KEGG Mapping Tools for Uncovering Hidden Features in Biological Data. *Protein Science* **2022**, *31*, 47–53, doi:10.1002/pro.4172.
295. Bateman, A. The Pfam Protein Families Database. *Nucleic Acids Research* **2004**, *32*, 138D – 141, doi:10.1093/nar/gkh121.
296. Alcock, B.P.; Huynh, W.; Chalil, R.; Smith, K.W.; Raphenya, A.R.; Wlodarski, M.A.; Edalatmand, A.; Petkau, A.; Syed, S.A.; Tsang, K.K.; et al. CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research* **2023**, *51*, D690–D699, doi:10.1093/nar/gkac920.
297. Gupta, A.; Jeyakumar, E.; Lawrence, R. Journey of Limonene as an Antimicrobial Agent. *J. Pure Appl. Microbiol.* **2021**, *15*, 1094–1110, doi:10.22207/JPAM.15.3.01.
298. Bortolaia, V.; Kaas, R.S.; Ruppe, E.; Roberts, M.C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R.L.; Rebelo, A.R.; Florensa, A.F.; et al. ResFinder 4.0 for Predictions of Phenotypes from Genotypes. *Journal of Antimicrobial Chemotherapy* **2020**, *75*, 3491–3500, doi:10.1093/jac/dkaa345.
299. Wiegand, I.; Hilpert, K.; Hancock, R.E.W. Agar and Broth Dilution Methods to Determine the Minimal Inhibitory Concentration (MIC) of Antimicrobial Substances. *Nat Protoc* **2008**, *3*, 163–175, doi:10.1038/nprot.2007.521.
300. Madeo, M.; O'Riordan, N.; Fuchs, T.M.; Utratna, M.; Karatzas, K.A.G.; O'Byrne, C.P. Thiamine Plays a Critical Role in the Acid Tolerance of Listeria Monocytogenes. *FEMS Microbiol Lett* **2012**, *326*, 137–143, doi:10.1111/j.1574-6968.2011.02442.x.
301. Murugkar, P.; Dimise, E.; Stewart, E.; Viala, S.N.; Clardy, J.; Dewhirst, F.E.; Lewis, K. Identification of a Growth Factor Required for Culturing Specific Fastidious Oral Bacteria. *Journal of Oral Microbiology* **2023**, *15*, 2143651, doi:10.1080/20002297.2022.2143651.
302. Baker, G.C.; Smith, J.J.; Cowan, D.A. Review and Re-Analysis of Domain-Specific 16S Primers. *Journal of Microbiological Methods* **2003**, *55*, 541–555, doi:10.1016/j.mimet.2003.08.009.
303. Livadaras, I.; Koidou, V.; Pitsili, E.; Moustaka, J.; Vontas, J.; Siden-Kiamos, I. Stably Inherited Transfer of the Bacterial Symbiont *Candidatus* Erwinia Dacicola from Wild Olive Fruit Flies *Bactrocera Oleae* to a Laboratory Strain. *Bull. Entomol. Res.* **2021**, *111*, 379–384, doi:10.1017/S0007485321000031.

