Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Elena K. Matveishina

born in: Moscow, Russia

Oral examination: 17-12-2025

# Consensus approach for assessing and resolving uncertainty in genome-scale metabolic models: advancing systems-level understanding of microbial metabolism

Referees: Prof. Dr. Ursula Kummer

Dr. Mikhail Savitski

# Abstract

Genome-scale metabolic models (GEMs) are an important methodology in systems biology: they play a major role in investigating microbial metabolism and predicting responses to perturbations by representing bacterial metabolism as a whole system. GEMs can be automatically reconstructed from bacterial genomes with computational tools that employ distinct methodological approaches. My initial analyses of GEMs reconstructed by different tools for a small set of diverse gut microbial species revealed that automated reconstruction pipelines often produce GEMs with different structures and predictive behaviour, even for the same organism. Because individual models can excel at different tasks and capture distinct metabolic capabilities, I hypothesised that combining them can increase confidence in network content and improve performance. In this thesis, I present a consensus approach implemented in GEMsembler, a Python package for cross-tool model comparison and assembly of consensus models from any subset of input GEMs. Alongside the consensus strategy, GEMsembler offers broad analysis functionality, including detection and visualisation of biosynthetic pathways, growth evaluation, and an agreement-driven curation workflow. In a use-case study, consensus models curated with this workflow, combining four automatically reconstructed GEMs for *Lactiplantibacillus plantarum* and *Escherichia coli*, outperform gold-standard models on auxotrophy and gene-essentiality benchmarks. Moreover, enabled by GEMs comparison, optimising gene–protein–reaction (GPR) rule combinations derived from input and consensus models improves gene-essentiality predictions even for manually curated gold-standard models. GEMsembler also helps explain model performance by highlighting relevant metabolic pathways and, together with the consensus principle, supports the assessment of network uncertainty and informs the design of targeted experiments to resolve it. Finally, I apply the consensus approach to de novo reconstruction of two of the most abundant human gut bacteria, *Bacteroides uniformis* and *Phocaeicola vulgatus*, yielding first-iteration curated models that reproduce their growth and major metabolic phenotypes. In agreement with experimental data, the *B. uniformis* model secretes more lactate and malate but grows less, whereas the *P. vulgatus* model secretes less of these metabolites, grows more, and also secretes succinate. Together, these results show that the consensus approach facilitates building metabolic models that are more accurate, concise, and biologically informed, advancing systems-level understanding of microbial metabolism.

# Zusammenfassung

Genome-scale metabolic models (GEMs) sind eine wichtige Methodik in der Systembiologie: sie spielen eine zentrale Rolle bei der Untersuchung des mikrobiellen Metabolismus und der Vorhersage von Reaktionen auf Störungen, indem sie den bakteriellen Stoffwechsel als ein gesamtes System darstellen. GEMs können automatisch aus bakteriellen Genomen mit rechnergestützten Werkzeugen rekonstruiert werden, die unterschiedliche methodische Ansätze verwenden. Meine ersten Analysen von GEMs, die mit verschiedenen Tools für eine kleine Auswahl unterschiedlicher Darmmikrobenarten rekonstruiert wurden, zeigten, dass automatisierte Rekonstruktionspipelines oft GEMs mit unterschiedlichen Strukturen und unterschiedlichem Vorhersageverhalten erzeugen, selbst für denselben Organismus. Da einzelne Modelle in verschiedenen Aufgabenbereichen besonders leistungsfähig sein und unterschiedliche metabolische Fähigkeiten abbilden können, stellte ich die Hypothese auf, dass ihre Kombination das Vertrauen in den Netzwerkinhalt erhöhen und die Leistungsfähigkeit verbessern kann. In dieser Arbeit präsentiere ich einen Konsensusansatz implementiert in GEMsembler, einem Python-Paket für werkzeugübergreifenden Modellvergleich und die Erstellung von Konsensusmodellen aus beliebigen Teilmengen von Eingabe-GEMs. Neben der Konsensusstrategie bietet GEMsembler umfangreiche Analysefunktionen, einschließlich der Erkennung und Visualisierung von Biosynthesewegen, Wachstumsauswertung und einem auf Übereinstimmung basierenden Kurations-Workflow. In einer Fallstudie zeigen Konsensusmodelle, die mit diesem Workflow kuratiert wurden, und vier automatisch rekonstruierte GEMs für *Lactiplantibacillus plantarum* und *Escherichia coli* kombinieren, bessere Ergebnisse als Goldstandardmodelle in Auxotrophie- und Genessenzialitäts-Benchmarks. Darüber hinaus ermöglicht der Vergleich von GEMs die Optimierung von Gen-Protein-Reaktions-(GPR)-Regelkombinationen, die aus Eingabe- und Konsensusmodellen abgeleitet werden, wodurch die Vorhersage der Genessenzialität selbst für manuell kuratierte Goldstandardmodelle verbessert wird. GEMsembler trägt zudem zum Verständnis der Modellleistung bei, indem es relevante Stoffwechselwege hervorhebt, und unterstützt zusammen mit dem Konsensusprinzip die Bewertung von Netzwerkunsicherheiten sowie die Planung gezielter Experimente zu deren Auflösung. Schließlich wende ich den Konsensusansatz auf die de novo-Rekonstruktion von zwei

der häufigsten menschlichen Darmbakterien, *Bacteroides uniformis* und *Phocaeicola vulgatus*, an, was zu erstkuratierten Modellen führt, die deren Wachstum und zentrale metabolische Phänotypen reproduzieren. In Übereinstimmung mit experimentellen Daten sezerniert das *B. uniformis*-Modell mehr Laktat und Malat, wächst jedoch weniger, während das *P. vulgatus*-Modell geringere Mengen dieser Metabolite produziert, stärker wächst und zusätzlich Succinat freisetzt. Zusammen zeigen diese Ergebnisse, dass der Konsensusansatz die Erstellung metabolischer Modelle erleichtert, die präziser, kompakter und biologisch aussagekräftiger sind, und damit das systemische Verständnis des mikrobiellen Metabolismus vorantreibt.

# Contents

# Introduction

# Genome-scale metabolic models (GEMs) as a versatile tool to study microbial metabolism

## 1 Importance of human gut bacteria and their metabolism

Human gut harbors trillions of bacteria from more than fifty bacterial phyla [1, 2]. Among them, *Firmicutes* and *Bacteroidetes* dominate, followed by *Proteobacteria*, *Verrucomicrobia*, and others [3]. These gut bacteria contribute to numerous metabolic functions, shape host physiology, and play a crucial role in human health and diseases [2, 4, 5].

Human gut bacteria function as a "metabolic organ" that complements human physiology by digesting otherwise indigestible dietary substrates, synthesising micronutrients, and chemically modifying host compounds, participating in the metabolism of carbohydrates, proteins, bile acids, vitamins, and other compounds [2, 6]. Initial degradation of insoluble carbohydrates can be performed by colonic microbes such as *Bacteroides*, *Ruminococcus*, and others [2, 6, 7]. This fermentation results in the production of short-chain fatty acids (SCFAs) such as acetate, propionate, and butyrate [2, 6]. Gut bacteria also supply essential nutrients to the host. For example, certain commensals (e.g. Bifidobacterium, Lactobacillus) can *de novo* synthesise B-vitamins, including folate, riboflavin, and vitamin B12, which cannot be synthesised by humans [2, 8]. In addition, the microbiota participates in host lipid metabolism by transforming primary bile acids (made in the liver) into secondary bile acids that play key roles in fat digestion and act as hormone-like signalling molecules in the gut [2, 6].

Beyond metabolic digestion, commensal bacteria have coevolved for mutually beneficial symbiosis with humans and profoundly shape host physiology and immune development [4, 5]. Microbial metabolites serve as messengers in host endocrine and

metabolic pathways. A prime example is SCFAs, which bind to G protein-coupled receptors on host cells to modulate gut barrier function, appetite regulation, glucose homeostasis, and adiposity [4–6]. Gut microbes also help maintain the integrity of the intestinal barrier and calibrate the host's immune responses. For example, neonatal gut microbiome dysbiosis can promote CD4+ T cell dysfunction associated with childhood atopy and asthma [9]. Another example is the anti-inflammatory activity of *B. thetaiotaomicron*, which regulates nuclear–cytoplasmic shuttling of transcription factors involved in inflammatory signalling [10]. In these ways, the gut microbiota plays an integral role in shaping normal physiology, ranging from immune equilibrium to energy balance.

An imbalance in the gut microbiome (dysbiosis) has been implicated in a wide range of diseases, most notably metabolic and inflammatory disorders [4, 5]. In obesity and type 2 diabetes, characteristic shifts in microbial composition, such as an increased *Firmicutes*-to-*Bacteroidetes* ratio, enhance the capacity to degrade polysaccharides and generate fermentative products like short-chain fatty acids (SCFAs), thereby promoting energy harvest and influencing host lipid metabolism [5, 11, 12]. In inflammatory bowel disease (IBD), dysbiosis manifests as reduced microbial diversity and depletion of beneficial SCFA-producing species, such as *Faecalibacterium prausnitzii*, which normally support barrier integrity and exert anti-inflammatory effects [13, 14]. The consequent reduction in butyrate availability is thought to impair mucosal health and worsen inflammation, consistent with the therapeutic benefits of butyrate supplementation or butyrate-producing probiotics [14, 15].

Microbial metabolism can even influence diseases beyond the gastrointestinal tract. A notable example is cardiovascular disease: certain gut bacteria metabolize dietary choline (from foods like red meat or eggs) into trimethylamine (TMA), which is absorbed and oxidized by the host liver into trimethylamine-N-oxide (TMAO) – a metabolite strongly linked to atherosclerosis risk [5, 16]. Elevated TMAO levels contribute to plaque formation in arteries, suggesting that TMA-producing gut microbes (e.g. within the *Clostridial* and *Proteobacteria* l groups) can affect host cardiovascular health via their metabolites [5, 16].

Furthermore, microbial metabolites have been implicated in cancer development. For example, bacteria in the colon convert primary bile acids into secondary bile acids; these microbial-derived bile acids can cause DNA damage and promote oxidative stress in colon epithelial cells [17]. Such effects may increase the risk of colorectal cancer, especially when certain bile-tolerant bacteria (like *Clostridia* species with 7α-dehydroxylation activity) are overrepresented in the gut [17, 18]. On the other hand, some fermentation byproducts (like butyrate) can induce protective anti-cancer mechanisms, illustrating the dualistic influence of microbial metabolism on tumorigenesis [17, 19].

In summary, the human gut microbiota profoundly influences host nutrition and physiology through its metabolic activities, with bacterial metabolites and biochemical signals tightly integrated into host metabolic pathways.

# 2 Studying bacterial metabolism: from metabolites to metabolic networks

Given the pivotal role of the human gut microbiota, the number of studies devoted to human-associated bacteria has drastically increased over the past two decades. Pioneering studies such as the Human Microbiome Project (HMP) [20] and MetaHIT (METAgenomics of the Human Intestinal Tract) [21] paved the way to acquiring wide-scale genomic data and assessing microbial diversity in human organisms.

Since one key way in which microbes interact with the host is through the production and consumption of metabolites, microbial studies aim not only to characterise bacterial diversity, but also to assess their metabolic functions by measuring these metabolites in metabolomics assays [4]. Metabolites can be measured with liquid chromatography–mass spectrometry (LC–MS) or gas chromatography–mass spectrometry (GC–MS) platforms, enabling both untargeted and targeted analyses that provide semi-quantitative or even quantitative information on metabolite concentrations [4, 22, 23].

Although numerous studies have examined associations between metabolite levels and various conditions, concentrations alone do not reveal the underlying metabolic processes occurring in biological systems, whereas metabolic fluxes and corresponding reaction rates can serve as more informative biomarkers [24, 25]. However, measuring these reaction rates remains a very challenging task. Common experimental approaches to gain insights into metabolite fluxes rely on $^{13}$C-labelling techniques and tracking the propagation of the label through metabolic pathways [25, 26]. Nevertheless, these methods are applicable only to a very limited set of metabolites and pathways, primarily focusing on central carbon metabolism.

Overall, experimental measurements of metabolite concentrations and fluxes provide valuable yet incomplete insights into microbial metabolism. To overcome the limitations of experimental techniques, various computational modelling approaches have been developed to investigate the mechanisms underlying metabolic changes [26]. Pharmacokinetic modelling enables the description of the dynamics of selected metabolites [27], whereas approaches such as genome-scale metabolic models (GEMs) and flux balance analysis (FBA) are employed to capture the complexity of whole-system metabolism at the organismal or community level [28].

# 3 What are genome-scale metabolic models and flux balance analysis?

Genome-scale metabolic models (GEMs) represent the comprehensive metabolic network of an organism, connecting metabolites through biochemical reactions catalysed by enzyme-encoded genes, and serve as a framework to estimate reaction fluxes (v) across the system using flux balance analysis (FBA) [13]. In these models, metabolites participate in reactions with coefficients that reflect the stoichiometry of each reaction, which enables the formulation of a stoichiometric matrix (S) where rows correspond to metabolites, columns to reactions, and entries to stoichiometric coefficients. The steady-state assumption ensures that, for each metabolite, the total amount of production equals the total amount of consumption, resulting in the condition $S \cdot v = 0$. However, this equation alone is insufficient to calculate unique reaction rates, since the number of reactions (variables) typically exceeds the number of metabolites (equations), leading to an underdetermined system. FBA addresses this challenge by introducing an objective function that is optimised through linear programming, most often it means maximising the flux through a selected reaction, thereby narrowing the solution space [28].

For GEMs representing single bacteria, the most common objective function is an artificial biomass reaction, reflecting the principle that bacterial growth can be modelled by maximising the yield of biomass [28]. The biomass reaction comprises essential cellular components required for growth, such as nucleotides, amino acids, various lipids, vitamins, and cofactors, together with an energy requirement in the form of adenosine triphosphate (ATP) [29]. The growth-associated ATP maintenance (GAM) is implemented as ATP hydrolysis within the lumped biomass reaction, while the non-growth-associated ATP maintenance (NGAM) is incorporated as a separate ATP hydrolysis reaction [29]. Under certain conditions, alternative objective functions, such as maximising overall ATP production or ATP yield per unit of flux, may be more appropriate [30]. In addition, optimisation of artificial demand reactions for specific metabolites of interest can be used to test whether these metabolites can be produced and at what rate.

In addition to the choice of objective function, the definition of environmental and physiological conditions plays a crucial role in FBA-based simulations [28, 31]. These conditions are implemented as lower and upper bounds on specific reactions, thereby defining their directionality through the sign of the flux and constraining the magnitude of flux values. In GEMs, metabolites are distributed and reactions take place in distinct cellular compartments, which in bacteria are typically defined as cytosolic, periplasmic, and extracellular. Transport reactions connect these compartments by enabling the transfer of metabolites between them. Therefore different growth media and nutrient availability can be

simulated by adjusting the bounds of artificial exchange reactions, introducing metabolites into the extracellular compartment, which represent the uptake and secretion of metabolites permitted in the model [28]. With various constraints narrowing the feasible solution space, GEMs and their simulations can also be referred to as constraint-based modelling approaches. [32]

The connection between the bacterial genome and the reactions in the GEM is implemented via gene-protein-reaction (GPR) associations, also called gene-reaction rules [29, 33]. These rules describe the logical relationships between genes encoding enzyme subunits or isoenzymes and the corresponding metabolic reactions. In GEMs, GPRs are formulated as Boolean expressions, where an 'AND' operator links genes that encode different subunits of the same enzyme complex (all of which are required for the reaction to occur), while an 'OR' operator represents isoenzymes, where the presence of any one gene is sufficient to catalyse the reaction [29]. This formalism allows systematic translation of gene presence/absence, mutations, or differential expression into constraints on reaction activity. For example, gene knock-outs are simulated by blocking (or deleting) the associated reaction(s), whereas reduced enzyme activity can be modelled by imposing tighter bounds on the corresponding flux [34]. In this way, GPR rules provide a crucial link between genotype and metabolic phenotype.

For a more nuanced analysis of metabolic flexibility and robustness under different conditions, GEMs can be used not only with standard FBA but also with several extended methodologies [35]. For example, flux variability analysis (FVA) [36] quantifies the possible range of each reaction flux by maximising and minimising individual fluxes while constraining the original objective function to a defined fraction of its optimal value obtained from the initial FBA. Flux sampling [37], instead, generates random solutions within the feasible flux space, thereby producing a distribution of possible flux values for each reaction rather than a single optimal value. Parsimonious FBA (pFBA) [38] introduces the principle of minimal enzyme usage by selecting among alternative flux distributions those that minimise the total flux while still achieving the optimal objective. Furthermore, optimising for the production of a specific metabolite with pFBA can reveal a set of reactions that approximate a biosynthetic pathway. Together, these and other extensions provide complementary perspectives on network capabilities and allow a more detailed exploration of the feasible flux space.

In practice, GEMs are most commonly distributed in SBML (Systems Biology Markup Language) format, [39], which provides a standardised way to represent models and ensures their compatibility across different software platforms. These models can then be manipulated and simulated using the MATLAB COBRA Toolbox [32] or the Python implementation COBRApy [40]. There are also various extensions and packages that

enable advanced analyses, simulations and data integration, such as CellNetAnalyzer [41], MetaboTools [42], and many others [43].

# 4 Genome-scale metabolic models (GEMs) in system biology

Overall, with all these GEMs and FBA methodologies, systems biology has gained a rigorous framework for bridging experimental data with mechanistic models of metabolism, iteratively generating hypotheses with genome-scale metabolic reconstructions and validating them with multi-omic datasets [31] (Fig. 1A). With this approach, various experimental observations can be converted into mathematical constraints on the reaction network, resulting in cell-type- or condition-specific GEMs (Fig. 1C).

The most direct experimental input comes from growth rate measurements, as GEMs typically use a biomass objective function to represent cellular proliferation (Fig. 1B). The closest experimental set-up to the steady-state assumption of GEMs is growing bacteria in chemostat cultures, where the growth rate equals the dilution rate [44]. In practice, however, growth rates are more commonly estimated from growth curves obtained by monitoring the optical density (OD) of a microbial batch culture over time, with the exponential phase of these curves providing a growth rate ($h^{-1}$) estimate [45]. The experimentally determined growth rate can then be directly compared with the biomass flux predicted by the model, enabling both validation and calibration of the GEM. In this way, growth rate measurements serve as a crucial benchmark, ensuring that computational predictions are anchored in the physiological reality of the system under study [29, 31].

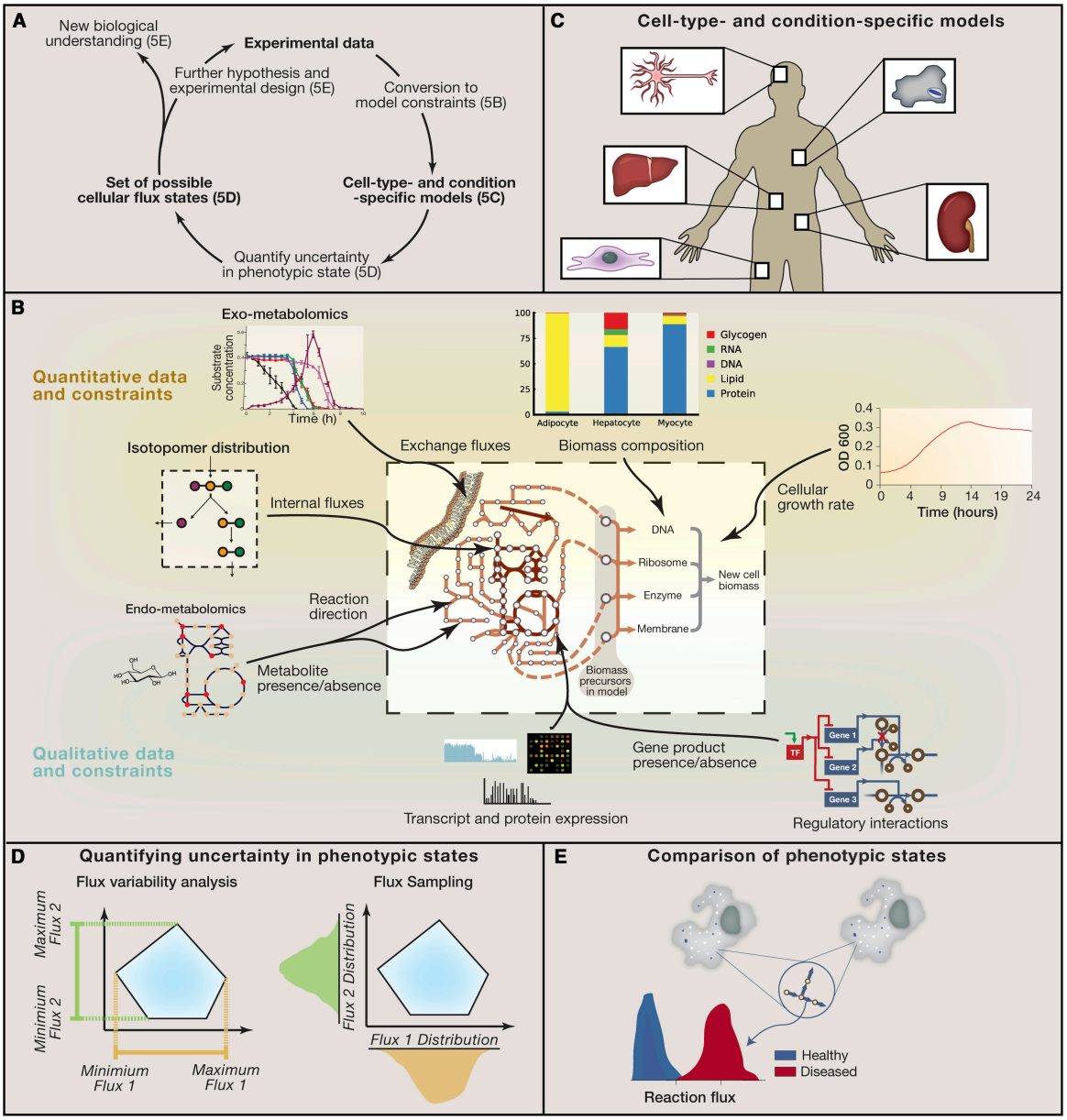


**Figure 1.** Iterative usage of genome-scale metabolic models (GEMs) for generating biological insights with multi-omic data integration (from Edward J O'Brien et. al., Cell 2015 [31]). A - Overview of the GEM usage workflow with stages corresponding to other panels. B - Integration of various experimental data as model constraints. C - Derivation of condition-specific models from integrated data. D - GEM simulation and flux quantification. E - Hypothesis generation through comparison of fluxes under different conditions.

Time-course profiles of extracellular metabolites quantify uptake and secretion and thereby define the nutritional environment for a GEM by constraining its exchange reactions [31] (Fig. 1B). In typical workflows, bacterial supernatant is sampled at multiple time points during exponential growth, and concentrations of metabolites of interest are measured either by targeted mass spectrometry or by other appropriate assays, for example enzymatic assay for determining glucose concentration [22,46]. These measurements allow the computation of uptake or secretion rates, which, after accounting for the culture volume (typically 1 L

by convention), are expressed as mmol $h^{-1}$. To compare these empirical rates with model variables, units must match the convention used for exchange reactions in GEMs, which are defined per unit biomass. Dividing the empirical rates by the measured biomass, converted from optical density (OD) to grams of dry weight (gDW), results in the rates in mmol $gDW^{-1}$ $h^{-1}$. These resulting uptake and secretion rates can then be imposed as lower or upper bounds on the corresponding exchange reactions, tightening the feasible flux space and grounding simulations in the assayed nutritional condition [45, 47].

Intracellular fluxes derived from $^{13}C$ tracer experiments can constrain specific internal reactions or pathway branch points, complementing uptake and secretion measurements (Fig. 1B). They resolve otherwise underdetermined intracellular routes and thereby narrow the solution space, particularly within central carbon metabolism, where such techniques are most effective [31].

Transcriptomic and proteomic data provide condition-specific evidence for reaction activity and capacity, and can be integrated into GEMs to tailor networks to the measured state by constraining reactions via gene–protein–reaction (GPR) rules (Fig. 1B). In typical workflows, mRNA abundance is quantified by RNA-seq or microarrays and reported as normalised counts, while protein abundance is obtained using shotgun or targeted mass spectrometry [48–50]. These measurements are then converted into model constraints either by thresholding to define reactions as active or inactive, or by applying enzyme kinetics to scale reaction bounds in proportion to expression levels [51–53]. Compared with exchange fluxes or $^{13}C$-derived internal fluxes, expression data are indirect and do not uniquely determine fluxes due to various regulatory mechanisms. Nevertheless, they are valuable for constructing condition-specific subnetworks and for further narrowing the feasible solution space.

Beyond simply constraining models with experimental data, GEMs are employed in a comparative framework (Fig. 1E). By constructing condition- or cell-type-specific models and exploring the feasible flux spaces with flux variability analysis (FVA) or flux sampling (Fig. 1D), it becomes possible to contrast metabolic states across various conditions. These comparisons can be made, for instance, between healthy and diseased cells, between pathogenic and mutualistic bacteria, or under distinct environmental conditions. Differences in the predicted metabolic capabilities highlight molecular mechanisms that distinguish the states and naturally generate biological hypotheses about altered pathways. These hypotheses can then be subjected to targeted experimental validation, thereby closing the loop between computation and experimental work.

In this way, data integration with GEMs is not a one-time procedure but rather an iterative discovery cycle: models are progressively refined as new data resolve uncertainties, while simultaneously guiding the design of new experiments that probe the most uncertain

or divergent predictions (Fig. 1). This feedback loop ensures that GEMs function both as structured repositories of current knowledge and as engines of discovery, continuously improving their predictive scope while driving biological insight.

# 5 Reconstruction of genome-scale metabolic models (GEMs)

The accuracy and predictive power of GEM simulations strongly depend on model quality, making reconstruction crucial for their further use and development. All reconstructions are initially based on the genome of the organism of interest. Performing genome annotation provides information on enzymatic functions, including the substrates and products of each enzyme, their stoichiometric coefficients, the reversibility of corresponding reactions, and the cellular compartments in which these reactions occur [54]. These elements are then integrated into a single, connected metabolic network that forms the foundation of the GEM.

Nowadays, a variety of computational tools exist for GEM reconstruction, many of which can automate parts or even the entire workflow (Fig. 2), thereby streamlining the process [55]. A central step relies on homology searches, in which the gene or protein sequences of the organism of interest are aligned with previously annotated enzymes of known function in organisms available in a reference database. These functional annotations are then mapped onto metabolic reactions using biochemical databases, for example, BiGG [33], KEGG [56], MetaCyc [57], or ModelSEED [58], resulting in a draft network of metabolites and reactions together with their corresponding gene–protein–reaction (GPR) associations. Reactions that are required for the synthesis of biomass components under a given or predicted growth medium but lack genetic evidence (i.e., no enzyme assigned) can be identified and incorporated through gap-filling procedure (Fig. 2). Reconstruction pipelines that integrate all of these steps automatically are capable of generating GEMs ready for simulation, typically in SBML format, directly from genome or proteome FASTA sequences.

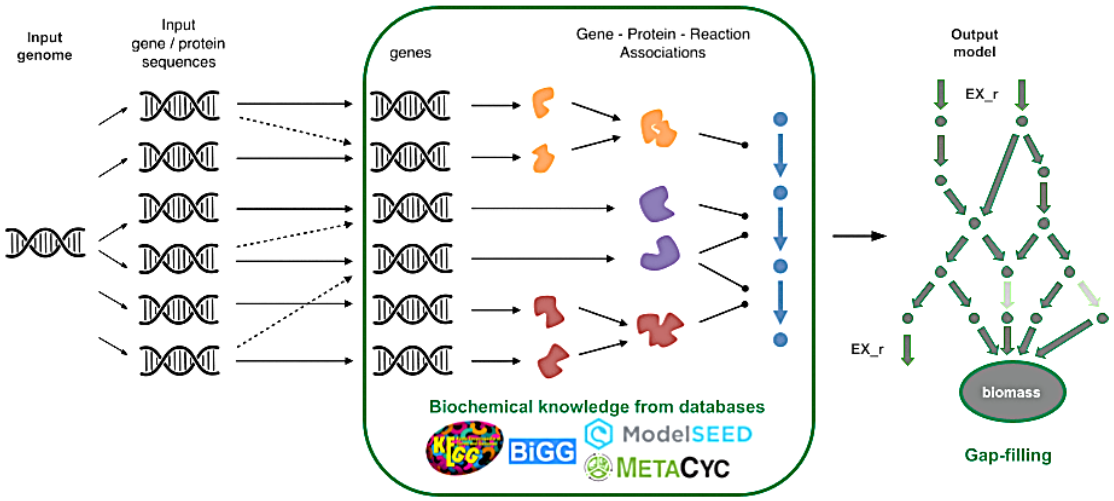


**Figure 2.** Automatic genome-scale metabolic model (GEM) reconstruction (adapted from Machado et. al., Nucleic Acid Research 2018 [59])

Different tools for automatic GEM reconstruction utilise distinct approaches, algorithms, and databases for these steps, and none of them invariably outperforms the others [55]. For example, merlin [60] provides a semi-automated framework that integrates genome annotation and database mapping with user-guided curation. Its main strength lies in the high degree of control it offers, allowing variation in the annotation algorithms, refinement of gene–protein–reaction associations and adjustments to biomass formulations. Another example is MetaDraft [61, 62], which builds draft reconstructions by leveraging selected reference models for related organisms. While this strategy speeds up the reconstruction process, it still requires user involvement to curate the transferred reactions and ensure consistency. Overall, GEM reconstruction tools differ considerably in terms of user experience and the extent of manual involvement they demand, ranging from platforms that emphasise detailed curation to those that prioritise efficiency.

In contrast to these semi-automated platforms, other reconstruction tools aim for a fully automated workflow with minimal user intervention (Table 1). Among these, CarveMe [59] represents a unique top-down approach that begins with a curated template model based on the BiGG database. Unnecessary reactions are then removed according to enzyme gene presence, determined by DIAMOND [63] alignment of the input bacterial protein sequences. CarveMe high-speed reconstructions are useful in large-scale studies, as generating a model from a bacterial genome containing 4000-6000 genes typically takes 1-3 minutes. gapseq [64] does not utilise protein sequences as input, but instead identifies genes directly from the genome. It uses a manually curated reaction set compiled from several databases (e.g., MetaCyc, ModelSEED), and extensive built-in pathway definitions to improve the coverage of metabolic capabilities. However, the procedure is time-consuming and typically requires 8–24 hours per model. ModelSEED [65] supports protein sequence alignment and is one

of the earliest and most established automated pipelines, producing draft reconstructions through a web service of the corresponding database [58]. The KBase [66], online platform builds on the updated ModelSEED pipeline but requires built-in RAST [67] annotation of the entire bacterial genome, providing an integrated environment for model reconstruction. All of the above tools support either universal or gram-positive/gram-negative bacterial model templates and can perform default gap filling or allow user-defined gap-filling media. Finally, AGORA [68, 69], although not a reconstruction tool per se, is a large-scale resource of semi-automatically built GEMs for human gut bacteria, available for download from the Virtual Metabolic Human (VMH) database [70]. Collectively, such tools (Table 1) provide efficient solutions for large-scale or high-throughput reconstruction tasks, offering researchers rapid access to functional draft models across a wide range of organisms.

| Tool | Type | Input | Nomencl. Reactions | Databases | Nomenclature Genes | Gap-filling | Gram-templ | Ref |
|---|---|---|---|---|---|---|---|---|
| CarveMe | Command line tool | Proteins | BiGG | BiGG | Adjusted IDs from input sequence | Top-down (universal model) | + | [59] |
| gapseq | Command line tool | Genome | modelSEED | modelSEED, KEGG, MetaCyc | Gene coordinates | Bottom-up | + | [64] |
| modelSEED | Web-service | Proteins | modelSEED | modelSEED | IDs from input | Bottom-up | + | [65] |
| Kbase | Web-portal | Genome | modelSEED | modelSEED | Coordinates from internal annotation | Bottom-up | + | [66], [65] |
| AGORA | Database | NaN | Recon | modelSEED and literature | Internal IDs and sequence | Bottom-up and curation | + | [68], [69] |

**Table 1.** Different tools for automatic reconstruction of functional genome-scale metabolic models (GEMs).

Despite their advantages, fully automated reconstruction tools also come with important limitations [55]. The resulting models often contain reactions with weak or lacking genetic support, inconsistencies in biomass composition, or incomplete representation of organism-specific pathways [71]. In many cases, these draft models are functional enough for simulations but still require post-reconstruction curation to achieve biological accuracy [72]. Automated pipelines also struggle with less-studied organisms, where genome annotation is sparse or database coverage is limited, leading to gaps or incorrect assignments [73]. Therefore, while these tools are invaluable for high-throughput applications and rapid generation of GEMs, they are best viewed as a starting point that should be refined through targeted manual curation and experimental validation [29].

# 6 Curation of GEMs and challenges for *de novo* reconstructions

Currently, the established practice is to base GEM reconstruction on a previously curated model from a related organism, or on a draft generated by a single reconstruction tool, followed by manual curation and verification [29]. This process becomes particularly challenging in *de novo* reconstructions of non-model organisms. While draft models may contain two to three thousand reactions, the limited biological knowledge available for non-model species often provides little guidance for the curation process, making it difficult to determine which reactions are reliable and which may be erroneous or irrelevant.

The most common experimental data used for model curation include measurements of biomass composition, auxotrophies for essential biomass components, carbon source utilisation, and gene essentiality screens [29]. Biomass composition experiments quantify the relative amounts of macromolecules and key metabolites, such as amino acids, nucleotides, lipids, and cofactors, required for cellular growth, providing the basis for constructing or refining the stoichiometry of the biomass reaction [45, 74]. Auxotrophy assays determine which metabolites an organism cannot synthesise autonomously by testing growth in minimal media supplemented with specific compounds; these data are used to verify the presence or absence of biosynthetic pathways and guide addition or removal of reactions [75]. Carbon source utilization experiments assess the ability of the organism to grow on different substrates by measuring growth rates or biomass yields in media containing a single carbon source, informing constraints on exchange reactions and the network's substrate usage capabilities. Gene essentiality screens identify genes that are indispensable for survival under defined conditions, typically through systematic knockout experiments (e.g., using transposon mutant libraries) [76]. Comparison with GEM predictions allows validation and refinement of gene–protein–reaction (GPR) associations and detection of missing or misannotated reactions [77]. Integrating these datasets ensures that the GEM reflects experimentally observed metabolic capabilities.

While integrating experimental datasets into GEMs ensures that the model reflects overall metabolic capabilities, it does not provide direct evidence for the activity of individual reactions or the correctness of specific gene–protein–reaction (GPR) associations. Achieving a model that is fully consistent with experimental observations requires additional steps, including manual curation, literature-based verification, and, where possible, incorporation of complementary data sources. This process often involves deciding which reactions should be included or removed, which GPR associations need adjustment, and how to reconcile inconsistencies between annotation and experimental phenotypes. For

poorly studied organisms, where genome annotation is limited and many genes have unknown functions, this curation becomes particularly challenging.

Another challenge arises when starting from a single draft model, as it provides only one initial network configuration, limiting the exploration of alternative metabolic arrangements, flux distributions or GPR rules that could potentially satisfy the experimental data. This makes it difficult to formulate hypotheses about how the network could function or produce specific metabolites. It also constrains the ways in which the model can be systematically adjusted to align with observed phenotypes.

All these challenges make the draft GEMs and the reconstruction tool used for their generation highly consequential for final model structure and accuracy. Moreover, emerging cross-tool studies [78–80] have revealed that models of the same organism reconstructed with different tools often exhibit more divergence than similarity. In fact, technical factors introduced by the reconstruction pipeline can shape model content more strongly than the underlying biological information, such as genome annotation or ecological context. For example, recent work [78] on environmental bacterial communities demonstrated that models reconstructed with the same tool were more similar to each other in their predictions of metabolite exchange than models of comparable communities generated with different tools. This study also shows that models built with different tools can capture complementary aspects of metabolic behaviour, motivating initial efforts to integrate them [78, 79].

Currently, integrating or even systematically comparing GEMs reconstructed with different tools or protocols remains a major challenge due to the limited degree of standardisation. Models frequently employ distinct nomenclatures for reactions and metabolites, depending on the underlying biochemical database (Table 1), which creates classical issues of identifier conversion. They may also differ in their representation of compartments; for example, a periplasmic compartment is included in CarveMe and gapseq reconstructions but is absent in ModelSEED and AGORA. Similarly, exchange reactions are implemented in tool-specific ways, such as the introduction of additional artificial boundary metabolites in ModelSEED. Finally, inconsistencies in gene integration generate variation both in gene identifiers and in the formulation of Boolean rules, which are often expressed differently across GEMs, further complicating direct comparison and integration.

Overall, GEM reconstruction remains challenging: drafts for non-model organisms offer limited reliability, and standard experimental datasets align models with phenotypes without proving direct reaction activity or GPR validation. Reliance on a single tool narrows the search space of potential GEM structures and weak standardization hinders comparison and integration of different GEMs. These issues argue for multi-tool workflows in GEMs reconstruction and curation to reduce pipeline bias and improve model performance.

# 7 Aims of the thesis

In this thesis, I present my work that addresses the methodological challenges associated with the reconstruction of genome-scale metabolic models (GEMs), particularly the lack of standardised frameworks for systematically comparing and integrating models generated with different tools. I introduce a consensus approach that, by enabling structural comparison across GEMs, enhances reconstruction and curation, moving beyond the constraints of individual tools.

Chapter I illustrates my initial attempts to reconstruct models using different tools and to compare them, highlighting the discrepancies observed and motivating the transition from single draft reconstructions to a consensus-based strategy. Chapter II introduces the methodology and the GEMsembler Python package (already published: Matveishina et al. mSystems 2025 [81]), which I developed for systematic comparison and integration of GEMs. By assessing structural differences and generating consensus models, this framework enables new strategies for reconstruction and curation, improving GEM performance with use cases in Lactobacillus plantarum and *Escherichia coli*. Finally, in Chapter III, I apply this consensus-based approach to the *de novo* reconstruction of two of the most abundant human gut bacteria, *Bacteroides uniformis* and *Phocaeicola vulgatus* (formerly *Bacteroides vulgatus*). Thus, this work establishes a consensus approach and demonstrates its utility in GEM reconstruction, systematic assessment of metabolic network uncertainties, and advancing knowledge of bacterial metabolism.

# Chapter I
# Initial comparison of genome-scale metabolic models (GEMs) of representative human gut bacteria reconstructed with different tools.

## 1.1 Introduction

Human gut bacteria exhibit remarkable metabolic diversity that underlies their ability to process nutrients, modulate host physiology, and influence health and disease [2, 4, 5]. Unravelling this diversity is challenging, as many species differ in genomic content, regulation, and environmental constraints, while experimental characterisation, including growth phenotypes, secreted metabolites, and dietary responses, remains limited and often inconsistent [82, 83]. To bridge this gap, several studies have combined genome-scale metabolic models (GEMs) with experimental data to explore the metabolic behaviour of different gut microbes and to systematically characterise and compare bacterial metabolism across species [84, 85].

Comparative analysis of GEMs has been used to characterise variation in microbial metabolism. For example, in constraint-based modelling analysis of the metabolism of two *Pelobacter* species [86], models of closely related bacteria were compared at the level of reactions and pathways, revealing both shared and species-specific metabolic capabilities, particularly in amino acid and cofactor metabolism. A subsequent study [87] reconstructed models for over 300 gut microbial species and examined differences in reaction content and inferred pathways, showing that metabolic repertoires correlate with phylogeny and ecological niche. More recently, a comparative work [88] systematically analysed multiple bacterial GEMs under simulated perturbations, identifying conserved and variable reactions and demonstrating that differences among networks can influence predicted metabolic

phenotypes. Together, these studies illustrate how systematic model comparison can uncover conserved and species-specific pathways, providing insight into metabolic diversity in the human gut microbiota.

My initial goal was to compare the metabolic capabilities of representative gut bacterial species in order to identify common and species-specific metabolic functions and pathways. Given the diversity of the human gut microbiome, nine bacteria were selected to represent three dominant phyla [3]. These include *Bacteroides uniformis*, *Bacteroides thetaiotaomicron*, *Parabacteroides distasonis*, and *Prevotella copri* from *Bacteroidetes*; *Akkermansia muciniphila* from *Verrucomicrobia*; and *Eubacterium rectale*, *Ruminococcus gnavus, Clostridium sporogenes*, and *Lactobacillus gasseri* from *Firmicutes*. Together, they reflect a broad functional and phylogenetic spectrum of the gut community.

In this chapter, I aim to reconstruct genome-scale metabolic models for these representative bacterial species using the three different tools: CarveMe [59], gapseq [64], and AGORA [68], each representing a distinct methodological strategy. I then compare the resulting metabolic networks with each other and assess which factors (e.g., the reconstruction approach or the input bacterial genome) have more influence on the network structure. The results of this comparison are the main motivation for my subsequent work on developing a method for systematic comparison of metabolic networks, and they lay the foundation for the consensus modelling approach proposed in my thesis.

## 1.2 Materials and Methods

### 1.2.1 Reconstruction of GEMs with different tools

Genome and protein sequence files for nine representative gut bacterial species were retrieved by NCBI assembly ID (Table 1.1) from RefSeq with ncbi_genome_download package [89]. The phylogenetic tree for these nine bacteria (Fig. 1.1) is obtained from NCBI taxonomy web service.

| Species | Type Strain | NCBI RefSeq assembly | Gram |
|---|---|---|---|
| *Akkermansia muciniphila* | ATCCBAA-835 | GCF_017504145.1 | Negative |
| *Bacteroides thetaiotaomicron* | VPI-5482 | GCF_900624795.1 | Negative |
| *Bacteroides uniformis* | ATCC8492 | GCF_900896415.1 | Negative |
| *Clostridium sporogenes* | ATCC15579 | GCF_000155085.1 | Positive |
| *Eubacterium rectale* | DSM17629 | GCA_000209935.1 | Positive |
| *Lactobacillus gasseri* | ATCC33323 | GCF_008868295.1 | Positive |
| *Parabacteroides distasonis* | ATCC8503 | GCF_900683725.1 | Negative |
| *Prevotella copri* | ATCC18205 | GCF_020735445.1 | Negative |
| *Ruminococcus gnavus* | ATCC29149 | GCF_009831375.1 | Positive |

**Table 1.1.** Nine human gut bacterial representatives and their genomes

Genome-scale metabolic models (GEMs) were reconstructed using CarveMe and gapseq tools (Fig. 1.1). For both pipelines, the Gram status of each organism was specified using the appropriate Gram-negative or Gram-positive template. Reconstructions were performed under anaerobic conditions and without any medium-specific gap-filling, to preserve the intrinsic metabolic capabilities of each genome. CarveMe models were generated from the protein sequence files, whereas gapseq models were reconstructed from the genome sequences, both in FASTA format. In parallel, corresponding AGORA1 models and their associated genome files were retrieved from the VMH database, resulting in a total of 27 GEMs, three per organism (Fig. 1.1).