304. Walter, J.; Tannock, G.W.; Tilsala-Timisjarvi, A.; Rodtong, S.; Loach, D.M.; Munro, K.; Alatossava, T. Detection and Identification of Gastrointestinal *Lactobacillus* Species by Using Denaturing Gradient Gel Electrophoresis and Species-Specific PCR Primers. *Appl Environ Microbiol* **2000**, *66*, 297–303, doi:10.1128/AEM.66.1.297-303.2000.
305. Qaraleh, S.Y.; Karajeh, M.R.; Al-Ameiri, N.S. Molds Associated with Olive Fruits Infested with Olive Fruit Fly (Bactrocera Oleae) and Their Effects on Oil Quality. *JJBS* **2015**, *8*, 217–220, doi:10.12816/0026961.
306. Boersch, M.; Rudrawar, S.; Grant, G.; Zunk, M. Menaquinone Biosynthesis Inhibition: A Review of Advancements toward a New Antibiotic Mechanism. *RSC Adv.* **2018**, *8*, 5099–5105, doi:10.1039/C7RA12950E.
307. Zhang, Z.; Liu, L.; Liu, C.; Sun, Y.; Zhang, D. New Aspects of Microbial Vitamin K2 Production by Expanding the Product Spectrum. *Microb Cell Fact* **2021**, *20*, 84, doi:10.1186/s12934-021-01574-7.
308. Nitzschke, A.; Bettenbrock, K. All Three Quinone Species Play Distinct Roles in Ensuring Optimal Growth under Aerobic and Fermentative Conditions in E. Coli K12. *PLoS ONE* **2018**, *13*, e0194699, doi:10.1371/journal.pone.0194699.
309. McCutcheon, J.P.; Moran, N.A. Parallel Genomic Evolution and Metabolic Interdependence in an Ancient Symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19392–19397, doi:10.1073/pnas.0708855104.
310. Mao, M.; Yang, X.; Poff, K.; Bennett, G. Comparative Genomics of the Dual-Obligate Symbionts from the Treehopper, Entylia Carinata (Hemiptera: Membracidae), Provide Insight into the Origins and Evolution of an Ancient Symbiosis. *Genome Biology and Evolution* **2017**, *9*, 1803–1815, doi:10.1093/gbe/evx134.
311. Wang, N.; Trivedi, P. Citrus Huanglongbing: A Newly Relevant Disease Presents Unprecedented Challenges. *Phytopathology®* **2013**, *103*, 652–665, doi:10.1094/PHYTO-12-12-0331-RVW.
312. Merfa, M.V.; Pérez-López, E.; Naranjo, E.; Jain, M.; Gabriel, D.W.; De La Fuente, L. Progress and Obstacles in Culturing ' *Candidatus* Liberibacter Asiaticus', the Bacterium Associated with Huanglongbing. *Phytopathology®* **2019**, *109*, 1092–1101, doi:10.1094/PHYTO-02-19-0051-RVW.
313. Sazama, E.J.; Ouellette, S.P.; Wesner, J.S. Bacterial Endosymbionts Are Common Among, but Not Necessarily Within, Insect Species. *Environmental Entomology* **2019**, *48*, 127–133, doi:10.1093/ee/nvy188.
314. Otero-Bravo, A.; Sabree, Z.L. Multiple Concurrent and Convergent Stages of Genome Reduction in Bacterial Symbionts across a Stink Bug Family. *Sci Rep* **2021**, *11*, 7731, doi:10.1038/s41598-021-86574-8.
315. Shigenobu, S.; Watanabe, H.; Hattori, M.; Sakaki, Y.; Ishikawa, H. Genome Sequence of the Endocellular Bacterial Symbiont of Aphids Buchnera Sp. APS. *Nature* **2000**, *407*, 81–86, doi:10.1038/35024074.
316. Xu, X.; Li, C.; Cao, W.; Yan, L.; Cao, L.; Han, Q.; Gao, M.; Chen, Y.; Shen, Z.; Jiang, J.; et al. Bacterial Growth and Environmental Adaptation via Thiamine Biosynthesis and Thiamine-Mediated Metabolic Interactions. *The ISME Journal* **2024**, *18*, wrae157, doi:10.1093/ismejo/wrae157.
317. Martin Říhová, J.; Gupta, S.; Darby, A.C.; Nováková, E.; Hypša, V.

Arsenophonus Symbiosis with Louse Flies: Multiple Origins, Coevolutionary Dynamics, and Metabolic Significance. *mSystems* **2023**, *8*, e00706-23, doi:10.1128/msystems.00706-23.
318. Belda, E.; Moya, A.; Bentley, S.; Silva, F.J. Mobile Genetic Element Proliferation and Gene Inactivation Impact over the Genome Structure and Metabolic Capabilities of Sodalis Glossinidius, the Secondary Endosymbiont of Tsetse Flies. *BMC Genomics* **2010**, *11*, 449, doi:10.1186/1471-2164-11-449.
319. Koskinioti, P.; Ras, E.; Augustinos, A.A.; Beukeboom, L.W.; Mathiopoulos, K.D.; Caceres, C.; Bourtzis, K. Manipulation of Insect Gut Microbiota towards the Improvement of *Bactrocera Oleae* Artificial Rearing. *Entomologia Exp Applicata* **2020**, *168*, 523–540, doi:10.1111/eea.12934.
320. Thaochan, N.; Drew, R.A.I.; Hughes, J.M.; Vijaysegaran, S.; Chinajariyawong, A. Alimentary Tract Bacteria Isolated and Identified with API-20E and Molecular Cloning Techniques from Australian Tropical Fruit Flies, *Bactrocera Cacuminata* and *B. Tryoni*. *Journal of Insect Science* **2010**, *10*, 1–16, doi:10.1673/031.010.13101.
321. Wang, H.; Jin, L.; Zhang, H. Comparison of the Diversity of the Bacterial Communities in the Intestinal Tract of Adult Bactrocera Dorsalis from Three Different Populations: Bacterial Communities in B. Dorsalis Gut. *Journal of Applied Microbiology* **2011**, *110*, 1390–1401, doi:10.1111/j.1365-2672.2011.05001.x.
322. Liu, L.J.; Martinez-Sañudo, I.; Mazzon, L.; Prabhakar, C.S.; Girolami, V.; Deng, Y.L.; Dai, Y.; Li, Z.H. Bacterial Communities Associated with Invasive Populations of *Bactrocera Dorsalis* (Diptera: Tephritidae) in China. *Bull. Entomol. Res.* **2016**, *106*, 718–728, doi:10.1017/S0007485316000390.
323. Bai, Z.; Liu, L.; Noman, M.S.; Zeng, L.; Luo, M.; Li, Z. The Influence of Antibiotics on Gut Bacteria Diversity Associated with Laboratory-Reared *Bactrocera Dorsalis*. *Bull. Entomol. Res.* **2019**, *109*, 500–509, doi:10.1017/S0007485318000834.
324. Santona, M.; Sanna, M.L.; Multineddu, C.; Fancello, F.; De La Fuente, S.A.; Dettori, S.; Zara, S. Microbial Biodiversity of Sardinian Oleic Ecosystems. *Food Microbiology* **2018**, *70*, 65–75, doi:10.1016/j.fm.2017.09.004.
325. Fusté, E.; Galisteo, G.J.; Jover, L.; Vinuesa, T.; Villa, T.G.; Viñas, M. Comparison of Antibiotic Susceptibility of Old and Current *Serratia*. *Future Microbiol.* **2012**, *7*, 781–786, doi:10.2217/fmb.12.40.
326. Mendoza, S.N.; Olivier, B.G.; Molenaar, D.; Teusink, B. A Systematic Assessment of Current Genome-Scale Metabolic Reconstruction Tools. *Genome Biol* **2019**, *20*, 158, doi:10.1186/s13059-019-1769-1.
327. Manni, M.; Berkeley, M.R.; Seppey, M.; Simão, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **2021**, *38*, 4647–4654, doi:10.1093/molbev/msab199.
328. Giani, A.M.; Gallo, G.R.; Gianfranceschi, L.; Formenti, G. Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly. *Computational and Structural Biotechnology Journal* **2020**, *18*,

9–19, doi:10.1016/j.csbj.2019.11.002.
329. Tutar, Y. Pseudogenes. *Comparative and Functional Genomics* **2012**, *2012*, 1–4, doi:10.1155/2012/424526.
330. Norsigian, C.J.; Pusarla, N.; McConn, J.L.; Yurkovich, J.T.; Dräger, A.; Palsson, B.O.; King, Z. BiGG Models 2020: Multi-Strain Genome-Scale Models and Expansion across the Phylogenetic Tree. *Nucleic Acids Res* **2020**, *48*, D402–D406, doi:10.1093/nar/gkz1054.
331. Smid, E.J.; Enckevort, F.J.H.; Wegkamp, A.; Boekhorst, J.; Molenaar, D.; Hugenholtz, J.; Siezen, R.J.; Teusink, B. Metabolic Models for Rational Improvement of Lactic Acid Bacteria as Cell Factories. *Journal of Applied Microbiology* **2005**, *98*, 1326–1331, doi:10.1111/j.1365-2672.2005.02652.x.
332. Gänzle, M.G. Lactic Metabolism Revisited: Metabolism of Lactic Acid Bacteria in Food Fermentations and Food Spoilage. *Current Opinion in Food Science* **2015**, *2*, 106–117, doi:10.1016/j.cofs.2015.03.001.
333. Siezen, R.J.; van Enckevort, F.H.; Kleerebezem, M.; Teusink, B. Genome Data Mining of Lactic Acid Bacteria: The Impact of Bioinformatics. *Current Opinion in Biotechnology* **2004**, *15*, 105–115, doi:10.1016/j.copbio.2004.02.002.
334. Salvetti, E.; Torriani, S.; Felis, G.E. The Genus Lactobacillus: A Taxonomic Update. *Probiotics & Antimicro. Prot.* **2012**, *4*, 217–226, doi:10.1007/s12602-012-9117-8.
335. Ayed, L.; M'hir, S.; Nuzzolese, D.; Di Cagno, R.; Filannino, P. Harnessing the Health and Techno-Functional Potential of Lactic Acid Bacteria: A Comprehensive Review. *Foods* **2024**, *13*, 1538, doi:10.3390/foods13101538.
336. Coelho, M.C.; Malcata, F.X.; Silva, C.C.G. Lactic Acid Bacteria in Raw-Milk Cheeses: From Starter Cultures to Probiotic Functions. *Foods* **2022**, *11*, 2276, doi:10.3390/foods11152276.
337. Garcia-Gonzalez, N.; Bottacini, F.; van Sinderen, D.; Gahan, C.G.M.; Corsetti, A. Comparative Genomics of Lactiplantibacillus Plantarum: Insights Into Probiotic Markers in Strains Isolated From the Human Gastrointestinal Tract and Fermented Foods. *Front. Microbiol.* **2022**, *13*, 854266, doi:10.3389/fmicb.2022.854266.
338. Duar, R.M.; Lin, X.B.; Zheng, J.; Martino, M.E.; Grenier, T.; Pérez-Muñoz, M.E.; Leulier, F.; Gänzle, M.; Walter, J. Lifestyles in Transition: Evolution and Natural History of the Genus Lactobacillus. *FEMS Microbiology Reviews* **2017**, *41*, S27–S48, doi:10.1093/femsre/fux030.
339. Zheng, J.; Wittouck, S.; Salvetti, E.; Franz, C.M.A.P.; Harris, H.M.B.; Mattarelli, P.; O'Toole, P.W.; Pot, B.; Vandamme, P.; Walter, J.; et al. A Taxonomic Note on the Genus Lactobacillus: Description of 23 Novel Genera, Emended Description of the Genus Lactobacillus Beijerinck 1901, and Union of Lactobacillaceae and Leuconostocaceae. *International Journal of Systematic and Evolutionary Microbiology* **2020**, *70*, 2782–2858, doi:10.1099/ijsem.0.004107.
340. Mazzoli, R.; Bosco, F.; Mizrahi, I.; Bayer, E.A.; Pessione, E. Towards Lactic Acid Bacteria-Based Biorefineries. *Biotechnology Advances* **2014**, *32*, 1216–1236, doi:10.1016/j.biotechadv.2014.07.005.
341. Tanaka, K.; Komiyama, A.; Sonomoto, K.; Ishizaki, A.; Hall, S.; Stanbury, P. Two Different Pathways for D -Xylose Metabolism and the Effect of