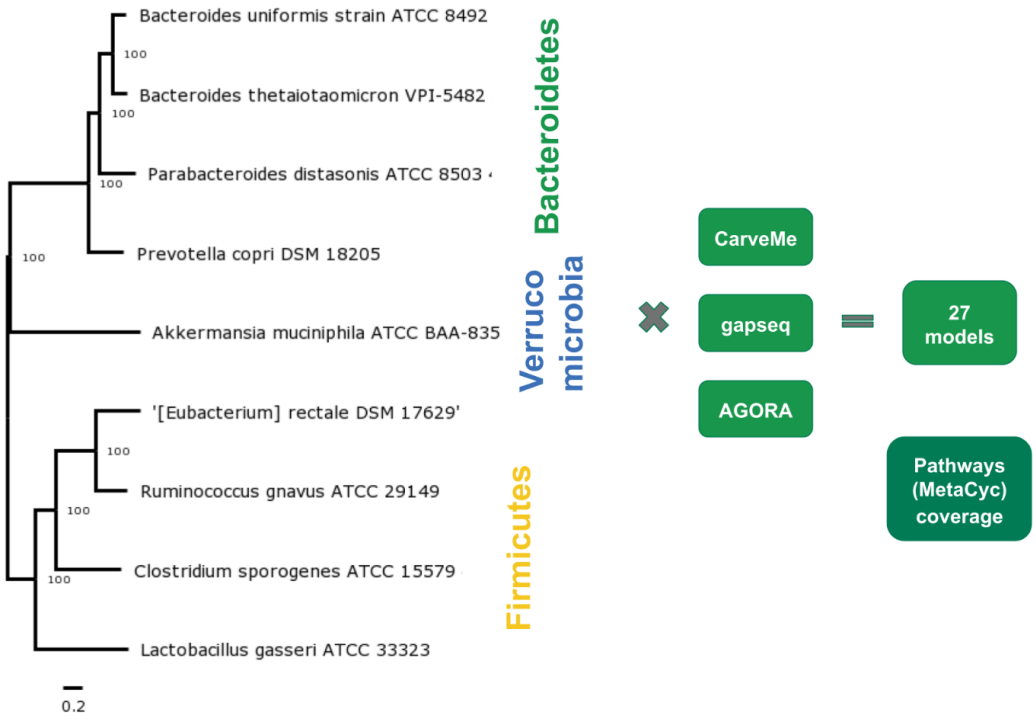


**Figure 1.1.** Taxonomic tree of nine human gut bacterial representatives and the workflow with their genome-scale metabolic model (GEM) reconstruction and pathway coverage comparison.

### 1.2.2 GEMs comparison with pathway coverage

For the comparison of reactions and metabolites included in different models, all identifiers were converted into the BiGG nomenclature. For AGORA models, outdated BiGG identifiers were updated to the current version using information from the BiGG database. For gapseq models, metabolites and reactions were mapped via cross-references from the ModelSEED and MetaNetX databases, incorporating all available conversion matches.

For the pathway-based comparison, the complete set of MetaCyc pathways and their associated reaction lists were downloaded from the MetaCyc database (Fig. 1.1). Reaction identifiers from all models were then converted to MetaCyc IDs through BiGG and

MetaNetX cross-references. All accessible crosslinks were used to collect corresponding MetaCyc reaction identifiers. Pathway coverage for each model was calculated as the proportion of reactions present in the model relative to the total number of reactions defined for each MetaCyc pathway. The resulting pathway abundance matrix was subjected to hierarchical clustering using the ComplexHeatmap R package, applying Euclidean distance and complete linkage as default parameters of the Heatmap function.

## 1.3 Results

### 1.3.1 GEMs reconstructed with different tools cluster according to the tool

As a first step in the metabolic modelling, I generated genome-scale metabolic models (GEMs) for nine representative human gut bacterial species and assessed the differences in their resulting metabolic networks.

To evaluate the overall similarity between models reconstructed with CarveMe, gapseq, and AGORA, the sets of reaction and metabolite identifiers were compared across the three reconstruction approaches (Fig. 1.2). The number of shared reactions and metabolites was largely comparable among any pair or combination of methods. Across all species, approximately half of the reactions and metabolites were common between different reconstruction pipelines, representing a conserved metabolic core. The remaining fraction consisted of method-specific reactions and metabolites, reflecting differences in reconstruction strategies (Fig. 1.2).

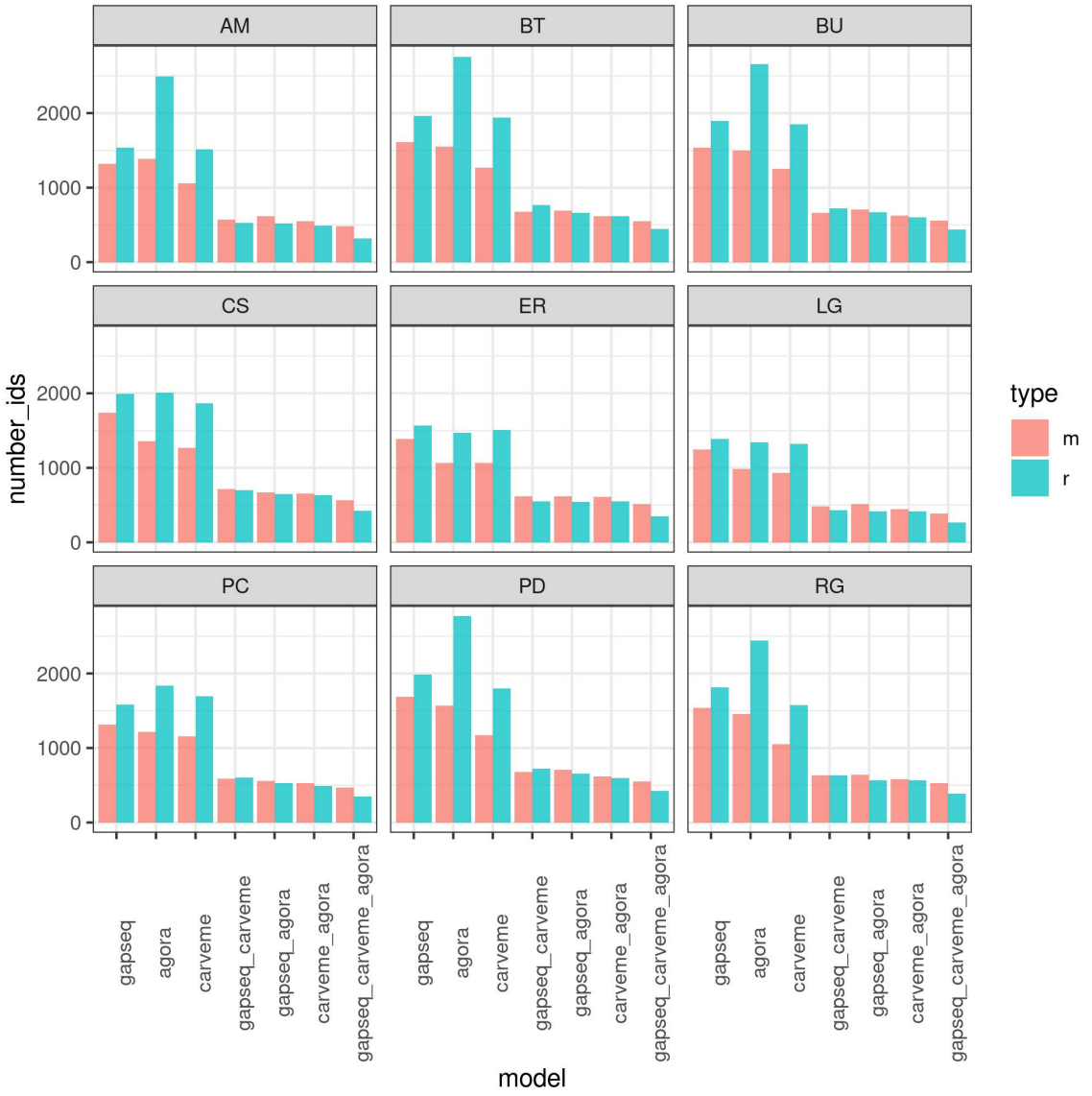


**Figure 1.2.** Comparison of metabolites (m), reactions (r) in CarveMe, gapseq, AGORA models for 9 species. Columns for one model show the original number of ids in the model. Columns for two or three models show the number of ids after conversion and intersection. AM - *A. muciniphila,* BT - *B. thetaiotaomicron* BU - *B. uniformis*, CS - *C. sporogenes*, ER - *E. rectale*, LG - *L. gasseri*, PC - *P. copri*, PD - *P. distasonis*, RG - *R. gnavus*.

Given the large number of reactions that differed among the reconstructed models, I next compared them at the pathway level to better summarise the variation in metabolic content (Fig. 1.3). In total, 246 MetaCyc pathways showed a coverage greater than 0.5 and contained more than two reactions present in at least one model of at least one species. Based on hierarchical clustering of the pathway coverage, only the models of *Akkermansia muciniphila* and *Lactobacillus gasseri* clustered by species, whereas all other models grouped according to the reconstruction method rather than their taxonomic origin. This pattern indicates that the technical factor, the choice of the reconstruction tool, can

influence the overall network structure even more strongly than the biological factor, i.e., the underlying bacterial genome.

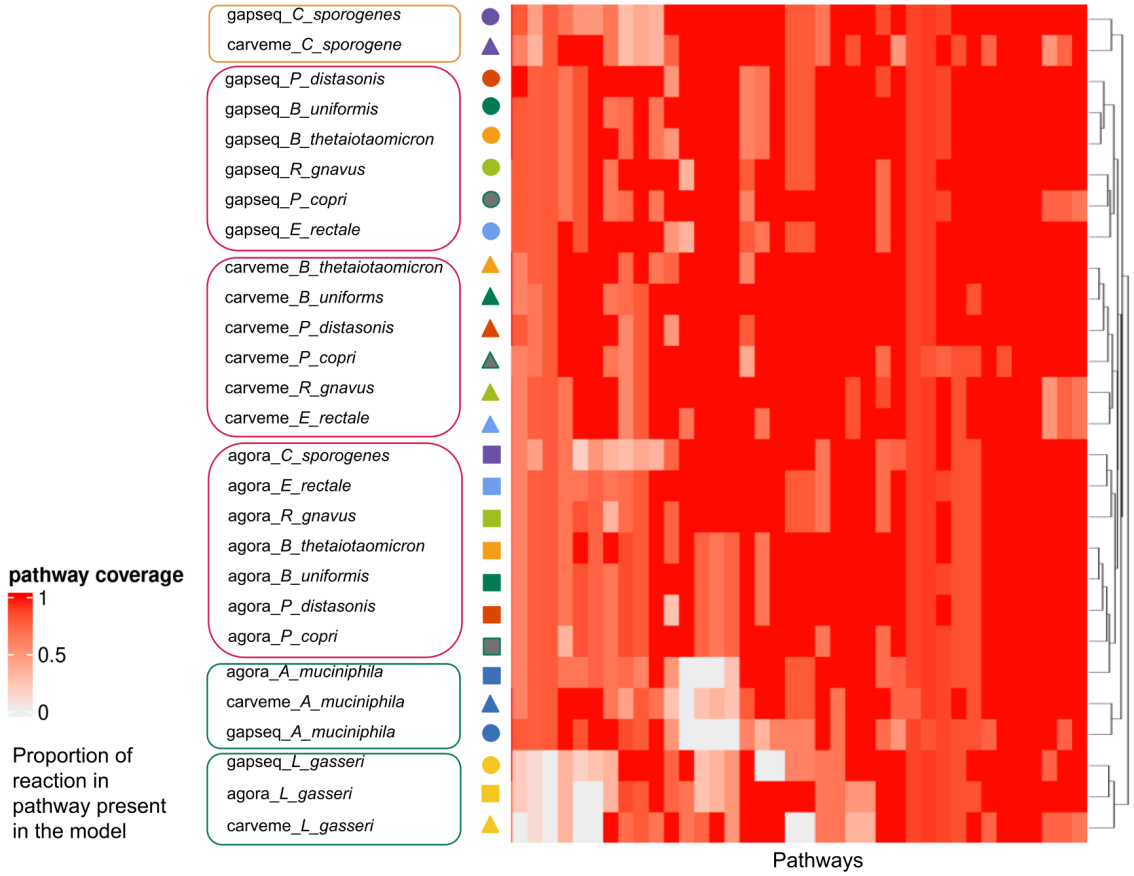


**Figure 1.3.** GEMs cluster by the reconstruction tool instead of the corresponding bacterial species when considering all reactions to calculate pathway coverage. Pathways with the highest coverage are shown. Green rectangles highlight GEMs clustering by species, orange rectangles highlight a subset of GEMs clustering by species, while red rectangles highlight GEMs clustering by the reconstruction tool. Symbols correspond to the model type (circle for gapseq, triangle for carveme, and square for agora), while their colours correspond to different species.

### 1.3.2 GEMs reconstruction differs in GPR rule assignment

Next, I hypothesised that the network differences among different types of models could result from the gap-filling procedures employed by each reconstruction tool. To test this, I evaluated whether models were more consistent when considering only reactions with genetic evidence (GPRs), i.e., reactions associated with enzyme-coding genes in the corresponding bacterial genome. Contrary to my hypothesis, clustering of models based on pathways with known genes was still largely driven by the reconstruction method, rather than by the species-specific genomic information (Fig. 1.4). Only *Clostridium sporogenes* formed a species-specific cluster after filtering for GPR-containing reactions, in addition to the species-specific clusters previously observed for *Akkermansia muciniphila*

and *Lactobacillus gasseri*. These results indicate that the majority of differences among the networks arise not from gap-filling procedures, but from the enzyme annotation and GPR rule assignment.

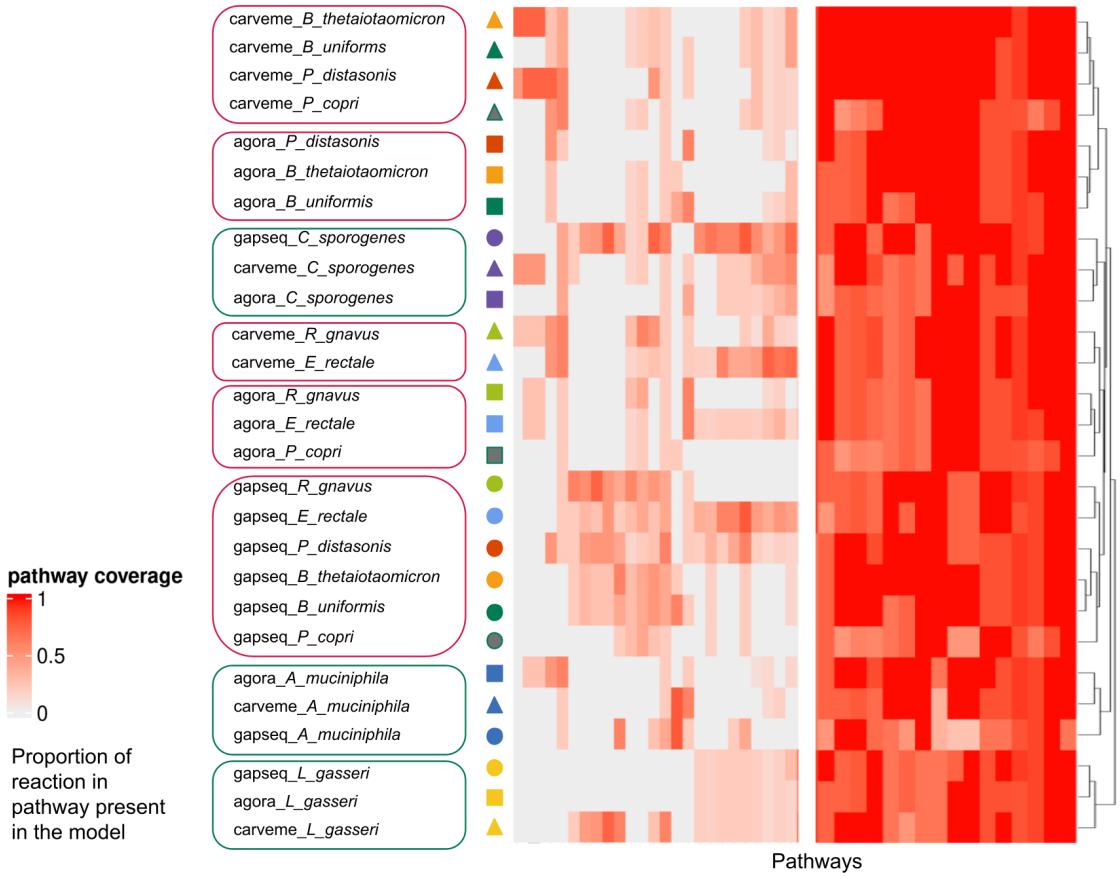


**Figure 1.4.** GEMs cluster by the reconstruction tool instead of the corresponding bacterial species when considering only not gap-filled reactions with genetic evidence (GPRs) to calculate pathway coverage. Pathways with the highest and the lowest coverage are shown. Green rectangles highlight GEMs clustering by species, while red rectangles highlight GEMs clustering by the reconstruction tool. Symbols correspond to the model type (circle for gapseq, triangle for carveme, and square for agora), while their colours correspond to different species.

To further test the hypothesis that differences among reconstruction methods arise from the genes included in the models, I focused only on reactions associated with genes common to all three model types (CarveMe, gapseq, and AGORA) based on genetic evidence. Under this restriction, the models clustered predominantly according to species, with the exception of *Bacteroides thetaiotaomicron* and *Parabacteroides distasonis*, which did not form distinct species clusters (Fig. 1.5). These results confirm that the majority of variation among models generated by different reconstruction tools originates from differences in genome-based enzyme annotation, together with the downstream reconstruction procedures (e.g. gap-filling).

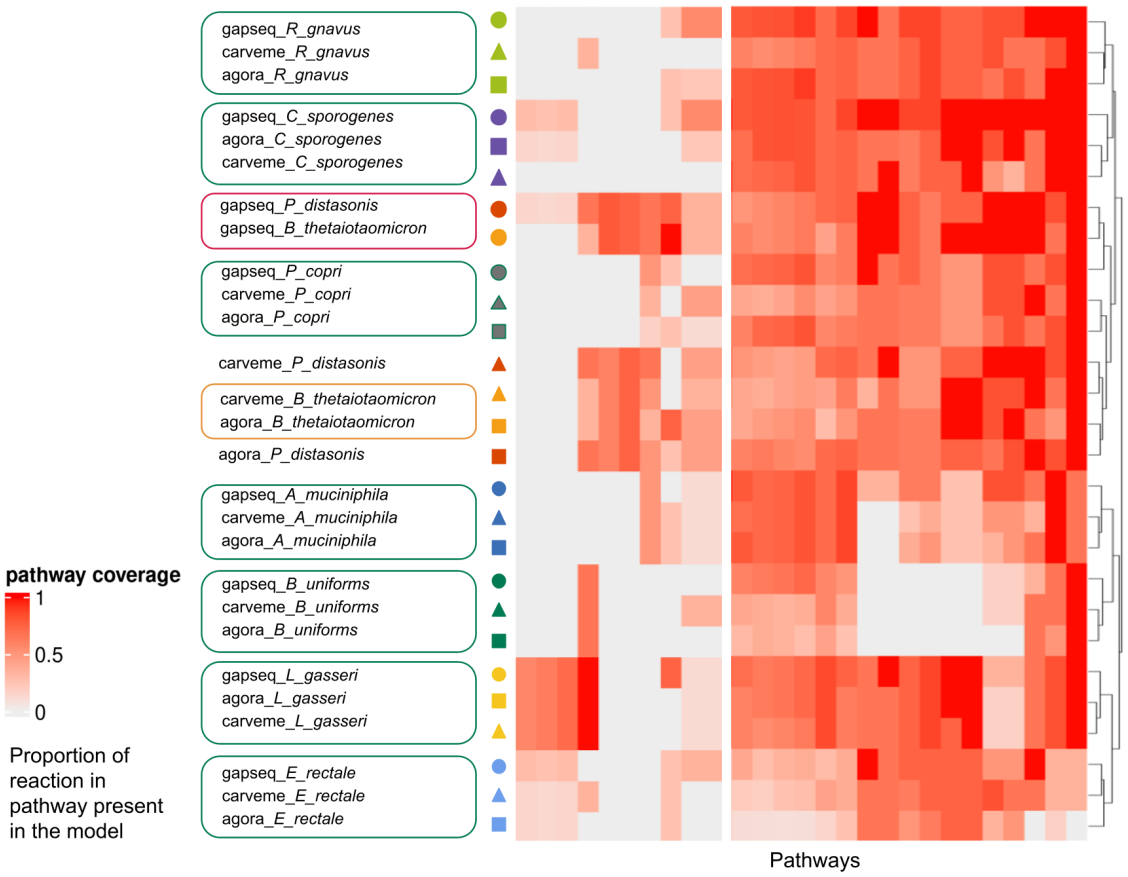


**Figure 1.5.** GEMs cluster by the corresponding bacterial species when considering only reactions with genes common between all three different GEMs to calculate pathway coverage. Pathways with the highest and the lowest coverage are shown. Green rectangles highlight GEMs clustering by species, orange rectangles highlight a subset of GEMs clustering by species, while red rectangles highlight GEMs clustering by the reconstruction tool. Symbols correspond to the model type (circle for gapseq, triangle for carveme, and square for agora), while their colours correspond to different species.

## 1.4 Discussion

The results of this chapter demonstrate substantial differences in GEMs reconstructed with different tools (CarveMe, gapseq, and AGORA), reflecting variations in reconstruction algorithms, templates, and database dependencies. While it was expected that the models built with different tools would be different from each other, because of the differences in the underlying gap-filling procedures, it was surprising that models clustered by tool even when considering only reactions associated with genes. This indicates the decisions on gene inclusion and GPR definitions are the primary drivers of the observed model difference. Such substantial variation indicates several layers of uncertainty in automated GEM reconstructions.

This work also revealed that systematically comparing GEMs generated by different reconstruction tools is far from straightforward. The complexity of model structure,

variation in all its components, and ambiguity in nomenclature cross-links make direct model-to-model comparisons challenging. To address the challenge, I aimed to develop a tool that enables a structured and systematic comparison of metabolic networks and to establish a consensus approach for GEMs analysis. This approach provides a framework to assess and resolve uncertainties in the models, ultimately improving understanding of bacterial metabolism.

# Chapter II
# GEMsembler: consensus model assembly and structural comparison improve functional performance of GEMs

## 2.1 Introduction

Genome-scale metabolic models (GEMs) have become central tools for studying microbial metabolism, yet their growing number and the diversity of reconstruction pipelines create a pressing need for systematic comparison [31, 55]. While individual reconstructions provide valuable insights into metabolism, differences in the outputs of reconstruction tools, model structures, and underlying assumptions, together with limited standardisation, make it difficult to assess their consistency or integrate them within a unified framework [55]. Existing efforts to compare models remain limited in their applicability.