Xylose Concentration on the Yield Coefficient of L -Lactate in Mixed-Acid Fermentation by the Lactic Acid Bacterium Lactococcus Lactis IO-1. *Applied Microbiology and Biotechnology* **2002**, *60*, 160–167, doi:10.1007/s00253-002-1078-5.
342. Kleerebezem, M.; Boekhorst, J.; van Kranenburg, R.; Molenaar, D.; Kuipers, O.P.; Leer, R.; Tarchini, R.; Peters, S.A.; Sandbrink, H.M.; Fiers, M.W.E.J.; et al. Complete Genome Sequence of Lactobacillus Plantarum WCFS1. *Proceedings of the National Academy of Sciences* **2003**, *100*, 1990–1995, doi:10.1073/pnas.0337704100.
343. Siezen, R.J.; Tzeneva, V.A.; Castioni, A.; Wels, M.; Phan, H.T.K.; Rademaker, J.L.W.; Starrenburg, M.J.C.; Kleerebezem, M.; Molenaar, D.; van Hylckama Vlieg, J.E.T. Phenotypic and Genomic Diversity of *Lactobacillus Plantarum* Strains Isolated from Various Environmental Niches. *Environmental Microbiology* **2010**, *12*, 758–773, doi:10.1111/j.1462-2920.2009.02119.x.
344. Goffin, P.; van de Bunt, B.; Giovane, M.; Leveau, J.H.J.; Höppener-Ogawa, S.; Teusink, B.; Hugenholtz, J. Understanding the Physiology of *Lactobacillus Plantarum* at Zero Growth. *Mol Syst Biol* **2010**, *6*, 413, doi:10.1038/msb.2010.67.
345. Tonsbeek, C.H.T.; Plancken, A.J.; Weerdhof, T.V.D. Components Contributing to Beef Flavor. Isolation of 4-Hydroxy-5-Methyl-3(2H)-Furanone and Its 2,5-Dimethyl Homolog from Beef Broth. *J. Agric. Food Chem.* **1968**, *16*, 1016–1021, doi:10.1021/jf60160a008.
346. Remus, D.M.; van Kranenburg, R.; van Swam, I.I.; Taverne, N.; Bongers, R.S.; Wels, M.; Wells, J.M.; Bron, P.A.; Kleerebezem, M. Impact of 4 Lactobacillus Plantarum Capsular Polysaccharide Clusters on Surface Glycan Composition and Host Cell Signaling. *Microb Cell Fact* **2012**, *11*, 149, doi:10.1186/1475-2859-11-149.
347. Jain, C.; Rodriguez-R, L.M.; Phillippy, A.M.; Konstantinidis, K.T.; Aluru, S. High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nat Commun* **2018**, *9*, 5114, doi:10.1038/s41467-018-07641-9.
348. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **2016**, *33*, 1635–1638, doi:10.1093/molbev/msw046.
349. Wang, Y.; Chen, C.; Ai, L.; Zhou, F.; Zhou, Z.; Wang, L.; Zhang, H.; Chen, W.; Guo, B. Complete Genome Sequence of the Probiotic *Lactobacillus Plantarum* ST-III. *J Bacteriol* **2011**, *193*, 313–314, doi:10.1128/JB.01159-10.
350. Jiang, Y.; Yang, Z. A Functional and Genetic Overview of Exopolysaccharides Produced by Lactobacillus Plantarum. *Journal of Functional Foods* **2018**, *47*, 229–240, doi:10.1016/j.jff.2018.05.060.
351. Tomita, S.; Irisawa, T.; Tanaka, N.; Nukada, T.; Satoh, E.; Uchimura, T.; Okada, S. Comparison of Components and Synthesis Genes of Cell Wall Teichoic Acid among *Lactobacillus Plantarum* Strains. *Bioscience, Biotechnology, and Biochemistry* **2010**, *74*, 928–933, doi:10.1271/bbb.90736.
352. Bron, P.A.; Tomita, S.; van Swam, I.I.; Remus, D.M.; Meijerink, M.; Wels, M.; Okada, S.; Wells, J.M.; Kleerebezem, M. Lactobacillus Plantarum Possesses the Capability for Wall Teichoic Acid Backbone Alditol

Switching. *Microb Cell Fact* **2012**, *11*, 123, doi:10.1186/1475-2859-11-123.
353. Fallico, V.; McAuliffe, O.; Fitzgerald, G.F.; Ross, R.P. Plasmids of Raw Milk Cheese Isolate Lactococcus Lactis Subsp. Lactis Biovar Diacetylactis DPC3901 Suggest a Plant-Based Origin for the Strain. *Appl Environ Microbiol* **2011**, *77*, 6451–6462, doi:10.1128/AEM.00661-11.
354. Parks, D.H.; Chuvochina, M.; Rinke, C.; Mussig, A.J.; Chaumeil, P.-A.; Hugenholtz, P. GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy. *Nucleic Acids Research* **2021**, doi:10.1093/nar/gkab776.
355. Chaumeil, P.-A.; Mussig, A.J.; Hugenholtz, P.; Parks, D.H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. *Bioinformatics* **2019**, btz848, doi:10.1093/bioinformatics/btz848.
356. Carpi, F.M.; Coman, M.M.; Silvi, S.; Picciolini, M.; Verdenelli, M.C.; Napolioni, V. Comprehensive Pan-genome Analysis of *Lactiplantibacillus Plantarum* Complete Genomes. *J of Applied Microbiology* **2022**, *132*, 592–604, doi:10.1111/jam.15199.
357. Li, K.; Wang, S.; Liu, W.; Kwok, L.-Y.; Bilige, M.; Zhang, W. Comparative Genomic Analysis of 455 Lactiplantibacillus Plantarum Isolates: Habitat-Specific Genomes Shaped by Frequent Recombination. *Food Microbiology* **2022**, *104*, 103989, doi:10.1016/j.fm.2022.103989.
358. Hayashi Sant'Anna, F.; Bach, E.; Porto, R.Z.; Guella, F.; Hayashi Sant'Anna, E.; Passaglia, L.M.P. Genomic Metrics Made Easy: What to Do and Where to Go in the New Era of Bacterial Taxonomy. *Critical Reviews in Microbiology* **2019**, *45*, 182–200, doi:10.1080/1040841X.2019.1569587.
359. Solioz, M.; Mermod, M.; Abicht, H.K.; Mancini, S. Responses of Lactic Acid Bacteria to Heavy Metal Stress. In *Stress Responses of Lactic Acid Bacteria*; Tsakalidou, E., Papadimitriou, K., Eds.; Springer US: Boston, MA, 2011; pp. 163–195 ISBN 978-0-387-92770-1.
360. Almeida, S.; Sousa, C.; Abreu, V.; Diniz, C.; Dorneles, E.M.S.; Lage, A.P.; Barh, D.; Azevedo, V. Exploration of Nitrate Reductase Metabolic Pathway in *Corynebacterium Pseudotuberculosis*. *International Journal of Genomics* **2017**, *2017*, 1–12, doi:10.1155/2017/9481756.
361. Brooijmans, R.J.W.; de Vos, W.M.; Hugenholtz, J. *Lactobacillus Plantarum* WCFS1 Electron Transport Chains. *Appl Environ Microbiol* **2009**, *75*, 3580–3585, doi:10.1128/AEM.00147-09.
362. Pepoyan, A.Z.; Manvelyan, A.M.; Balayan, M.H.; McCabe, G.; Tsaturyan, V.V.; Melnikov, V.G.; Chikindas, M.L.; Weeks, R.; Karlyshev, A.V. The Effectiveness of Potential Probiotics Lactobacillus Rhamnosus Vahe and Lactobacillus Delbrueckii IAHAHI in Irradiated Rats Depends on the Nutritional Stage of the Host. *Probiotics & Antimicro. Prot.* **2020**, *12*, 1439–1450, doi:10.1007/s12602-020-09662-7.
363. Patil, A.; Dubey, A.; Malla, M.A.; Disouza, J.; Pawar, S.; Alqarawi, A.A.; Hashem, A.; Abd_Allah, E.F.; Kumar, A. Complete Genome Sequence of Lactobacillus Plantarum Strain JDARSH, Isolated from Sheep Milk. *Microbiol Resour Announc* **2020**, *9*, e01199-19, doi:10.1128/MRA.01199-19.
364. Sewell, E.W.; Brown, E.D. Taking Aim at Wall Teichoic Acid Synthesis: New Biology and New Leads for Antibiotics. *J Antibiot* **2014**, *67*, 43–51,