In most cases, approaches to comparing GEMs do not go beyond converting and matching metabolite and reaction identifiers (IDs). In this context, MetaNetX [90, 91] is a particularly useful resource: an online platform that links a wide variety of namespaces for metabolites and reactions. It extends cross-database mappings already present in commonly used biochemical databases for GEM reconstruction, such as ModelSEED [58], KEGG [56], and BiGG [33]. For example, MetaNetX has been used to compare IDs across models reconstructed with different tools in the COMMIT pipeline [79], which generates consensus models as part of microbial community analyses.

When relying on identifier mappings, several well-known sources of ambiguity limit the reliability of cross-model comparisons. Cross-references between databases are rarely one-to-one (often many-to-one, one-to-many, or even many-to-many), for example because

of differences in tautomeric form, stereochemistry, or the level of generality of a compound or reaction. Out-of-date or missing IDs and duplicate reactions add further complications. Indirect mappings that pass through multiple databases, where IDs are converted into an intermediate namespace, like in MetNetX [90, 91], propagate and can amplify mistakes. Altogether, naive ID-based matching with the selection of a single identifier is prone to false conversions, which hides the real biological agreement and disagreement between GEMs.

One example that sought to mitigate this problem comes from a study, comparing GEMs reconstruction tools [55], where identifiers were first converted via MetaNetX [90, 91] and reactions were then matched by their reaction equations, with equations reconstructed from metabolite identifiers and compared across models. However, the code used for this procedure was a custom implementation developed for that study, not released as a reusable tool or package. Another attempt is modelBorgifier [92], an addition to the COBRA Toolbox [32] that assists in merging two GEMs in a semi-automatic manner, relying on user review and confirmation of conversion results. While such efforts can improve matching within a study, they remain labour-intensive and they do not resolve identifier ambiguity in a systematic or reproducible way.

On current evidence, the first dedicated tool for merging heterogeneous GEMs appears to be the recently published Python package mergem [80]. It merges input models into a single union by matching metabolite identifiers and reaction equations. The package is geared towards comparing particular models rather than harmonising them to a single nomenclature. It processes models sequentially, retaining the identifier from the first occurrence; as a result, the merged model can contain a mixture of namespaces. When an identifier originates from a later model namespace, it becomes impossible to tell whether this reflects a technical failure to convert it into the first model's namespace or a genuine difference in model content. More importantly, mergem [80] does not account for key attributes of GEMs, such as genes, gene–protein–reaction (GPR) rules, or reaction boundaries and therefore directionality. Consequently, to date, no framework can provide rigorous model comparison, integrate all model features, and support fine-tuned, flexible combinations of GEMs.

In this chapter, I introduce my recently published work (Matveishina et al. mSystems 2025 [81]) on GEMsembler, a Python package designed to make models built by different reconstruction pipelines directly comparable and to enable a consensus approach in GEM reconstruction and curation. The package centres on three capabilities: first, structural comparison of GEMs produced by different tools, while tracking the origin of each feature; second, systematic assessment of network confidence, defined by agreement between models on metabolites, reactions, and genes; third, assembly of alternative model combinations followed by evaluation of their predictive performance for growth,

auxotrophy, and gene essentiality. In addition, I provide a protocol for semi-automated curation based on a consensus approach across multiple GEMs.

I demonstrate the framework on two case studies, *Escherichia coli* and *Lactiplantibacillus plantarum* (formerly *Lactobacillus plantarum*). In both cases, GEMsembler-curated consensus models achieve higher accuracy than current gold-standard, manually curated models for auxotrophy and gene essentiality. The workflow also identifies the specific features that drive these improvements, providing clear targets for experimental follow-up to close the current knowledge gaps and address remaining uncertainties in metabolic network reconstructions.

## 2.2 Materials and Methods

### 2.2.1 GEMsembler package development and overview

I designed GEMsembler as a Python package to interoperate tightly with the major GEM manipulation and analysis library, COBRApy [40]. A bioinformatician in my group, Bartosz J. Bartmański, assisted with files handling and the packaging of the software for PyPI. GEMsembler depends on standard Python libraries, BLAST for gene mapping [93], and the MetQuest package for topological network analysis [94]. Of these, only BLAST must be installed separately by the user, because its installation procedures differ across environments.

I organized GEMsembler as a four-stage pipeline: (i) conversion of input GEMs, (ii) supermodel assembly, (iii) comparison and consensus model generation, and (iv) assessment of model agreement followed by downstream functional analysis (Fig. 2.1). I decided that input GEMs in SBML format must be accompanied by metadata specifying model type (the reconstruction tool), which informs GEMsembler how each model should be handled. In the current implementation, models produced by CarveMe [59], gapseq [64], MetaNetX [91] and modelSEED [65] are supported, as well as reconstructions retrieved from the AGORA [69] and BiGG [33] databases. To accommodate additional sources, Bartosz J. Bartmański introduced a custom model type that enables user-defined import logic.

To enable accurate cross-model mapping of gene–protein–reaction (GPR) rules, I added optional genomic inputs (Fig. 2.1). These comprise: (i) the bacterial genome or protein FASTA files used to generate each model, and (ii) either a local genome file or an NCBI assembly identifier for automated download. The latter serves as the target reference to which gene IDs in the final GPRs are converted.

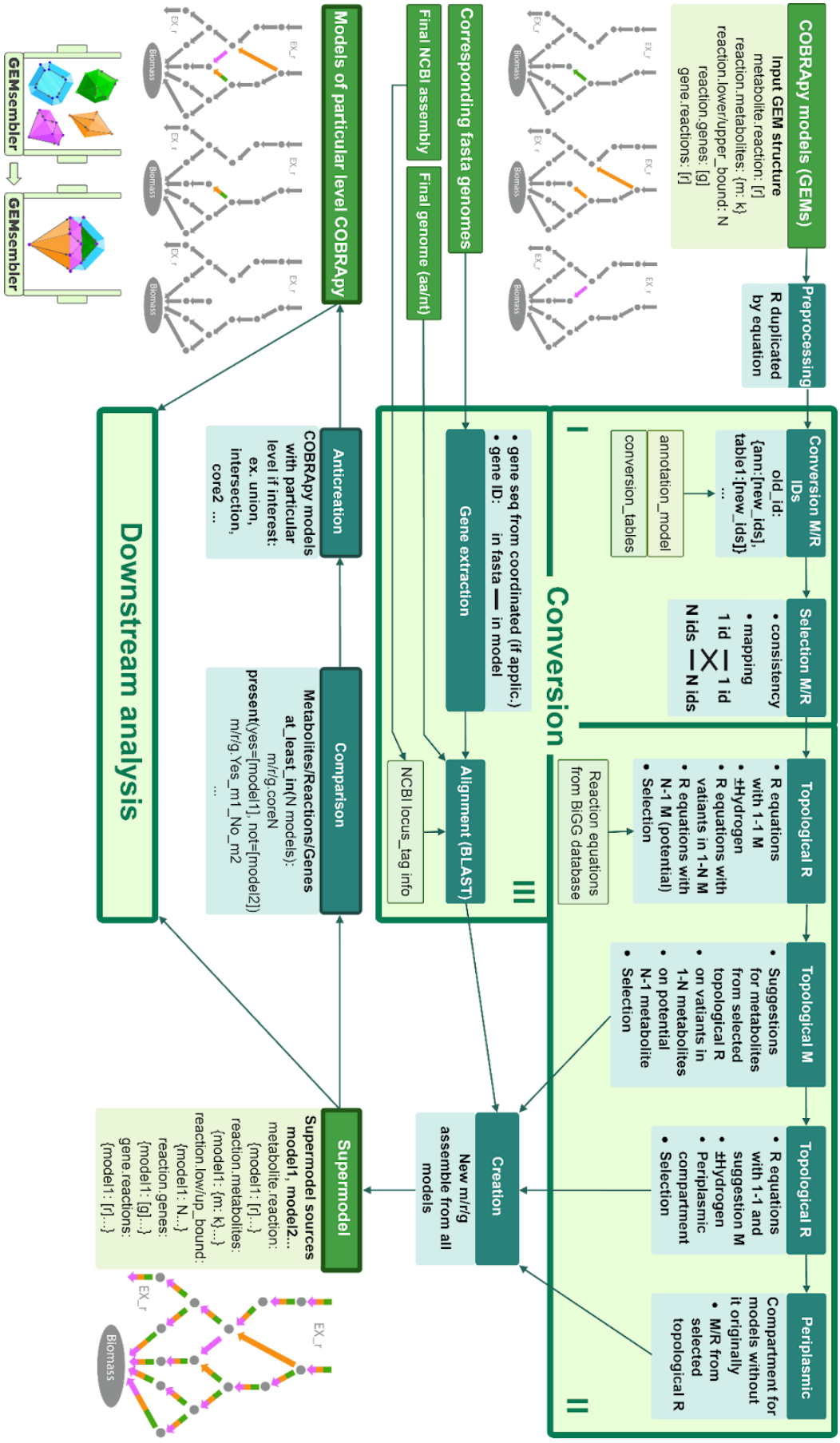


**Figure 2.1.** GEMsembler workflow. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

I designed GEMsembler to emit two principal outputs (Fig. 2.1): a supermodel, that tracks the origin of each feature, and consensus models that are the results of comparisons among selected input models. Using COBRApy, I implemented consensus models as standard GEMs in SBML format. The supermodel was initially stored as a Python object in pickle format, but later my husband, Gleb E. Gavrish, implemented JSON export for the supermodel, enabling cross-platform integration of GEMsembler. For the downstream analysis stage, the required inputs typically include a specification of the growth medium and the metabolites or pathways of interest, and my package produces plots, tables, and interactive pathway maps in HTML format.

### 2.2.2 GEMs conversion step

GEMsembler workflow begins by converting metabolite identifiers in all input models to BiGG IDs [33] (Fig. 2.1). To do this, I prioritised multiple cross-reference sources and applied them in a fixed order to maximise precision and reproducibility:

1. Intersection first. When BiGG IDs were present both in the model's own annotations and in the originating database (ModelSEED [58] or BiGG [33]), I used the intersection as the highest-priority mapping.
2. Model annotations. If there was a discrepancy, the BiGG IDs listed in the model annotation field served as the second priority.
3. Origin database crosslinks. If model annotations were missing, crosslinks from the model's source database were used as the third priority.
4. Meta-databases. When neither of the above yielded a match, I consulted additional crosslink resources, notably MetaNetX [90], as the fourth priority.
5. Pattern rules. If no BiGG ID was recovered, heuristic renaming rules were applied as a fifth priority (for example, transforming certain AGORA IDs by replacing a trailing single underscore with a double underscore).
6. Direct lookup. As a last resort, the unmodified identifier was checked directly against the BiGG namespace.

For each supported model type, I implemented a dedicated conversion routine that reflects its native nomenclature and attributes. For custom model types, the conversion step must be adapted to the submitted naming scheme.

Because cross-database mappings are often non-bijective, in the selection stage (Fig. 2.1) I explicitly tracked ambiguity: one-to-many, many-to-one, and many-to-many relations. When different input models shared the same biochemical nomenclature (for example, gapseq and ModelSEED models, both rooted in the ModelSEED database), I checked consistency across models to eliminate ambiguous results, if possible. For the main

analysis path, only metabolites with a unique one-to-one mapping to BiGG were advanced; ambiguous cases were retained and handled separately.

I decided to convert reactions based on their equations (Fig. 2.1), following the approach from a study comparing GEM reconstruction tools [55], to preserve correct network topology. To increase the number of successful mapping I composed reaction equations using only metabolite identifiers, ignoring stoichiometry, and tolerated variation in hydrogen inclusion. First, I reconstructed equations in the BiGG namespace using only metabolites with one-to-one mappings to establish an unambiguous baseline. Metabolites without unique conversions (e.g., one-to-many mappings) were then resolved by enumerating the possible mappings to compose candidate reaction equations and automatically checking whether any of these equations exist in the BiGG reaction set. When a candidate equation matched, the corresponding mapping was fixed for that metabolite and propagated to all reactions containing it. For models without a periplasmic compartment, I allowed compartment reassignment if moving a metabolite from the cytosolic or extracellular compartment to a periplasmic compartment yielded a reaction present in BiGG. Finally, reactions with identical equations across input models and BiGG were merged, retaining the identifier used by the largest number of models in the BiGG database so as to favor the most widely adopted identifiers.

To make the conversion transparent and to leave room for potential user control, I recorded every conversion stage in a dedicated GatheredModels class that precedes supermodel assembly. Although only one converted ID, if selected, is used in supermodel assembly, GatheredModels retains all alternative mappings for inspection, including reaction identifiers obtained solely via MetaNetX [90] or other databases that were not confirmed by equation-level evidence.

I designed GEMsembler to optionally accept the genomes used for the original reconstructions together with either an NCBI assembly ID or a FASTA genome file for the output; when these inputs are provided, the software converts model gene identifiers to the target genome (Fig. 2.1). If an NCBI assembly ID is supplied, the assembly is downloaded automatically with the ncbi_genome_download package [89] and locus tags from that assembly are used as gene IDs; otherwise, genes are mapped to the identifiers present in the user-provided FASTA genome file. Tool-specific changes in gene naming are handled during this step, including punctuation changes introduced by some tools, such as added dots or replaced underscores. For gapseq, gene sequences must be additionally retrieved, since genes are represented by coordinates. Actual gene conversion is performed by sequence alignment with BLAST [93].

To handle the diversity in how GPRs were formulated across different model types, I standardised all GPR rules to a canonical two-level form: each rule is a Boolean expression

where genes are grouped inside parentheses with AND, and those groups are connected by OR. This corresponds to a disjunctive normal form (DNF), which guarantees a single, comparable structure rather than arbitrary nesting. CarveMe models already provided GPRs in this format; other inputs could contain any amount of nesting and mixed operators. I therefore implemented a standardisation step that parses each GPR rule and converts it to DNF by applying Boolean distributive law via the SymPy library [95].

### 2.2.3 The supermodel structure and functionality

I assembled all information from the input GEMs into a Python structure I call the supermodel, which resembles the COBRApy implementation for a GEM (Fig. 2.1). Metadata from every source model are preserved in a "notes" attribute. The supermodel comprises three core classes, metabolites, reactions, and genes, each with three main fields: "assembly", "comparison", and "not_converted". The "assembly" field aggregates, for each entity, the instances drawn from all input models; "comparison" stores comparison results and is empty at the time of supermodel creation; and "not_converted" contains elements that could not be mapped. For direct access, I also retained per-model subfields within metabolites, reactions, and genes. By default, GEMsembler sacrifices non-converted entities in favour of a unified nomenclature and higher conversion confidence; however, if users prefer not to lose non-converted content, the pipeline can mix these entries into "assembly" using their original identifiers from the input models via the "do_mix_conv_notconv" parameter.

To keep the structure familiar, I mirrored key COBRApy attributes for metabolites, reactions, and genes (Fig. 2.1). For example, metabolites include "name" and "reactions"; genes include "reactions"; reactions include "metabolites", "genes", "gene_reaction_rule", "lower_bound", and "upper_bound". Within each attribute, I added separate fields for every input model alongside their union. This layout enables direct comparison of attribute values per source model and their assembly, while preserving the origin for all features.

I implemented various comparison functionalities directly in the supermodel (Fig. 2.1). In particular, the methods "at_least_in" and "exactly_in" compute features supported by at least, or exactly, X input models, respectively, and "present" identifies features included by some models and excluded by others. Beyond set-level comparisons of metabolites, reactions, and genes, I also implemented comparisons of their attributes. For reaction boundaries, I intersected the lower–upper intervals across the selected models and then, based on possible model combinations, took the union of the resulting intersections. Because I ensured consistent orientation of reactants and products during supermodel assembly, reaction directionality is compared via their numeric bounds. For stoichiometric coefficients, I applied an arbitrary rule: for each metabolite, the coefficient is taken as the

most frequent value among the models of interest, with averaging within possible model combinations. In cases of agreement, the coefficient remains unchanged; where models disagree, all are taken into account, but that can lead to artificially unbalanced reactions. To address this issue, further I introduced a reaction-balance check and subsequent coefficient adjustment. I did not implement any special treatment for growth-related terms, so the Growth-Associated Maintenance (GAM) value is treated like any other coefficient and the Non-Growth-Associated Maintenance (NGAM) like the lower bound of the ATP maintenance reaction (ATPM). Thus, if GAM and NGAM values are known, they should be introduced into the GEMs outside GEMsembler. For gene–reaction rules, I decided to identify AND groups of genes on which all models of interest agree (a group can be treated as a single clause or genes within it can be intersected separately, controlled by the "and_as_solid" parameter), and then to unite these agreed parts with OR across the relevant model combinations. All comparison outputs are recorded in the "comparison" field of the supermodel with assigned comparison identification (for example, core3).

Each comparison can comprise a corresponding consensus model, and I implemented its extraction as a standard GEM in SBML format (Fig. 2.1). On top of the comparison that forms a given consensus model, I enabled targeted addition or removal of reactions according to user-specified lists, and, when a reaction is added, all of its attributes are taken from the assembly level. The consensus level for genes and for the biomass reaction can be set independently. To resolve orphan metabolites across compartments, additional transport reactions can be inserted from the overall assembly. To control reaction stoichiometry, GEMsembler evaluates mass and charge balance using metabolite formulas from the BiGG database, when available. If a reaction is unbalanced but can be balanced by adopting coefficients from the original models or by adding or removing hydrogen, the stoichiometry is adjusted accordingly.

### 2.2.4 Downstream analysis of the network topology and function

With supermodel and consensus models available, a wide range of downstream analyses become feasible. The analyses I consider most insightful I implemented directly in the GEMsembler package, enabling exploration of network topology and functional behaviour of consensus GEMs.

For topology-driven pathway search, I adapted the MetQuest package [94]. The package accepts a path to an SBML/XML model file, a growth-medium specification as a configuration file, and optional parameters such as a list of metabolites of interest for targeted pathway extraction. MetQuest enumerates all feasible paths up to a specified length from medium components to reachable metabolites, using only compounds already available. Because networks with over a thousand metabolites and reactions can require

more than 16 GB of RAM (at least 32 GB recommended), a bioinformatician in my group, Bartosz J. Bartmański, transferred this step to a separate "pathsfinding" module so it can be executed on a high-performance cluster, when available. The module writes an HDF5 file, "metquest.h5", containing the full dictionary of biosynthetic pathways up to the chosen length, and a "shortest_paths.pkl" file with a smaller dictionary of the shortest paths, three per metabolite by default, for either all metabolites or a user-specified subset. To analyse outputs from multiple original and consensus models simultaneously, I wrote the "run_metquest_results_analysis" function, that summarises what metabolites via what pathways in which models can be synthesised.

Another analysis to explore network topology that I implemented in GEMsembler is called "get_met_neighborhood", and it computes and visualises the network neighbourhood around a specified metabolite. The user provides the starting metabolite, a radius in number of reactions, and a threshold for highly connected nodes; the function returns the subnetwork induced by all reactions within that radius, stopping at metabolites involved in more reactions than the threshold, and generates the corresponding plots. This function is also used to compute each reaction's distance from a pathway's biosynthetic product, enabling distance-based summaries and filtering.

To assess functional behaviour of the models, I used the FBA and pFBA functions in COBRApy. In GEMsembler, I implemented growth simulation in specified media and testing of production either for all biomass components or for user-selected metabolites. Production of a metabolite is counted when the flux through its demand reaction, set as the simulation objective, exceeds 0.001 mmol $gDW^{-1}$ $h^{-1}$. In the corresponding pFBA solution, reactions carrying more than 0.001 mmol $gDW^{-1}$ $h^{-1}$ define a biosynthetic pathway for that metabolite. I implemented this biosynthesis analysis simultaneously, together with a summary across a dictionary of models, in the "run_growth_full_flux_analysis" function of GEMsembler.

As output, the functions "run_metquest_results_analysis" and "run_growth_full_flux_analysis" produce summary plots and tables of production and pathway agreement for metabolites of interest across all models. I also made GEMsembler automatically generate companion tables listing all the identified paths, together with interactive network maps built using the networkx (v 3.3) [96] and pyvis (v 0.3.2) [97] packages.

### 2.2.5 Case study: draft models reconstruction, supermodel, and consensus models generation for *E. coli* and *L. plantarum*

To demonstrate GEMsembler's functionality, I selected two bacteria, *Lactiplantibacillus plantarum* WCFS1 (LP) and *Escherichia coli* BW25113 (EC), and reconstructed draft GEMs for them using the CarveMe and gapseq command-line tools and the ModelSEED

web server. Protein sequences were used for CarveMe and modelSEED, whereas gapseq was run with nucleotide sequences. For *L. plantarum* I used assembly GCF_000203855.3. For *E. coli* I reconstructed models *de novo* from the BW25113 genome associated with the KEIO gene-essentiality dataset (available at https://fit.genomics.lbl.gov/cgi-bin/org.cgi?orgId=Keio). I selected the gram-positive template for *L. plantarum* and gram-negative for *E. coli* and used default gap filling without specifying a medium.

In parallel, species models were downloaded from the AGORA2 collection [69]at https://www.vmh.life/files/reconstructions/AGORA2/version2.01/sbml_files/individual_reconstructions/. Because AGORA2 does not distribute genomes, I paired the *L. plantarum* model with the genome from AGORA1 (https://www.vmh.life/files/reconstructions/AGORA/genomes/AGORA-Genomes.zip). For *E. coli*, I used assembly GCF_000750555.1, which matches the AGORA2 model's gene identifiers when locus tags are selected (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/750/555/GCF_000750555.1_ASM75055v1/GCF_000750555.1_ASM75055v1_cds_from_genomic.fna.gz)

For both organisms, I first converted the input models using the "GatheredModels" class via its "run" method. I then assembled two supermodels with the "assemble_supermodel" method of the "GatheredModels" class: a default version with "do_mix_conv_notconv" set to False, and a mixed version with "do_mix_conv_notconv" set to True. In each case, I ran input GEMs comparison with the supermodel method "get_all_confident_levels" and generated consensus GEMs with the function "get_models_with_all_confidence_levels". These models were subsequently used for topology-driven pathway searches and for growth and pFBA-based pathway analyses.