doi:10.1038/ja.2013.100.
365. Wu, X.; Han, J.; Gong, G.; Koffas, M.A.G.; Zha, J. Wall Teichoic Acids: Physiology and Applications. *FEMS Microbiology Reviews* **2021**, *45*, fuaa064, doi:10.1093/femsre/fuaa064.
366. Zeidan, A.A.; Poulsen, V.K.; Janzen, T.; Buldo, P.; Derkx, P.M.F.; Øregaard, G.; Neves, A.R. Polysaccharide Production by Lactic Acid Bacteria: From Genes to Industrial Applications. *FEMS Microbiology Reviews* **2017**, *41*, S168–S200, doi:10.1093/femsre/fux017.
367. Garcia-Gonzalez, N.; Battista, N.; Prete, R.; Corsetti, A. Health-Promoting Role of Lactiplantibacillus Plantarum Isolated from Fermented Foods. *Microorganisms* **2021**, *9*, 349, doi:10.3390/microorganisms9020349.
368. Mukherjee, S.; Stamatis, D.; Li, C.T.; Ovchinnikova, G.; Bertsch, J.; Sundaramurthi, J.C.; Kandimalla, M.; Nicolopoulos, P.A.; Favognano, A.; Chen, I.-M.A.; et al. Twenty-Five Years of Genomes OnLine Database (GOLD): Data Updates and New Features in v.9. *Nucleic Acids Research* **2023**, *51*, D957–D963, doi:10.1093/nar/gkac974.
369. Koduru, L.; Lakshmanan, M.; Lee, Y.Q.; Ho, P.-L.; Lim, P.-Y.; Ler, W.X.; Ng, S.K.; Kim, D.; Park, D.-S.; Banu, M.; et al. Systematic Evaluation of Genome-Wide Metabolic Landscapes in Lactic Acid Bacteria Reveals Diet- and Strain-Specific Probiotic Idiosyncrasies. *Cell Reports* **2022**, *41*, 111735, doi:10.1016/j.celrep.2022.111735.
370. Choi, Y.-M.; Lee, Y.Q.; Song, H.-S.; Lee, D.-Y. Genome Scale Metabolic Models and Analysis for Evaluating Probiotic Potentials. *Biochemical Society Transactions* **2020**, *48*, 1309–1321, doi:10.1042/BST20190668.
371. Choi, S.; Jin, G.-D.; Park, J.; You, I.; Kim, E.B. Pan-Genomics of Lactobacillus Plantarum Revealed Group-Specific Genomic Profiles without Habitat Association. *Journal of Microbiology and Biotechnology* **2018**, *28*, 1352–1359, doi:10.4014/jmb.1803.03029.
372. Evanovich, E.; De Souza Mendonça Mattos, P.J.; Guerreiro, J.F. Comparative Genomic Analysis of *Lactobacillus Plantarum* : An Overview. *International Journal of Genomics* **2019**, *2019*, 1–11, doi:10.1155/2019/4973214.
373. Fidanza, M.; Panigrahi, P.; Kollmann, T.R. Lactiplantibacillus Plantarum–Nomad and Ideal Probiotic. *Front. Microbiol.* **2021**, *12*, 712236, doi:10.3389/fmicb.2021.712236.
374. Martino, M.E.; Bayjanov, J.R.; Caffrey, B.E.; Wels, M.; Joncour, P.; Hughes, S.; Gillet, B.; Kleerebezem, M.; Van Hijum, S.A.F.T.; Leulier, F. Nomadic Lifestyle of *Lactobacillus Plantarum* Revealed by Comparative Genomics of 54 Strains Isolated from Different Habitats: The Genomes of 54 *Lactobacillus Plantarum* Strains Reveal Their Nomadic Lifestyle. *Environmental Microbiology* **2016**, *18*, 4974–4989, doi:10.1111/1462-2920.13455.
375. Reid, S.J.; Abratt, V.R. Sucrose Utilisation in Bacteria: Genetic Organisation and Regulation. *Appl Microbiol Biotechnol* **2005**, *67*, 312–321, doi:10.1007/s00253-004-1885-y.
376. Siezen, R.J.; Francke, C.; Renckens, B.; Boekhorst, J.; Wels, M.; Kleerebezem, M.; van Hijum, S.A.F.T. Complete Resequencing and Reannotation of the Lactobacillus Plantarum WCFS1 Genome. *Journal of*