As a baseline for comparison, I used four original models for each organism, AGORA, CarveMe, gapseq, and ModelSEED, converted to the BiGG nomenclature with GEMsembler's mixed approach and without adding transport reactions.

### 2.2.6 Models curation for *E. coli* and *L. plantarum*

To ensure GEMs reproduction of growth phenotypes I curated the models in two stages: (i) the biomass reaction composition, and (ii) the biosynthetic pathways for biomass components under the specified growth medium, PMM5 for *L. plantarum* and M9 for *E. coli*.

PMM5 = 'pi_e': 1000, 'glc__D_e': 10, 'na1_e': 1000, 'ac_e': 1000, 'ascb__L_e': 10, 'arg__L_e': 10, 'glu__L_e': 10, 'ile__L_e': 10, 'leu__L_e': 10, 'met__L_e': 10, 'phe__L_e': 10, 'thr__L_e': 10, 'trp__L_e': 10, 'tyr__L_e': 10, 'val__L_e': 10, 'nac_e': 1000, 'pnto__R_e': 10, 'ribflv_e': 1000, 'mg2_e': 1000, 'cl_e': 1000, 'ca2_e': 1000, 'mn2_e': 1000, 'fe3_e': 1000, 'fe2_e': 1000, 'zn2_e': 1000, 'so4_e': 1000, 'cobalt2_e': 1000, 'cu_e': 1000, 'mobd_e': 1000, 'k_e': 1000, 'cu2_e': 1000, 'h2o_e': 1000, 'h_e': 1000

M9 = ’glc__D_e’: 10, ’pi_e’: 1000, ’co2_e’: 1000, ’mg2_e’: 1000, ’cl_e’: 1000, ’ca2_e’: 1000, ’mn2_e’: 1000, ’fe3_e’: 1000, ’fe2_e’: 1000, ’zn2_e’: 1000, ’so4_e’: 1000, ’cobalt2_e’: 1000, ’mobd_e’: 1000, ’k_e’: 1000, ’cu2_e’: 1000, ’h2o_e’: 1000, ’h_e’: 1000, ’ni2_e’: 1000, ’sel_e’: 1000, ’nh4_e’: 1000, ’na1_e’: 1000, ’o2_e’: 1000, ’tungs_e’: 1000, ’slnt_e’: 1000

To decide which biomass components to retain, I first ran preliminary growth and biomass components synthesis analyses on the original models after conversion to the BiGG nomenclature using GEMsembler’s default and mixed approaches. I then used the agreement score from the GEMsembler “biomass” function to keep components present in at least three models with several exceptions. One exception is acyl carrier protein (ACP_c), which appeared in three of four models, except CarveMe, yet none could produce it in the preliminary analysis; ACP is also not included in the biomass of the BiGG models [98] so I excluded it. For *E. coli*, vitamin B12 (adenosylcobalamin, adocbl_c), core oligosaccharide lipid Al (colipa_c), and phosphatidylethanolamine (dioctadecanoyl, n-C18:0, pe180_c) were removed, as they were included by three of four models but not produced by any. I also reviewed metabolites supported by only one or two models, retaining some low-confidence lipids where they are essential for biomass formation despite inter-model disagreement. Conversely, two lipid metabolites and siroheme in *L. plantarum*, and one lipid in *E. coli*, were removed because their biosynthetic routes were long, non-linear, and present in only a single model, making their presence less likely.

In this case study, my biomass selection was driven primarily by the presence of components in the input GEMs and by whether their production was achievable, without considering experimental validation, stoichiometric refinement, or normalisation, because I was not using the models’ growth rate values and aimed to demonstrate the principle rather than report curated models.

I curated the *L. plantarum* and *E. coli* core3 models to produce all biomass components using the consensus approach as follows. For each precursor that core3 could not produce, I inspected pathway maps from the core2 model or from whichever original model was able to produce the target metabolite, then identified a minimal set of reactions missing in core3 and added them. When the pathway for one biomass component required another component upstream, I curated the upstream component first. During pathway-map inspection, I also added transport reactions for common metabolites that had agreement in fewer than three models. If, after introducing all reactions from the functional pathway, the curated core3 model still failed to produce the metabolite, I compared reaction bounds with those in the original models and adjusted them as needed. For example, in ATP biosynthesis for *L. plantarum*, a comparison of bounds between core3 and the CarveMe model revealed

a constraint on phosphoribosylaminoimidazole carboxylase (AIRCr). Making AIRCr bidirectional restored ATP production in the curated core3 model.

Instead of a manual curation procedure for baseline original models converted to BiGG I performed automatic gap-filling on the minimal PMM5 medium using the CarveMe gap-filling command. The numbers of gap-filled reactions added by CarveMe were 8, 0, 3, and 27 for the AGORA, CarveMe, gapseq, and ModelSEED *L. plantarum* models, and 0, 1, 1, and 19 for the corresponding *E. coli* models. Note that the *L. plantarum* CarveMe model and the *E. coli* AGORA model required no gap filling because they already grew in the tested medium. For *E. coli*, I used the biomass components from the supermodel's "assembly" field. For *L. plantarum*, I instead took biomass components from each original model and then modified them according to the curation procedure described above. This divergence was necessary because gap-filling was not feasible when the *L. plantarum* biomass reaction was taken from the "assembly" level.

### 2.2.7 Comparison with the gold-standard models for *E. coli* and *L. plantarum*

As gold-standard references, I used the original *Lactiplantibacillus plantarum* model iLP728 [99] and, for *Escherichia coli*, the iML1515/iML1515a models adapted by Bernstein et al. for the BW25113 strain by modifying the original MG1655-based iML1515 to account for strain differences [77, 100]. To assess basic similarity to these references in terms of overall reaction lists, I computed recall and precision for each evaluated model. At the reaction level, I formed the set intersection between reaction IDs in the gold-standard model and in the evaluated model; recall was calculated as the intersection divided by the total number of reactions in the gold standard, and precision was calculated as the intersection divided by the total number of reactions in the evaluated model. At the gene level, I applied the same definitions but restricted comparisons to genes associated with the same reaction in both models (i.e., each model's GPR contains the gene for that specific reaction, but GPRs do not have to be identical). Consequently, genes present in both models but linked to different reactions were excluded from the gene intersection set for recall and precision calculations.

To assess model similarity in a low-dimensional space, a postdoc in my group, Sara Benito-Vaquerizo, carried out principal component analysis (PCA) on reaction–presence matrices using the PCA implementation in the scikit-learn package [101] and visualised the models' projections onto the first two principal components. Two scopes were analysed separately, before and after gap filling/curation: first, the non-gap-filled GEMsembler-converted original GEMs, the consensus models, and the gold-standard

model; second, the gap-filled GEMsembler-converted original GEMs, the curated Core3 consensus model, and the gold-standard model.

### 2.2.8 Auxotrophy prediction for *L. plantarum*

For auxotrophy analysis in *L. plantarum*, I classified the experimental growth outcomes in CDPM medium lacking each of the 35 components reported in [75] into three categories: growth ($OD_{600} \geq 1$), no growth ($OD_{600} \leq 0.1$), and reduced growth ($0.1 < OD_{600} < 1$). For three metabolites, I adjusted these labels based on a previously published auxotrophy study [99]. Specifically, growth without isoleucine was set to zero ($OD_{600} = 0$ rather than 0.2), as the authors reported that *L. plantarum* cannot synthesise isoleucine and the observed growth was due to trace carryover in the medium [99]. Growth without phenylalanine and without tyrosine was set to the reduced-growth category ($OD_{600} = 0.2$ rather than 0.1), in line with the reported visible growth under these conditions [99].

To simulate auxotrophies, CDPM medium was used as follows:

CDPM = [”pi”, ”glc__D”, ”na1”, ”ac”, ”nh4”, ”cit”, ”ascb__L”, ”ala__L”, ”arg__L”, ”asp__L”, ”cys__L”, ”glu__L”, ”gly”, ”his__L”, ”ile__L”, ”leu__L”, ”lys__L”, ”met__L”, ”phe__L”, ”pro__L”, ”ser__L”, ”thr__L”, ”trp__L”, ”tyr__L”, ”val__L”, ”lipoate”, ”btn”, ”nac”, ”pnto__R”, ”4abz”, ”pydam”, ”pydxn”, ”ribflv”, ”thm”, ”adocbl”, ”ade”, ”gua”, ”ins”, ”xan”, ”orot”, ”thymd”, ”ura”, ”mg2”, ”cl”, ”ca2”, ”mn2”, ”fe3”, ”fe2”, ”zn2”, ”so4”, ”cobalt2”, ”cu”, ”mobd”, ”k”, ”cu2”, ”h2o”, ”h”]

For each GEM, I predicted auxotrophies by running FBA to maximise biomass production under each condition in which one of the tested CDPM nutrients was removed. Predicted outcomes were classified as growth or no growth using absolute or relative thresholds. Growth was assigned if the growth rate was at least 1 $h^{-1}$ (absolute) or at least 0.85 of the model’s maximum growth rate in unmodified CDPM (relative). No growth was assigned if the rate was at most 0.001 $h^{-1}$ or at most 0.15 of the corresponding maximum in unmodified CDPM. No cases of reduced growth were predicted for any model in any condition. Finally, I compared these predicted categories with those derived from the experimental data.

### 2.2.9 Gene essentiality prediction for *E. coli*

To evaluate gene essentiality prediction, I used the KEIO collection fitness dataset for *E. coli* BW25113 (https://fit.genomics.lbl.gov/cgi-bin/org.cgi?orgId=Keio) [102, 103], comprising gene knockout mutants grown on minimal media with diverse carbon and nitrogen sources. This dataset had previously been assembled in a study to assess four curated MG1655 models that were adapted to BW25113; in that work, iML1515 was further manually turned for improving essentiality predictions, yielding iML1515a [77]. For each

gene and condition, the dataset reports a fitness defect as the $\log_2$ fold change in growth relative to the wild type; measurements are available for 3,789 genes across 28 carbon sources and 13 nitrogen sources.

Because I reconstructed models directly from the *E. coli* BW25113 genome, no additional strain-adjustment was required. I therefore followed the same evaluation pipeline and predicted essentiality with COBRApy's single_gene_deletion for each tested model: the four original GEMs, the curated core3 consensus GEM and the gold-standard iML1515, and iML1515a. I counted a gene as essential if the predicted growth rate with its knockdown was less than 0.001 $h^{-1}$, and non-essential otherwise. I then ranked genes by increasing experimental fitness defect and computed precision–recall curves and the corresponding area (AUCPR) using scikit-learn package. The set of genes evaluated varies per model, depending on overlap with the experimental gene list. To ensure comparability across models, I restricted analysis to conditions on which all models grew: resulting in 15 of the 28 carbon sources and all 13 nitrogen sources. For counting improved predictions relative to the experimental benchmark, I set a fitness-defect threshold of $\log_2$ fold change to −2, as in the prior study [77].

### 2.2.10 Improving gene essentiality prediction

In order to improve gene essentiality predictions I implemented two strategies for GPR rules modification using GEMsembler supermodel (Fig. 2.2).

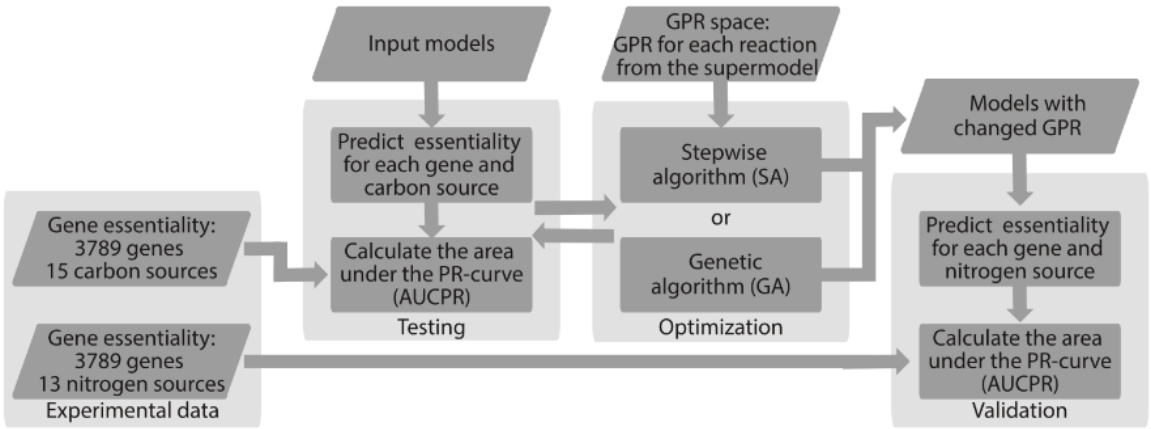


**Figure 2.2.** Schematic representation of the GPR modification to improve gene essentiality predictions by each model. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

The first strategy, a stepwise GPR-combination algorithm (SA (Fig. 2.2), enriches the target model's GPRs by replacing them with alternatives drawn from other models, tested in order of their overall predictive accuracy (AUCPR). I began by identifying reactions whose GPRs include a gene misclassified as essential or non-essential in the target model. For each

such reaction, I substituted the target GPR with the GPR from another model in which the gene was predicted correctly for that reaction. A substitution was retained only if it corrected the selected gene's prediction without degrading predictions for genes that the target model already classified correctly. If not, I tried the next candidate GPR from a different model.

As a second strategy, I implemented a genetic algorithm (GA) (Fig. 2.2) to search agnostically the space of cross-model GPR combinations and maximise AUCPR for gene-essentiality prediction in a given target model. In the GA encoding, each reaction that has alternative GPR variants across the input sources acts as a locus in the chromosome. For each such reaction, the choice of GPR source was encoded as an integer: 0 for the target model itself, 1 for Core3, 2 for ag_EC, 3 for ga_EC, 4 for ca_EC, and 5 for mo_EC. A chromosome is therefore a vector of these integers, one per reaction with alternatives, and the full solution space is the Cartesian product over all such choices. With such formulation I reduced potential space by eliminating cases in which different sources do not yield different GPRs.

I performed the search in this solution space, optimising a fitness function that computes AUCPR, using the PyGAD package with the following settings:

"num_generations" = 50, "num_parents_mating" = 40, "sol_per_pop" = 200, "parent_selection_type" = "tournament", "K_tournament" = 20, "keep_elitism" = 5, "crossover_type" = "two_points", "mutation_type" = "random", "mutation_by_replacement" = true, "mutation_probability" = 0.05, "random_seed" = 42, "parallel_processing" = 100.

A bioinformatician in my group, Bartosz J. Bartmański, assisted with running the GA on a high-performance cluster and resolved numerical-accuracy failures observed with the GLPK solver in COBRApy's "single_gene_deletion" function by switching to the CPLEX solver.

To limit unnecessary GPR changes, I post-processed GA outputs by intersecting solutions across generations and retaining only those reaction-level choices on which the last N generations agreed. Starting with N equal to 1 (the 50th generation), I progressively tightened this consensus by adding one earlier generation at a time and re-evaluating AUCPR (e.g., 50th and 49th agree; 50th, 49th, and 48th agree, etc.). The procedure stopped when the AUCPR of the consensus solution fell below the 50th-generation AUCPR, rounded down to two decimal places.

## 2.3 Results

### 2.3.1 Introduction to GEMsembler, generating cross-tool consensus models

GEMsembler provides a unified procedure to harmonise, combine, and analyse heterogeneous genome-scale metabolic reconstructions. In brief, it assembles GEMs through four main steps: (i) conversion of input-model features (metabolites, reactions, genes) into a unified nomenclature, (ii) assembling of the converted models into a single object, hereafter the supermodel, (iii) comparison of the input models and generation of consensus models comprising different combinations of input features, and (iv) downstream analysis of the resulting consensus models (Fig. 2.1, 2.3A).

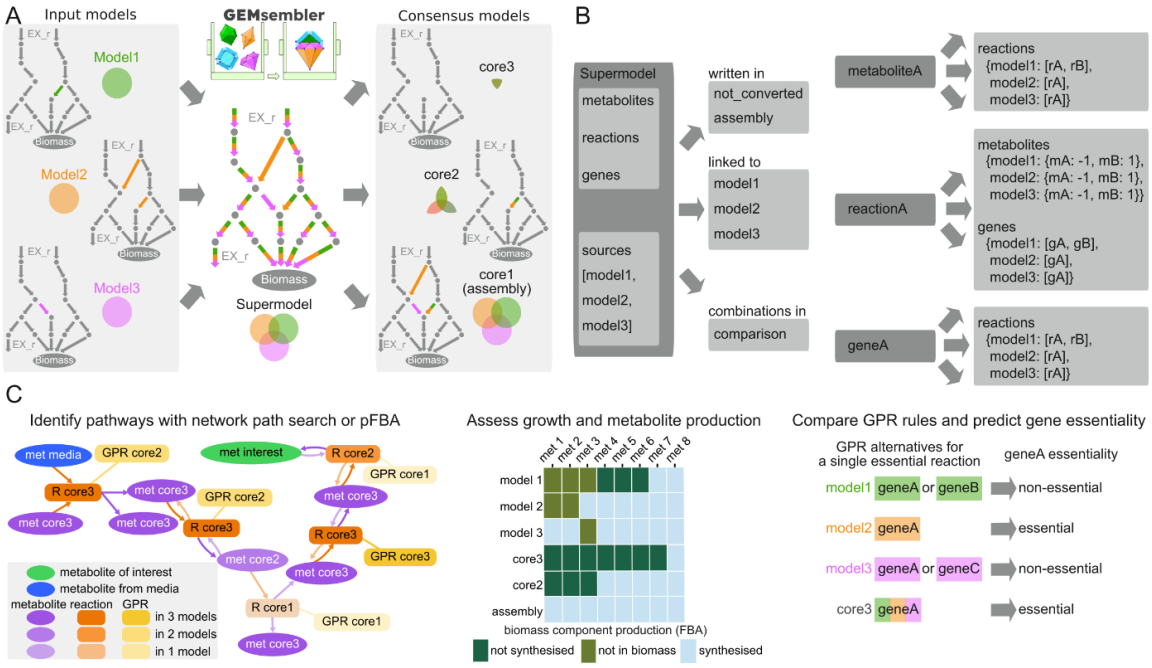


**Figure 2.3.** GEMsembler overview. (A) Schematic workflow with assembling supermodel and deriving confidence-stratified consensus models. (B) Supermodel layout mirroring the COBRApy model class, but extended with conversion details and source tracking for all features. (C) Major examples of downstream analyses: pathway visualisation (left), growth via metabolite production (centre), and assessing gene essentiality from GPR combinations (right). This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

GEMsembler first maps metabolite identifiers from the input models to BiGG IDs [33]using multiple cross-referencing sources. The resulting mappings are then used to rewrite reaction equations in the BiGG nomenclature, preserving the original network topology (Fig. 2.1). When genome sequences are supplied, genes from the input models are mapped to the locus tags or user-defined gene names of a selected output genome via

BLAST [93](Fig. 2.1). At each stage, GEMsembler records intermediate states, enabling inspection and troubleshooting of the conversion process.

GEMsembler then assembles the converted models into a single supermodel (Fig. 2.1, 2.3A). The supermodel mirrors the COBRApy class structure [40]and adds fields that record each converted element of the network (metabolite, reaction, gene) together with its source provenance (Fig. 2.1, 2.3B). Elements that could not be mapped are stored separately in a "not_converted" field. At the moment of creation, the supermodel contains only the union of input features, termed the "assembly", which includes everything present in at least one model. All other combinations of input GEMs, termed consensus models, can be generated at the next step with supermodel comparison functionalities. For example, "coreX" consensus models retain features present in at least X input models, so the assembly equals core1. Feature confidence is defined as the number of input models that include the feature (metabolite, reaction or gene), and attributes of each feature in consensus models follow the same agreement principle. For instance, if a reaction is unidirectional in three of four input GEMs and bidirectional in one, it is unidirectional in core4, core3, and core2, and bidirectional in the assembly. GPR attributes are reconciled by comparing the logical gene expressions from the original GEMs to construct new GPRs for the outputs. Consensus models are stored within the supermodel and can be exported as standalone SBML models for downstream analyses with COBRA tools, including flux balance analysis and gene essentiality prediction.

### 2.3.2 Investigating the structure and functions of the metabolic network with GEMsembler consensus models

The GEMsembler supermodel comprises metabolite, reaction, and gene objects that can be inspected interactively and that summarise agreement across the original models for each network feature, supporting assessment of the network uncertainties and highlighting gaps for potential experimental validation. Given the size of GEMs, pinpointing subnetworks or features of interest can be difficult. To aid this, I integrated additional functionality into GEMsembler (Fig. 2.3C) to systematically explore metabolic capabilities, assign confidence levels to network structure via consensus models, and identify subnetworks of interest.

GEMsembler supports exploration of network structure by offering a neighbourhood search that lists all reactions within a user-defined distance of a metabolite of interest, and by providing three ways to define a "pathway" as a proxy for a given metabolic capability whose confidence can be evaluated. First, a pathway can be user-defined, allowing investigation of expected functions of interest; central carbon metabolism pathways: glycolysis, the pentose phosphate pathway, and the tricarboxylic acid cycle, are already implemented in GEMsembler. Second, a pathway can reflect the capability to produce a

target metabolite and can therefore be inferred topologically by enumerating feasible routes from the supplied medium to that metabolite using the integrated MetQuest package [94], yielding routes that resemble classical pathways in KEGG [104]or MetaCyc [57]. These topology-derived routes are not guaranteed to carry flux, since no flux analysis is performed at this stage. Third, biosynthetic pathways can be derived by parsimonious flux balance analysis (pFBA) [38], which maximises production of the metabolite of interest while selecting a minimal-flux solution. For any of the three definitions, pathway confidence is quantified by agreement scores across input GEMs for the constituent metabolites, reactions, and GPR rules. Outputs include a summary table and an interactive pathway map for visual exploration (Fig. 2.3C).

To probe the functional behaviour of GEMs, GEMsembler simulates growth in a specified medium by running FBA with biomass production as the objective. Additionally, it redirects the objective to each biomass component in turn to test whether it can be produced, and, together with pFBA-based pathway identification, allows pinpointing of missing routes when growth is not achieved (Fig. 2.3C). Beyond reaction-network analysis, GEMsembler compares alternative GPR formulations supplied by different input GEMs and assembles their combinations. The resulting per-reaction GPR variants provide testable options and guidance for model curation aimed at improving gene-essentiality predictions (Fig. 2.3C).