Bacteriology **2012**, *194*, 195–196, doi:10.1128/JB.06275-11.
377. Teran, M.D.M.; De Moreno De LeBlanc, A.; Savoy De Giori, G.; LeBlanc, J.G. Thiamine-Producing Lactic Acid Bacteria and Their Potential Use in the Prevention of Neurodegenerative Diseases. *Appl Microbiol Biotechnol* **2021**, *105*, 2097–2107, doi:10.1007/s00253-021-11148-7.
378. Cui, Y.; Wang, M.; Zheng, Y.; Miao, K.; Qu, X. The Carbohydrate Metabolism of Lactiplantibacillus Plantarum. *IJMS* **2021**, *22*, 13452, doi:10.3390/ijms222413452.
379. Kanehisa, M. Toward Understanding the Origin and Evolution of Cellular Organisms. *Protein Science* **2019**, *28*, 1947–1951, doi:10.1002/pro.3715.
380. Koller, M.; Maršálek, L.; De Sousa Dias, M.M.; Braunegg, G. Producing Microbial Polyhydroxyalkanoate (PHA) Biopolyesters in a Sustainable Manner. *New Biotechnology* **2017**, *37*, 24–38, doi:10.1016/j.nbt.2016.05.001.
381. Khanna, S.; Srivastava, A.K. Recent Advances in Microbial Polyhydroxyalkanoates. *Process Biochemistry* **2005**, *40*, 607–619, doi:10.1016/j.procbio.2004.01.053.
382. Albuquerque, M.G.E.; Torres, C.A.V.; Reis, M.A.M. Polyhydroxyalkanoate (PHA) Production by a Mixed Microbial Culture Using Sugar Molasses: Effect of the Influent Substrate Concentration on Culture Selection. *Water Research* **2010**, *44*, 3419–3433, doi:10.1016/j.watres.2010.03.021.
383. Strazzera, G.; Battista, F.; Andreolli, M.; Menini, M.; Bolzonella, D.; Lampis, S. Influence of Different Household Food Wastes Fractions on Volatile Fatty Acids Production by Anaerobic Fermentation. *Bioresource Technology* **2021**, *335*, 125289, doi:10.1016/j.biortech.2021.125289.
384. Albuquerque, M.G.E.; Carvalho, G.; Kragelund, C.; Silva, A.F.; Barreto Crespo, M.T.; Reis, M.A.M.; Nielsen, P.H. Link between Microbial Composition and Carbon Substrate-Uptake Preferences in a PHA-Storing Community. *The ISME Journal* **2013**, *7*, 1–12, doi:10.1038/ismej.2012.74.
385. Frison, N.; Andreolli, M.; Botturi, A.; Lampis, S.; Fatone, F. Effects of the Sludge Retention Time and Carbon Source on Polyhydroxyalkanoate-Storing Biomass Selection under Aerobic-Feast and Anoxic-Famine Conditions. *ACS Sustainable Chem. Eng.* **2021**, *9*, 9455–9464, doi:10.1021/acssuschemeng.1c02973.
386. O'Leary, N.A.; Cox, E.; Holmes, J.B.; Anderson, W.R.; Falk, R.; Hem, V.; Tsuchiya, M.T.N.; Schuler, G.D.; Zhang, X.; Torcivia, J.; et al. Exploring and Retrieving Sequence and Metadata for Species across the Tree of Life with NCBI Datasets. *Sci Data* **2024**, *11*, 732, doi:10.1038/s41597-024-03571-y.
387. Mishra, N.K.; Chang, J.; Zhao, P.X. Prediction of Membrane Transport Proteins and Their Substrate Specificities Using Primary Sequence Information. *PLoS ONE* **2014**, *9*, e100278, doi:10.1371/journal.pone.0100278.
388. Ren, Q.; Paulsen, I.T. Comparative Analyses of Fundamental Differences in Membrane Transport Capabilities in Prokaryotes and Eukaryotes. *PLoS Comput Biol* **2005**, *1*, e27, doi:10.1371/journal.pcbi.0010027.
389. Zallot, R.; Harrison, K.; Kolaczkowski, B.; De Crécy-Lagard, V. Functional Annotations of Paralogs: A Blessing and a Curse. *Life* **2016**, *6*, 39, doi:10.3390/life6030039.
390. Kuriya, Y.; Araki, M. Dynamic Flux Balance Analysis to Evaluate the Strain