### 2.3.3 GEMsembler enables systematic characterisation of uncertainties in GEMs: a use case with *L. plantarum* and *E. coli*

In order to showcase GEMsembler functionality, I focused on two well-characterised bacteria: *Lactiplantibacillus plantarum* WCFS1 (LP), a gram-positive inhabitant of fermented foods and the gastrointestinal tract with multiple auxotrophies, and *Escherichia coli* BW25113 (EC), a gram-negative model organism with extensive gene-essentiality measurements. Both species have established minimal media: PMM5 for *L. plantarum* and M9 for *E. coli*, defining their nutritional requirements. Curated GEMs are available for both species [99, 100], providing gold-standard references against which to evaluate GEMsembler outputs. The workflow proceeded as follows: automatic reconstruction for each organism with three tools (CarveMe, gapseq, ModelSEED) together with downloading the corresponding AGORA models; assembly of supermodels in GEMsembler; assessment of agreement for reactions and GPRs across consensus models spanning the union of inputs (assembly) to the intersection of all four inputs (core4); and comparison of these consensus models to the gold-standard reconstructions.

Checking the conversion procedure by comparing the number of metabolites, reactions, and genes in the original GEMs, those incorporated into the supermodel, and those

not converted shows that the majority of features from the four original models were successfully converted and included in the supermodels (Fig. 2.4).

With regard to the basic agreement on network elements (Fig. 2.4), roughly half of all metabolites and reactions appeared in only a single model, and between one quarter and one third of genes were present in GPRs of only one model (LP: 339 of 1186 genes; EC: 663 of 1952 genes). Full agreement across all four models was uncommon, not exceeding one quarter of metabolites, reactions, or genes. For reactions with GPRs, a separate GPR-agreement score was computed, which for many reactions was lower than the reaction-level agreement. Overall concordance across model types was, as expected, limited, but higher among the *E. coli* models than among the *L. plantarum* models, reflecting *E. coli*'s status as the best-studied model bacterium (Fig. 2.4).

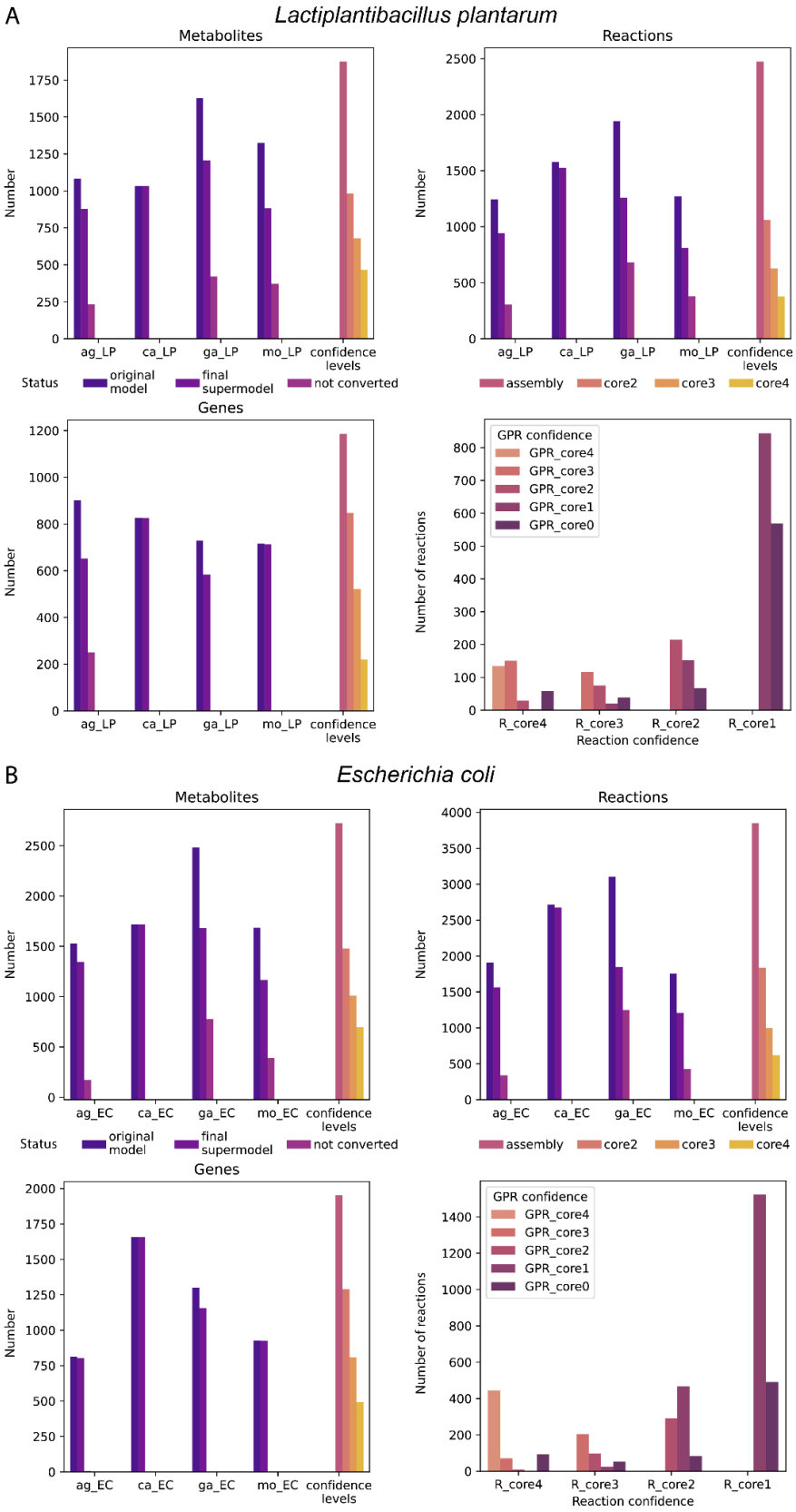


**Figure 2.4.** General characteristics of *L. plantarum* (A) and *E. coli* (B) models, showing agreement across metabolites, reactions, genes, and GPRs. ag: AGORA; ca: CarveMe; ga: gapseq; mo: ModelSEED. CoreX indicates features shared by X models. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

To compare pathway-level agreement, I first examined production routes for central carbon metabolites with available GEMsembler functionality (Fig. 2.5A, B). Topologically feasible biosynthetic routes were enumerated in PMM5 minimal medium for *L. plantarum* and in M9 minimal medium for *E. coli*. For each route, confidence was then evaluated by inspecting reaction- and GPR-level agreement along the path. Central carbon outputs were largely consistent across tools: for *L. plantarum* (Fig. 2.5A), all four models agreed on the ability to produce most of the metabolites (all four original GEMs producing the metabolite) and, in many cases, on the specific routes (the core4 consensus GEM also produces it). At the same time GPR concordance was lower than reaction-level agreement. (Fig. 2.5A). In *E. coli* case (Fig. 2.5B), models showed complete agreement in every case except β-D-glucose 6-phosphate, whose route includes several reactions present in only three of four models. GPRs likewise showed strong consensus: nearly all reactions achieved core4 agreement, with four exceptions: HEX1, EDA, ALKP (appearing in many of the tested routes), and FBA3 (specific to D-fructose 1,6-bisphosphate) – which reached core3 agreement (Fig. 2.5B). Examination of canonically defined glycolysis, the pentose phosphate pathway, and the TCA cycle led to the same conclusion: high agreement in glycolysis and the pentose phosphate pathway and the most discrepancies in the TCA. The discrepancies occur primarily at the GPR level in *E. coli* and, for *L. plantarum*, with many of the TCA cycle reactions absent or represented with large inconsistencies.

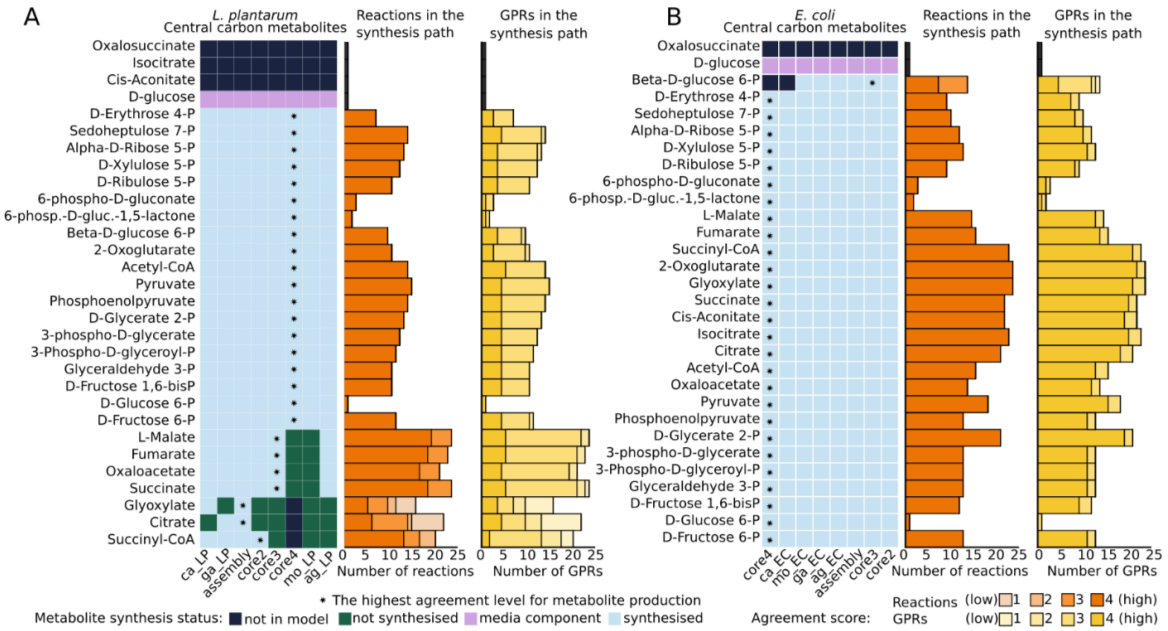


**Figure 2.5.** Central carbon metabolite production determined by the network topology (left panels) in converted original models and in consensus models, and agreement scores for reactions and GPRs (right panels) in the corresponding highest-confidence pathways (marked with an asterisk) for *L. plantarum* (A) and *E. coli* (B). This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

I then applied the same topology-based analysis to biomass components. Because the biomass reaction differs across models, I took the union of components to obtain a comprehensive set: 76 reactants across both species (Appendix I, II). Since medium nutrients and cofactors are required as starting points for the topology search, biomass components belonging to these categories were excluded from pathway analysis and marked separately (Fig. 2.6, 2.7). Compared with central carbon metabolites, producibility of biomass components showed markedly lower confidence in both organisms, though the *E. coli* models still exhibited higher agreement. For *L. plantarum*, only six components were produced with complete agreement (Fig. 2.6), whereas *E. coli* reached core4 agreement for eleven components (Fig. 2.7).

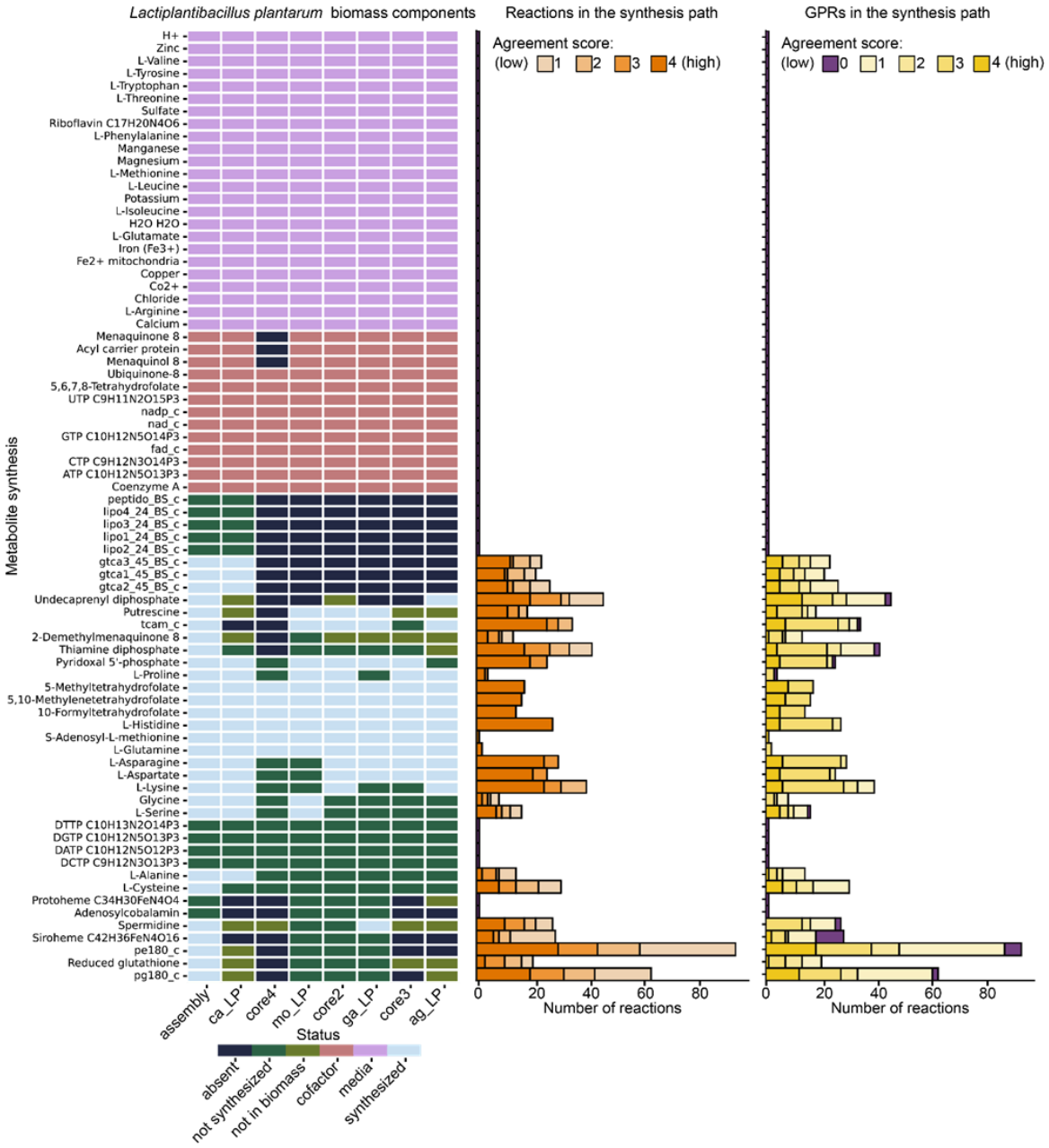


**Figure 2.6.** Biomass components production as determined by the network topology (left panel) in converted original models and in consensus models, and agreement scores for reactions and GPRs (right panels) in the corresponding highest-confidence pathways for *L. plantarum*. ag: AGORA, ca: CarveMe, ga: gapseq, mo: modelSEED. CoreX corresponds to the agreement of X models. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

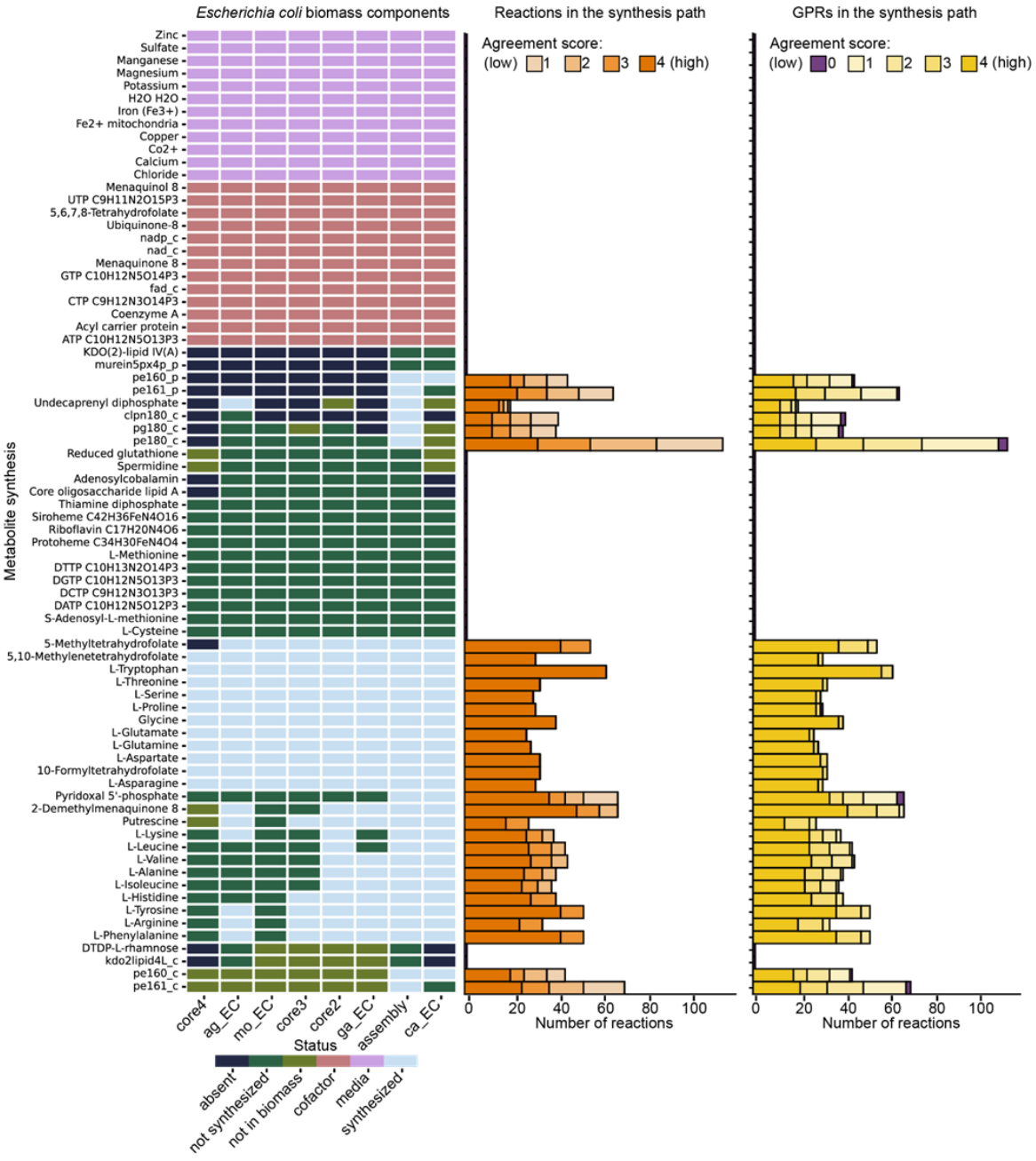


**Figure 2.7.** Biomass components production as determined by the network topology (left panel) in converted original models and in consensus models, and agreement scores for reactions and GPRs (right panels) in the corresponding highest-confidence pathways for *E. coli*. ag: AGORA, ca: CarveMe, ga: gapseq, mo: modelSEED. CoreX corresponds to the agreement of X models. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

Summarising the topology analysis highlights specific regions of uncertainty in the metabolic networks and motivates a stepwise follow-up, enabling a zoom-in from the global network to a handful of pathways and even individual reactions (Fig. 2.8A). First, the metabolites of interest (30 central-carbon compounds and 76 biomass components) were partitioned into “regular” metabolites versus categories that serve as inputs or enabling factors for the analysis (nutrients, other medium components, and cofactors), thereby

removing metabolites whose biosynthesis is either irrelevant or not assessable with a topology-only approach. Second, for the regular metabolites, an agreement score was computed across models and entries below core2 were excluded; this stage filters out metabolites whose presence is unlikely, making their biosynthesis less informative to investigate. This yielded 51 metabolites for *L. plantarum* and 60 for *E. coli*, which were then examined for the presence of production pathways. Third, metabolites showing complete agreement across input GEMs on either their biosynthetic pathway or its absence were set aside, as they carry high certainty and thus low priority for further analysis. Finally, metabolites lacking complete agreement on their biosynthetic routes were flagged for manual inspection (Fig. 2.8A).

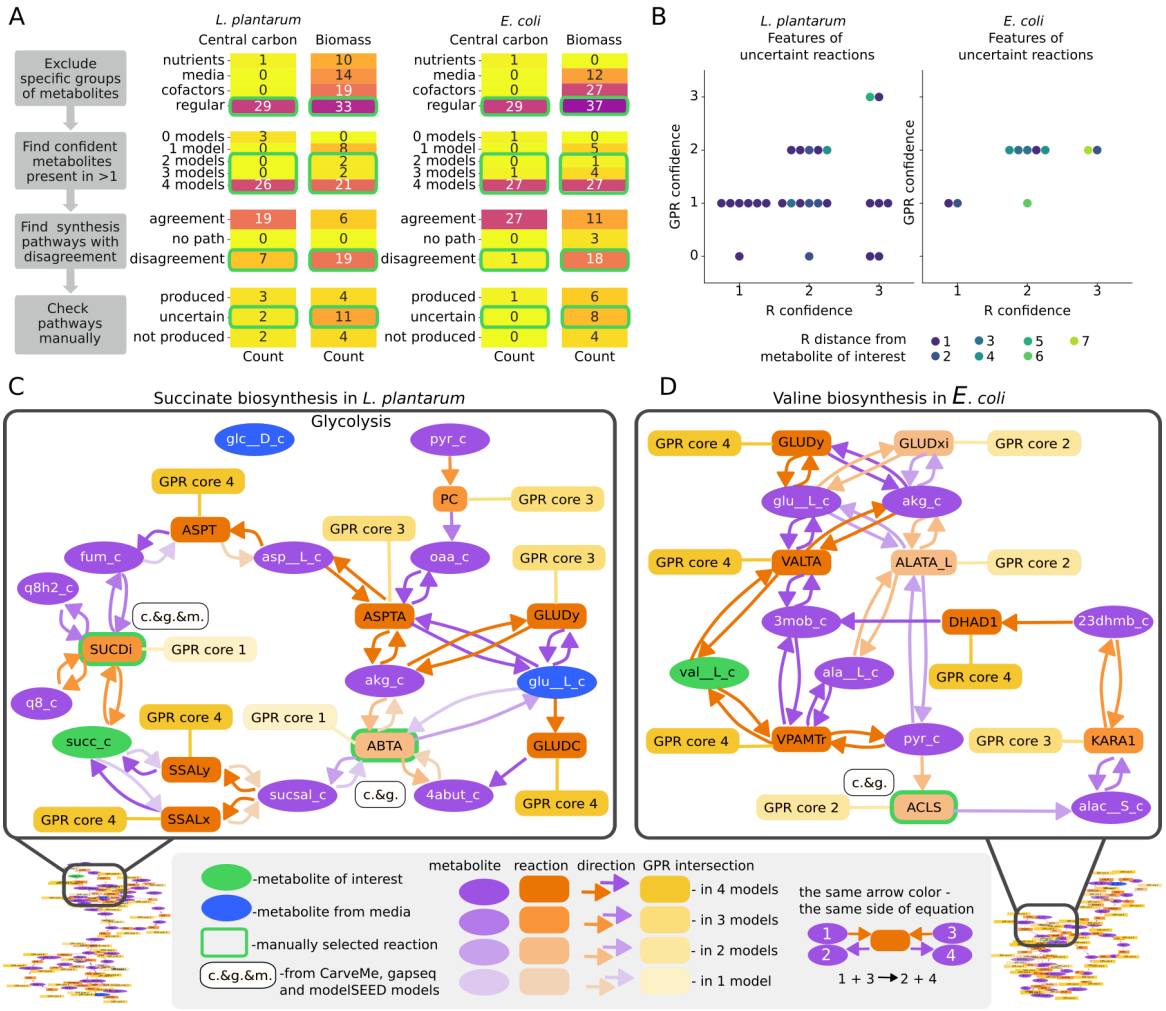


**Figure 2.8.** Mapping uncertainty in topology analyses of metabolic networks with pathway examples. (A) Overview of the step-wise prioritisation scheme: a left-panel decision cascade narrows the search from the full network to candidate pathways/reactions; the right panel reports outcomes for central carbon and biomass components sets in *L. plantarum* and *E. coli*. (B) Profile of reactions prioritised in panel C, summarising per-reaction agreement, GPR-rule agreement, and network distance from the target metabolite of the biosynthesis pathway. (C) *L. plantarum* example: ambiguous succinate synthesis driven by SUCDi and ABTA reactions. (D) *E. coli* example: uncertain valine synthesis driven by ACLS reaction. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

At the final stage, I inspected biosynthetic routes for 7 central carbon metabolites and 19 biomass components in *L. plantarum*, and for 1 central carbon metabolite and 18 biomass components in *E. coli*, using GEMsembler's interactive maps. Each case was assigned to one of three outcome classes: produced, not produced, or uncertain, reflecting the inferred biosynthesis (Fig. 2.8A). Pathways labelled uncertain contain reactions required for the biosynthesis of the target metabolite whose gene-level annotations disagree across at least two models. Such disagreement highlights these routes as areas of highest uncertainty in the network, so they should be prioritised for further investigation. In *L. plantarum*, the uncertain set includes two central-carbon metabolites (succinate and succinyl-CoA) and 11

biomass components; in *E. coli*, 8 biomass components (and no central carbon metabolites) fall into this category (Fig. 2.8A). Each uncertain pathway harbours at least one reaction responsible for the ambiguity, yielding 26 uncertain reactions in *L. plantarum* and 10 in *E. coli* (Fig. 2.8B). In *L. plantarum*, most of these reactions are supported by two models but have a GPR present in only one; in *E. coli*, confidence is slightly higher, with GPRs typically present in two models. The majority of uncertain reactions lie at the product node or one step upstream, although some occur as far as seven steps away along the biosynthetic route (Fig. 2.8B).

GEMsembler's interactive maps enable visual inspection of candidate routes and localisation of the reactions that drive uncertainty. For *L. plantarum* succinate synthesis, two alternative routes explain the ambiguity: one via succinate dehydrogenase (SUCDi) and another via 4-aminobutyrate transaminase (ABTA), with GPR support for these branches coming from only a single model (Fig. 2.8C). In *E. coli* valine synthesis, three reactions appear in only two models: ACLS, GLUDxi, and ALATA_L (Fig. 2.8D). The glutamate dehydrogenase reaction GLUDxi duplicates the role of the better-supported GLUDy but uses NAD instead of NADP, so it does not affect valine production; given that *E. coli* glutamate dehydrogenase is NADP-specific [105, 106], GLUDxi is likely an artefact of automatic reconstruction. The L-alanine transaminase reaction ALATA_L is essential for alanine synthesis and was therefore flagged within the alanine pathway; reports of multiple alanine transaminases in *E. coli* [107]suggest this reaction is plausibly present. The acetolactate synthase reaction ACLS is directly required for valine formation and has been identified in *E. coli* [108, 109], so it should be retained. In this way, the uncertain reactions identified can either be resolved using existing knowledge, where available, or serve as candidates for further experimental verification.

In this section, I showed how GEMsembler identifies uncertain regions of the network and characterises confidence across metabolites, reactions, and GPRs. By combining consensus-based agreement scores with topology-derived pathways, the workflow narrows broad uncertainty to a focused set of pathways and reactions for follow-up.

### 2.3.4 Curation of GEMs with GEMsembler to reproduce growth phenotypes

A principal step in GEM curation is to ensure the organism's metabolic capabilities by reproducing growth phenotypes via classical FBA simulations, and consensus models generated with GEMsembler provide a basis for this process. Both *L. plantarum* and *E. coli* grow in defined minimal media (PMM5 and M9, respectively), which implies that all other required metabolites must be synthesised under those constraints. I used this information to guide curation under the consensus approach.

GEMsembler's growth-analysis module applies FBA to test the producibility of each biomass component, thereby explaining growth failures under the given conditions. In this section, I define biosynthesis in FBA terms rather than topology terms to ensure relevance to growth simulations and to account for transport and cofactor usage. Preliminary simulations on mixed original and consensus models indicated that none of the models achieved growth, as at least one required biomass component could not be produced in each case.

To curate the models of *L. plantarum* and *E. coli*, I first revised the biomass reaction and then ensured that all biomass components were producible in FBA simulations. Using the agreement score from GEMsembler's "biomass" function, I retained components present in at least three models (Appendix I, II). I also kept several metabolites supported by only one or two models when their synthesis was feasible in the corresponding reconstructions (see Materials and Methods). In total, the curated biomass comprised 58 components for *L. plantarum* and 61 for *E. coli* (Appendix I, II). This approach does not replace experimental curation, but it provides criteria for deciding on biomass composition when experimental data are unavailable.

After revising the biomass reaction under the consensus approach, I re-ran FBA across all models using GEMsembler's functionality to test for growth. All models, except the assembly, failed to grow because at least one biomass component remained non-producible (Fig. 2.9A, B). Nevertheless since each biomass component is produced by at least one input GEM, combining different GEMs provides an opportunity to restore the required functions in any model. Complete core4 agreement on biomass-precursor biosynthesis was low in both organisms (LP: 10/58; EC: 12/61), with *E. coli* again showing higher concordance overall. In *E. coli*, the core2 consensus produced most precursors (Fig. 2.9A), whereas in *L. plantarum* the core2 and core3 consensus GEMs performed similarly and produced fewer than half of the targets (Fig. 2.9B), indicating lower inter-model consistency. Some metabolites, such as cysteine, glutathione, and 2-demethylmenaquinone, were producible in all input models but not in every consensus, reflecting divergent pathway realisations that introduce gaps and yield incomplete routes in the consensus GEM. As a curation strategy, I selected a single consensus model and restored its ability to grow. To balance confidence, complexity, and functionality, I chose the core3 consensus as the base and added a minimal set of reactions from other models with GEMsembler-guided procedure to ensure all biomass precursors could be produced.

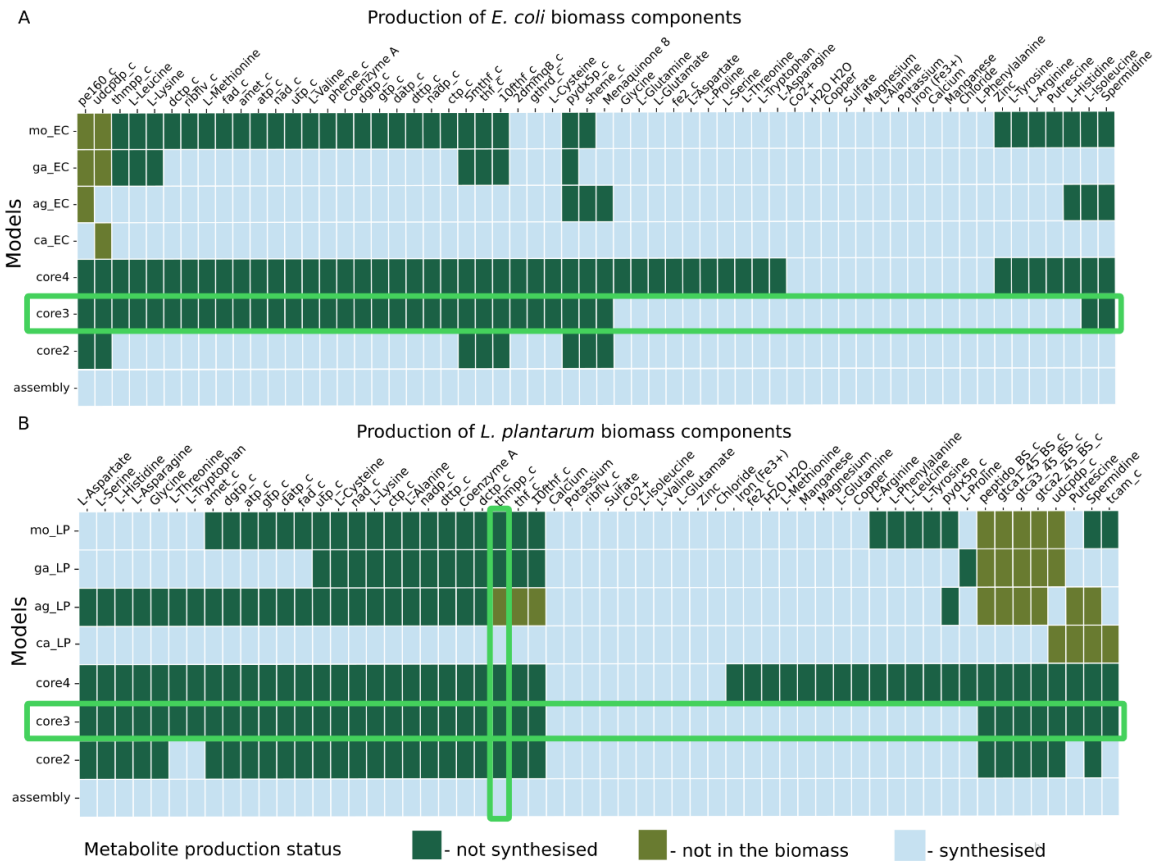


**Figure 2.9.** Production of revised biomass components for E.coli (A) and *L. plantarum* (B) estimated with GEMsembler growth analysis functionality based on FBA simulations. Green horizontal rectangles highlight the consensus models selected for further curation; vertical rectangle in B highlights metabolite for which the curation process is exemplified in Figure 2.10. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

Alongside the summary of biomass component production, GEMsembler's growth-analysis function automatically derives biosynthesis pathways for all biomass components that are producible in each GEM using pFBA simulations, and it generates corresponding tables and interactive maps. These outputs provide the information needed to identify which reactions should be added during curation. Accordingly, I inspected the interactive maps of pFBA-derived routes for components that the core3 model fails to produce but at least one other model can. Along these routes in the producing models, reactions with confidence below core3 are candidates for restoring the target metabolite's biosynthesis, even when they lie several steps away. Not all such reactions are necessary, therefore indiscriminately adding them would introduce unnecessary uncertainty, so I used visual examination of the maps to make an initial selection of curated reactions. Restoration of each pathway was then confirmed by rerunning FBA.

As an example, in *L. plantarum* the thiamine diphosphate pathway contains two to three low-confidence reactions in three segments that were curated (Fig. 2.10). Highly connected

precursors, such as ATP or NAD, pose an additional challenge: adding reactions may be insufficient and reaction bounds may require adjustment. In *L. plantarum*, only the CarveMe model produced ATP, owing to the bidirectional setting of phosphoribosylaminoimidazole carboxylase (AIRCr); adopting bidirectionality for AIRCr in the curated core3 model likewise enabled ATP production.

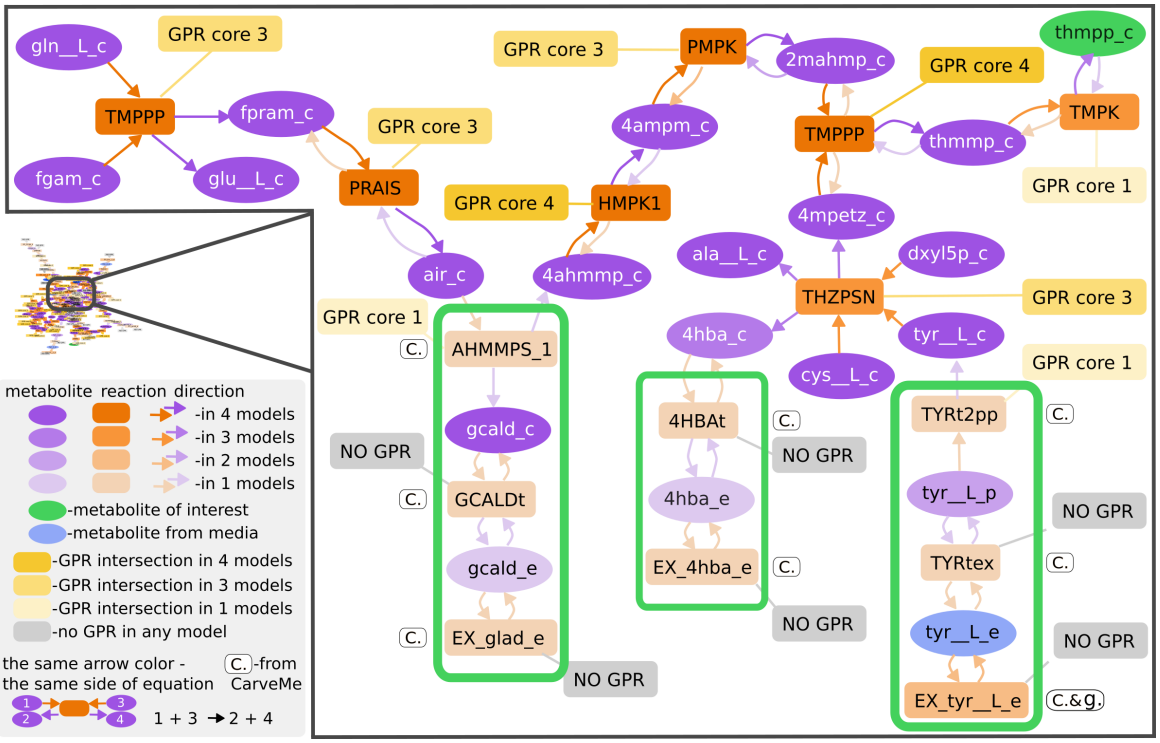


**Figure 2.10.** GEMsembler interactive map derived from CarveMe model for the thiamine diphosphate (tmpp_c) biosynthesis pathway in *L. plantarum*, with curated reactions that restore this function in the core3 model. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

In total, I added 72 reactions to the *L. plantarum* core3 model, including 28 transport and exchange steps, yielding a curated core3 model with 639 metabolites, 729 reactions, and 420 genes. The *E. coli* core3 model was curated analogously: 43 reactions were added, 11 of which were transport or exchange, resulting in a curated core3 model with 943 metabolites, 1217 reactions, and 644 genes.

In this section, I showed how GEMsembler supports semi-automated curation to recover a known growth phenotype. Using the consensus approach and FBA-based checks, I pinpointed missing functions, yielding high-confidence, functional core3 models for *L. plantarum* and *E. coli*.

### 2.3.5 Comparing GEMsembler-curated models to original GEMs and gold-standard models for *L. plantarum* and *E. coli*

After curating the core3 consensus models of *L. plantarum* and *E. coli*, I evaluated them against the original GEMs as a baseline and against previously reported curated models as the gold standards. The curated *L. plantarum* reference iLP728 [55, 99]and the latest curated *E. coli* model iML1515 [77]are already in the BiGG nomenclature. It should be noted that the iML1515 version used here was modified by Bernstein et al. from the original strain MG1655-based iML1515 [100]to reflect differences between MG1655 and BW25113, the strain I used in this study. The four original GEMs (AGORA, CarveMe, gapseq, ModelSEED) were converted to the BiGG namespace with GEMsembler using the mixed approach to preserve native structure while maximising comparability.

With respect to the growth phenotypes in minimal media, among the original GEMs only the CarveMe model for *L. plantarum* and the AGORA model for *E. coli* reproduced growth in PMM5 and M9, respectively. All other models required gap filling in the corresponding medium, which I performed automatically with the CarveMe tool, consistent with automatic draft reconstruction.

To quantify overall similarity of GEMsembler-curated core3 models and the original drafts to the gold standards, I compared reactions and genes (Fig. 2.11, Fig. 2.12). For each model, I computed the intersection of reaction and gene sets with the gold standard and derived precision (intersection divided by the size of the model under consideration), recall (intersection divided by the size of the gold standard), and the F1 score (the harmonic mean of precision and recall).

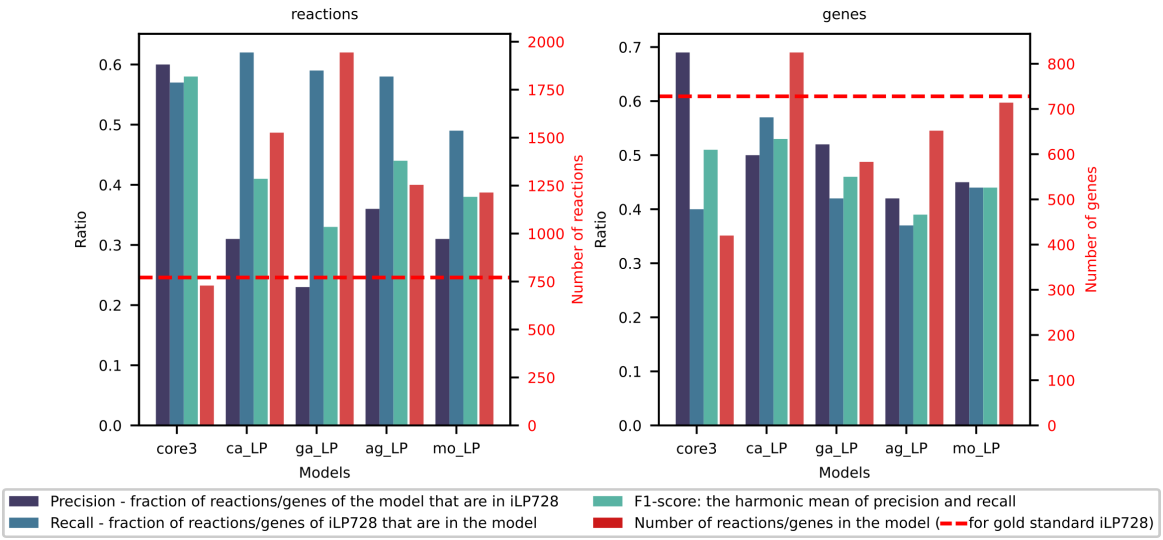


**Figure 2.11.** Comparison of reactions and genes in the GEMsembler-curated core3 model and the four original *L. plantarum* models (AGORA, CarveMe, gapseq, ModelSEED) against the gold-standard iLP728. Bars show set overlaps with iLP728 quantified as precision, recall, and F1 score for reactions and for genes (left axis). Red bars represent total number of reactions and genes in the models (right axis). This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

Relative to the original GEMs, the core3 model recalls a similar fraction of iLP728 reactions and their associated genes while including far fewer features not present in iLP728, yielding the highest reaction-level F1 score (0.58) and the second-highest gene-level F1 score (0.51) (Fig. 2.11). Overall, the GEMsembler-curated core3 model is closer to the iLP728 gold standard than any of the four original models, reflecting the consensus-based reconstruction that reduces unconfirmed elements.

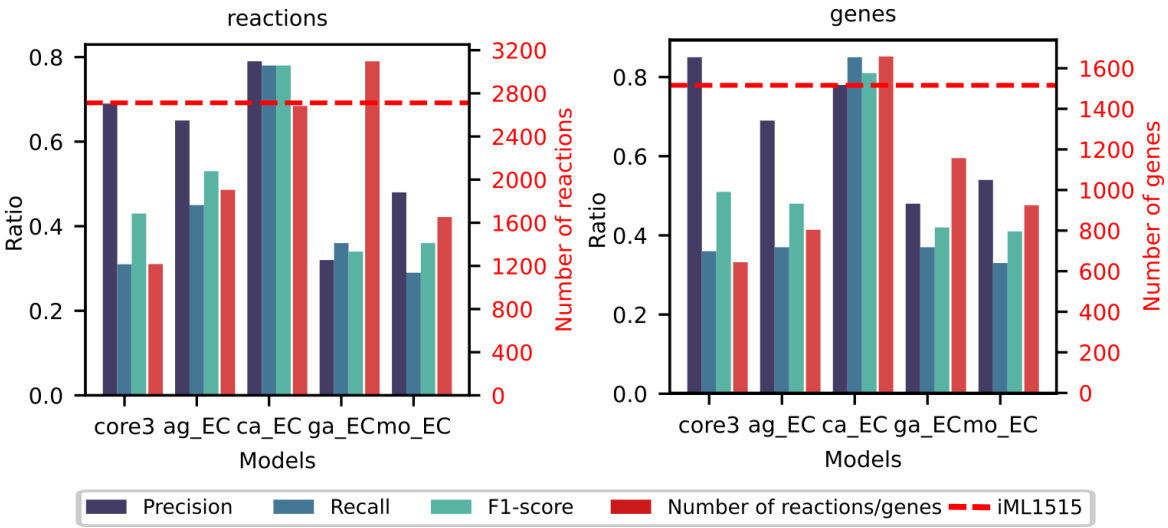


**Figure 2.12.** Comparison of reactions and genes in the GEMsembler-curated core3 model and the four original *E. coli* models (AGORA, CarveMe, gapseq, ModelSEED) against the gold-standard iML1515. Bars show set overlaps with iML1515 quantified as precision, recall, and F1 score for reactions and for genes (left axis). Red bars represent the total number of reactions and genes in the models (right axis). This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

For *E. coli*, the original CarveMe model is the closest to iML1515, which is expected given that CarveMe's universal template draws extensively on reactions from BiGG *E. coli* reconstructions (Fig. 2.12). This *E. coli* model-organism heritage yields substantial overlap with the curated reference. The Core3 consensus, despite containing the fewest reactions and genes, attains the highest gene-level precision and the second-highest reaction-level precision relative to iML1515 (Fig. 2.12), making it the second-closest model to the gold standard.