Production Performance on Shikimic Acid Production in Escherichia Coli. *Metabolites* **2020**, *10*, 198, doi:10.3390/metabo10050198.
391. Bachmann, H.; Molenaar, D.; Branco dos Santos, F.; Teusink, B. Experimental Evolution and the Adjustment of Metabolic Strategies in Lactic Acid Bacteria. *FEMS Microbiology Reviews* **2017**, *41*, S201–S219, doi:10.1093/femsre/fux024.
392. Schroeder, W.L.; Suthers, P.F.; Willis, T.C.; Mooney, E.J.; Maranas, C.D. Current State, Challenges, and Opportunities in Genome-Scale Resource Allocation Models: A Mathematical Perspective. *Metabolites* **2024**, *14*, 365, doi:10.3390/metabo14070365.
393. Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M.K.M.; Kerkhoven, E.J.; Nielsen, J. Deep Learning-Based Kcat Prediction Enables Improved Enzyme-Constrained Model Reconstruction. *Nat Catal* **2022**, *5*, 662–672, doi:10.1038/s41929-022-00798-z.

Ringraziamenti 🇮🇹

Ho scelto di fare questo dottorato nonostante le condizioni della ricerca universitaria in Italia siano “un po’ particolari”. Mosso da sincera passione, valeva la pena provarci: meglio l’eventualità di un fallimento che la certezza di un rimpianto! In ogni caso… studiare i modelli metabolici è stato davvero entusiasmante.

Sono stati tre anni al computer, ma non per questo solitari: sono tante le persone con cui ho fatto squadra e che desidero ringraziare. Comincio dalla Prof. Giovanna Felis, per il suo supporto decisivo, l’entusiasmo contagioso e le discussioni stimolanti. Grazie anche a Prof. Silvia Lampis, Dott. Mehrdad Jaberi, Prof. Elisa Salvetti e Dott. Ilaria Checchia, per i loro importanti contributi nei progetti condivisi. Un super grazie va sicuramente al Prof. Bas Teusink, per l’opportunità di trascorrere tre mesi al Systems Biology Lab della Vrije Universiteit. E grazie anche a Pranas, Daniëlle, Francesco, Tania, e a tutti gli altri ragazzi conosciuti ad Amsterdam, per aver reso il mio periodo all’estero un piacere. Estendo i ringraziamenti ai compagni di laboratorio, per aver contribuito ad un’atmosfera positiva. Ma, soprattutto, la mia gratitudine va al Prof. Nicola Vitulo: mi ha dato fiducia fin dall’inizio, mi ha sostenuto in questo percorso ad ostacoli, e si è dimostrato una persona di valore oltre che un professore competente: sono fiero di aver svolto il dottorato nel suo laboratorio.

Certo, grazie anche alla mia famiglia e alle classiche domande “Quando ti trovi un lavoro? Quando ti stabilizzi? Ma non ti stai trascurando?”. A parte gli scherzi, sento tutto il vostro affetto. Grazie agli amici, vicini e distanti, storici e nuovi, per avermi alleggerito. Tanta altra vita ci attende.

Valentina… credo non esistano parole adatte. Per quello che vale, questo dottorato è dedicato a te.

Gioele Lazzari
Peseggia, 6 dicembre 2024

Acknowledgements

I chose to pursue this PhD despite the fact that the conditions of university research in Italy are “a bit peculiar”. Driven by genuine passion, it was worth trying: better the possibility of failure than the certainty of regret! In any case… studying metabolic models has been truly exciting.

It has been three years at the computer, but not of solitude: there have been many people with whom I’ve worked closely, and to whom I want to express my gratitude. I begin with Prof. Giovanna Felis, for her crucial support, contagious enthusiasm, and stimulating discussions. Thanks also to Prof. Silvia Lampis, Dr. Mehrdad Jaberi, Prof. Elisa Salvetti, and Dr. Ilaria Checchia, for their important contributions to the shared projects. A huge thank you goes to Prof. Bas Teusink, for the opportunity to spend three months at the Systems Biology Lab of the Vrije Universiteit. And thanks also to Pranas, Daniëlle, Francesco, Tania, and all the other people I met in Amsterdam, for making my time abroad a pleasure. I extend my thanks to my lab mates for contributing to a positive atmosphere. But, above all, my gratitude goes to Prof. Nicola Vitulo: he trusted me from the very beginning, supported me throughout this obstacle course, and proved to be a person of great value besides being a competent professor: I am proud to have pursued my PhD in his lab.

Of course, thanks also to my family and the usual questions: “When will you find a job? When will you settle down? Aren’t you neglecting yourself?”. Jokes aside, I feel all your love. Thanks to my friends, near and far, old and new, for lightening my load. Much more life awaits us.

Valentina… I believe there are no suitable words. For whatever it’s worth, this PhD is dedicated to you.

Gioele Lazzari
Peseggia, 6 dicembre 2024