Principal component analysis (PCA) of reaction–presence matrices (Fig. 2.13) was conducted by a postdoc in my group, Sara Benito-Vaquerizo. The PCA places the GEMsembler-curated *E. coli* Core3 model nearest to iML1515 among all candidates (Fig. 2.13D). For *L. plantarum*, the GEMsembler-curated Core3 model is the closest along the first principal component (Fig. 2.13B).

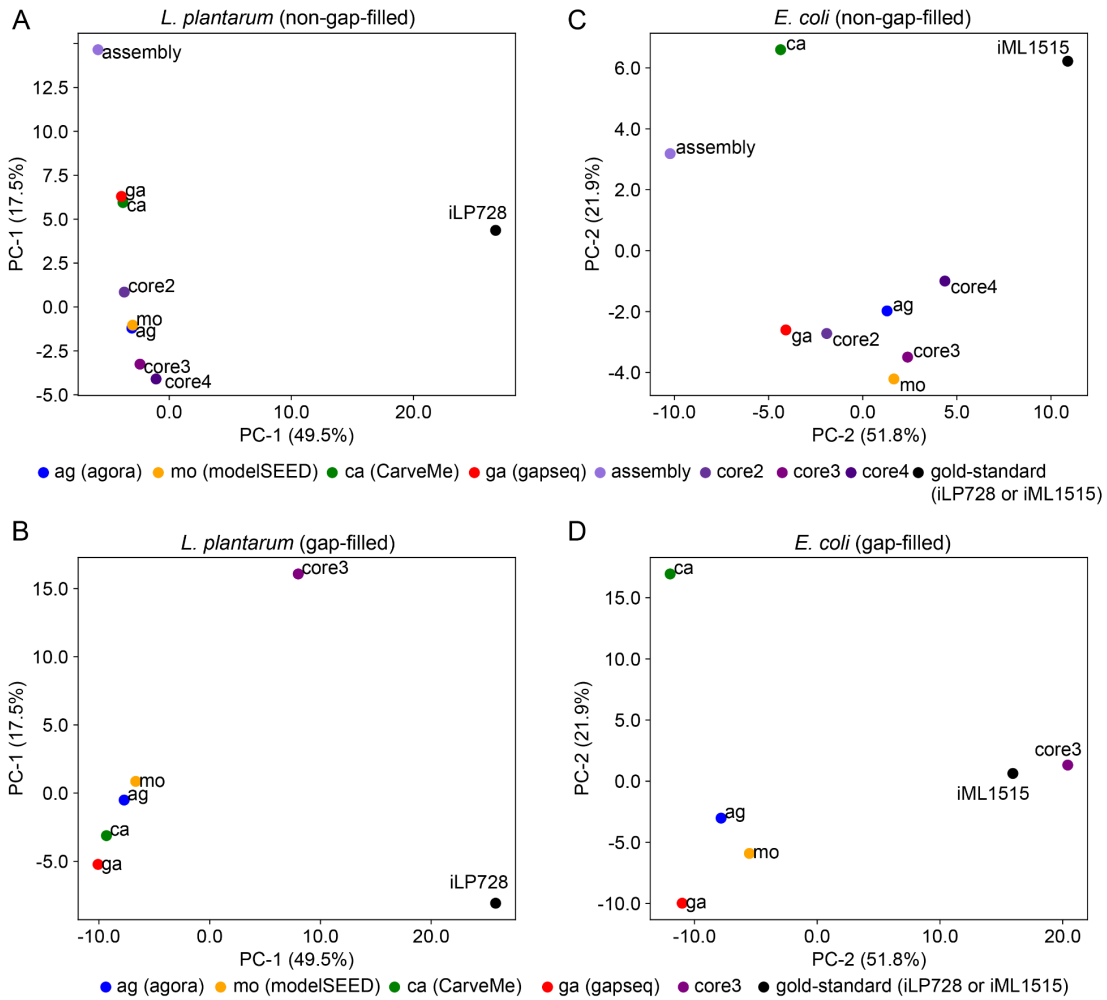


**Figure 2.13.** Principal component analysis of reaction-presence matrices for automatically reconstructed, consensus, and gold-standard models of *L. plantarum* (A, B) and *E. coli* (C, D).
A, C: GEMsembler-converted originals prior to gap filling, consensus models (core3 not curated), and the gold-standard references. B, D: GEMsembler-converted originals after CarveMe gap filling together with the GEMsembler-curated core3 model. Only the core3 consensus is shown in panels B and D, as this was the only consensus model curated for growth. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was the only plot generated not by me, but by a postdoc in my group, Sara Benito-Vaquerizo.

Overall, the consensus approach eliminated large numbers of low-confidence reactions typical of draft reconstructions and produced medium-sized, high-confidence models that more closely match the gold-standard reconstructions.

### 2.3.6 GEMsembler-curated model outperforms the gold-standard *L. plantarum* model in auxotrophy prediction

Following structural analysis of the models, I evaluated their functional performance through auxotrophy prediction. For *L. plantarum* greater similarity to the iLP728 reference does not automatically imply higher quality, since novel gene functions and pathways may have emerged since iLP728 was published. Models were tested for *L. plantarum*

auxotrophies in an alternative medium, CDPM and compared against experimental classifications of growth, no growth, and reduced growth [75,99]. For each model–condition pair, growth was predicted by running FBA in CDPM with one tested nutrient removed. Because no simulations yielded an intermediate outcome, predictions were binned into growth or no growth.

The GEMsembler-curated core3 model performed best overall, exceeding the accuracy of all other models, including iLP728, in two to four conditions. No metabolites were found for which core3 was wrong while another model was correct (Fig. 2.14). Four of the five errors made by core3 corresponded to experimentally observed "reduced growth," where the binary classification necessarily mislabels the outcome. For tryptophan, all models, including core3, predicted growth; the discrepancy likely reflects non-metabolic inhibition by other aromatic amino acids present in the medium [99]. Relative to iLP728, core3 correctly captured four phenotypes: growth in the absence of biotin and pyridoxamine, and no growth in the absence of glutamate and riboflavin (Fig. 2.14).

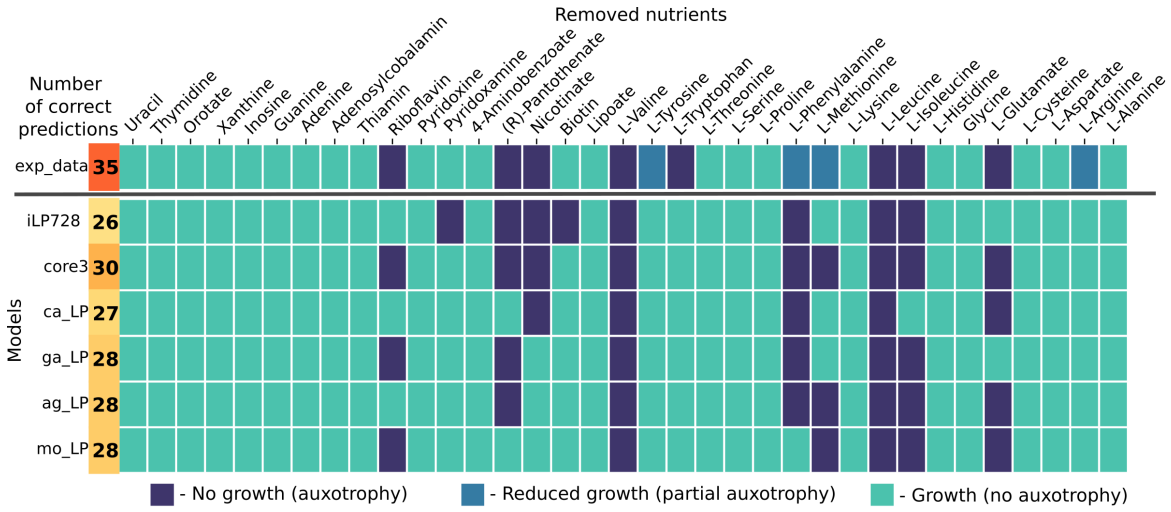


**Figure 2.14.** Auxotrophy predictions in CDPM medium for all models compared with experimental information. Labels: core3, GEMsembler-curated core3 model; ag_LP, AGORA; ca_LP, CarveMe; ga_LP, gapseq; mo_LP, ModelSEED (*L. plantarum* models); exp_data, experimental data. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

GEMsembler's network-structure analysis enables moving beyond mere comparison of simulation outputs to investigate the underlying causes of improvements in specific predictions.

For biotin, the key difference is compositional: iLP728 treats biotin as a biomass component, whereas none of the original reconstructions, and consequently the core3 model, include it in biomass (Appendix I). This change yields the correct prediction of growth without biotin supplementation in *L. plantarum*, but it does not clarify the organism's biosynthesis route to biotin. Given biotin's role in fatty acid synthesis in lactic acid bacteria

[75], the corresponding biosynthetic pathway warrants dedicated follow-up. Riboflavin shows the opposite pattern. The experimentally observed auxotrophy is missed by iLP728 because riboflavin is not part of its biomass reaction. In contrast, core3 and all original models include riboflavin in biomass (Appendix I). Among these, CarveMe and AGORA incorrectly predict its production, while core3, gapseq, and ModelSEED lack a riboflavin biosynthesis route, aligning with auxotrophy.

Two further discrepancies between core3 and iLP728 arise from differences in the pathway structure. For pyridoxamine, iLP728 incorrectly predicts its auxotrophy because the model requires it to make biomass component pyridoxal 5′-phosphate via ALATA_Lr, a reaction without a GPR (interactive map is available with GEMsembler publication as File S10). The core3 model instead uses PYDXS to form pyridoxal 5′-phosphate without pyridoxamine: an interactive map is available with GEMsembler publication on Zenodo and GitLab (Appendix III) as File S11. PYDXS appears in three original models, also without a GPR, but correctly implies that pyridoxamine is non-essential. Finally, glutamate auxotrophy is not predicted by iLP728 because it contains P5CD, which converts proline (present in CDPM) to glutamate (interactive map is available with GEMsembler publication as File S12). P5CD lacks a GPR in iLP728 and is absent from all original models; consequently, it is not present in core3, resulting in the correct prediction of glutamate auxotrophy.

This section shows that models generated with the consensus approach can capture experimental phenotypes and even outperform the curated gold-standard model. The consensus strategy is not a substitute for experimental validation; rather, together with GEMsembler's network-structure exploration, it provides a principled way to generate hypotheses and prioritise validation. As new evidence becomes available, GEMsembler can support an iterative curation loop between the data and the models.

### 2.3.7 Curating GPR rules with GEMsembler improves gene essentiality predictions in *E. coli*

Model quality depends not only on network topology (the reactions) but also on the gene content encoded in GPR rules. These rules can be evaluated via gene-essentiality tests: simulate growth with a given gene knocked out so that its associated reactions carry no flux, then compare predictions to experimental outcomes. Because GPR formulations differ across input reconstructions, even when built from the same genome, essentiality predictions also diverge. I therefore first assessed models performance in this regard. In addition, since GEMsembler aggregates GPRs from all original and consensus models, I investigated whether this pool of alternatives can be leveraged to improve essentiality predictions by recombining GPR rules from different models.

To evaluate functional performance, I compared modelled gene essentiality with experimental fitness defects for 3789 single-gene knockouts measured across 41 minimal-media conditions comprising 28 carbon and 13 nitrogen sources [102, 103]available for *E. coli* BW25113, which motivated selection of this strain for this project.. This dataset was previously used to benchmark four curated *E. coli* models and to produce an adjusted version of iML1515 (iML1515a) with improved essentiality predictions [77]. For comparability, I restricted analysis to the 15 carbon sources and all 13 nitrogen sources on which every model grows.

Functional performance for each model was evaluated as the area under the precision–recall curve (AUCPR) over gene–condition pairs ranked by the experimental growth defect (Fig. 2.15). Among the tested models, the AGORA reconstruction achieved AUCPR = 0.642, surpassed only by the adjusted gold standard iML1515a (AUCPR = 0.754), and exceeding the standard iML1515 (AUCPR = 0.593) (Fig. 2.15). The GEMsembler-curated Core3 model followed with AUCPR = 0.556, whereas the original gapseq, CarveMe, and ModelSEED drafts performed worse, with AUCPR values between 0.3 and 0.5 (Fig. 2.15).

Misclassifications of essential/nonessential genes can stem from either network topology or from GPR specification. To target the latter, I used the consensus approach to revise GPRs in each tested model, including iML1515 and its adjusted variant with one knowledge-driven and another agnostic strategy (see Materials and Methods and Fig. 2.2).

First, I applied a stepwise combination algorithm (SA): for a given model, reactions whose GPRs implicated mispredicted genes were reassigned to GPRs taken from alternative sources in the supermodel (meaning alternative models) that correctly classified those genes, testing candidates in order of decreasing AUCPR (see Materials and Methods and Fig. 2.2). Second, I used a genetic algorithm (GA) to select, for each reaction, one GPR from the pool offered by all models, optimising AUCPR on the carbon-source conditions. After tuning on carbon sources, I evaluated whether these changes also improved predictions on nitrogen sources (see Materials and Methods and Fig. 2.2).

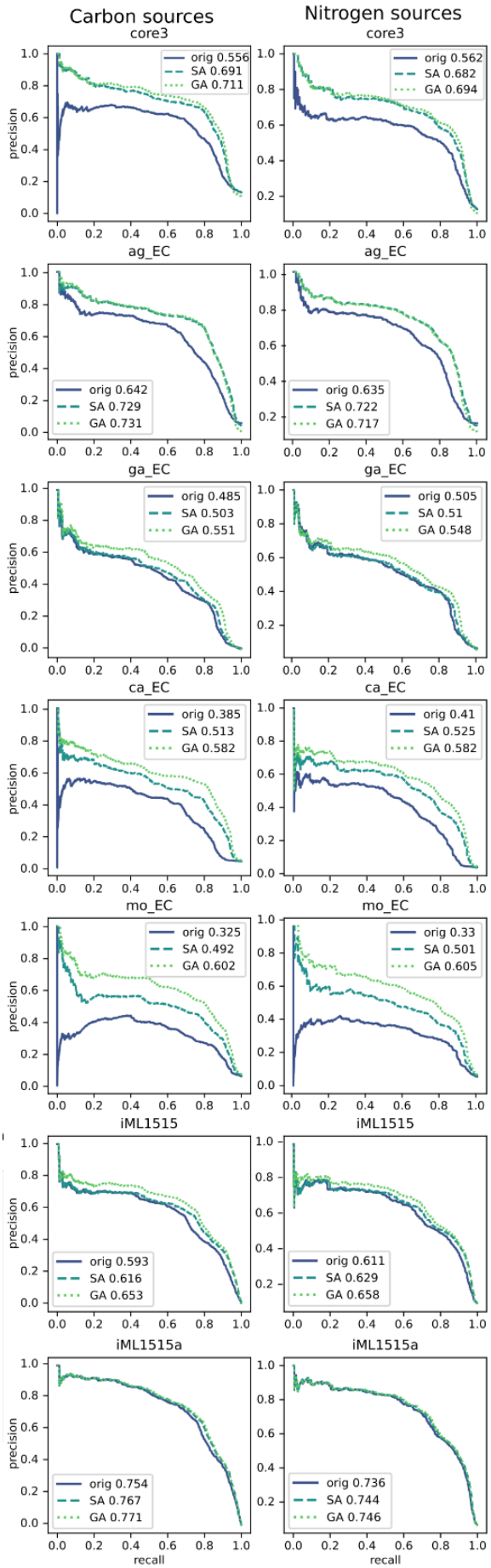


**Figure 2.15.** Combining GPR rules in *E. coli* models improves gene-essentiality predictions across carbon and nitrogen sources. Precision–recall curves and AUCPR are shown for carbon sources (left) and nitrogen sources (right) for the original models (blue, solid) and for models modified with the combination algorithm (turquoise, dashed) or the genetic algorithm (green, dotted). This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

The two GPR-modification strategies differ markedly in the extent of changes they introduce. The stepwise algorithm alters only a few dozen GPRs per model, from as few as three in iML1515a up to thirty-four in iML1515, whereas the genetic algorithm typically rewrites hundreds of GPRs (Fig. 2.16A). These edits also change gene content. The SA affects between one and ten genes per model, while the GA alters between six and 248 genes (Fig. 2.16B). Under SA, gene removals are rare, limited to six genes in iML1515a and one in the gapseq model. The GA likewise tends to add rather than remove genes, with the main exceptions being iML1515a and CarveMe, which already contain the largest gene sets (Fig. 2.16B).

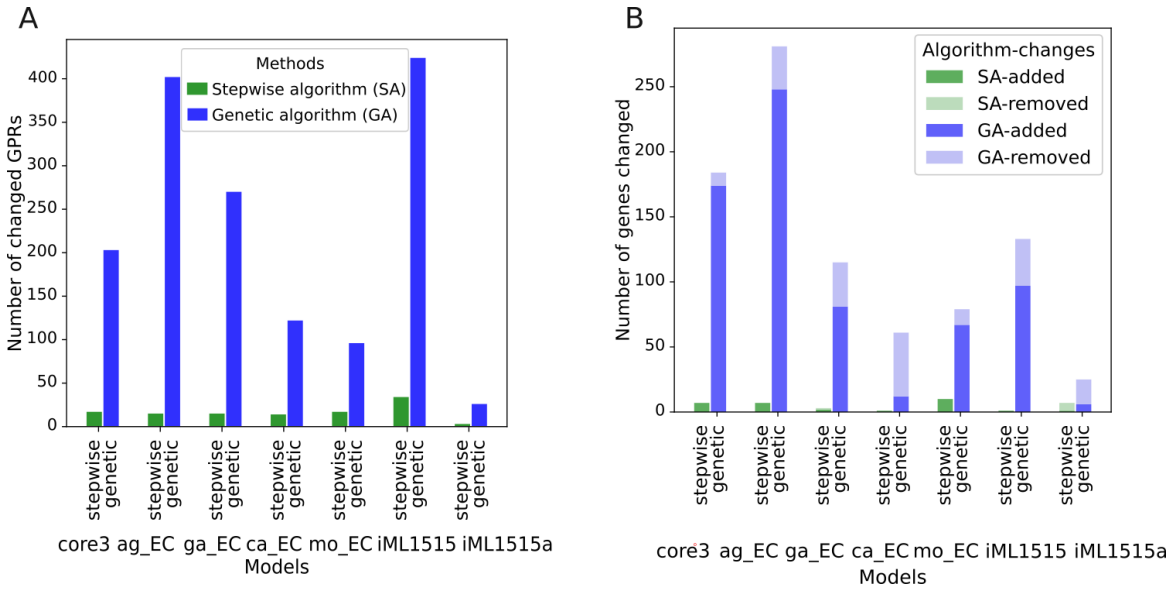


**Figure 2.16.** Curation of GPRs using the stepwise combination algorithm (SA) and the genetic algorithm (GA). (A) Number of GPRs changed in each model by SA and by GA. (B) Number of genes changed in each model by SA and by GA. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

Applying the stepwise combination algorithm (SA) improved gene-essentiality predictions across all models (Fig. 2.15). For the core3 model, AUCPR rose by 13.5% to 0.691, exceeding the gold-standard iML1515. AGORA increased by 8.7% to 0.729. Gapseq gained the smallest improvement 1.8%, while CarveMe and ModelSEED showed the largest relative jumps, +12.8% and +16.7%, reaching AUCPR = 0.513 and 0.492, respectively. Even the gold standards benefited from considering GPRs in the other GEMs: iML1515 reached 0.616 (+2.3%) and the adjusted iML1515a reached 0.767 (+1.3%). Notably, improvement was not one-way from strong to weak models. For example, despite its higher baseline performance, AGORA was still enhanced by combinations with core3, gapseq, CarveMe, and ModelSEED, each of which started with lower original performance.

Using the GA to select an optimal cross-model GPR set yielded even larger improvements in essentiality prediction than using the SA across all analysed models.

The modified core3 and AGORA models reached AUCPR values of 0.711 and 0.731, respectively, again outperforming the gold-standard iML1515, and by an even wider margin (Fig. 2.15). Gapseq improved by 6.6%, while CarveMe and ModelSEED again showed the largest boosts, 19.7% and 27.7%, achieving AUCPR = 0.582 and 0.602, respectively. The original gold-standard iML1515 benefited more considerably, by 6.0%, and the adjusted iML1515a also improved by 1.7% (Fig. 2.15).

Improvements in overall prediction quality were accompanied by more genes falling into the all-correct or mixed categories across the 15 carbon sources and by fewer genes with completely incorrect predictions (Fig. 2.17A). Across the analysed GEMs, between 3 and 15 mispredicted genes were corrected by either the SA or GA procedure, often across multiple conditions (Fig. 2.17B). Both approaches also introduced new genes with correct or mixed predictions: 1 to 8 genes with SA and 6 to 229 genes with GA (Fig. 2.17B). No genes in any model moved into the wrong category as a result of SA or GA. However, there was one new gene added by SA to the ModelSEED model that was mispredicted in all 15 carbon-source conditions.

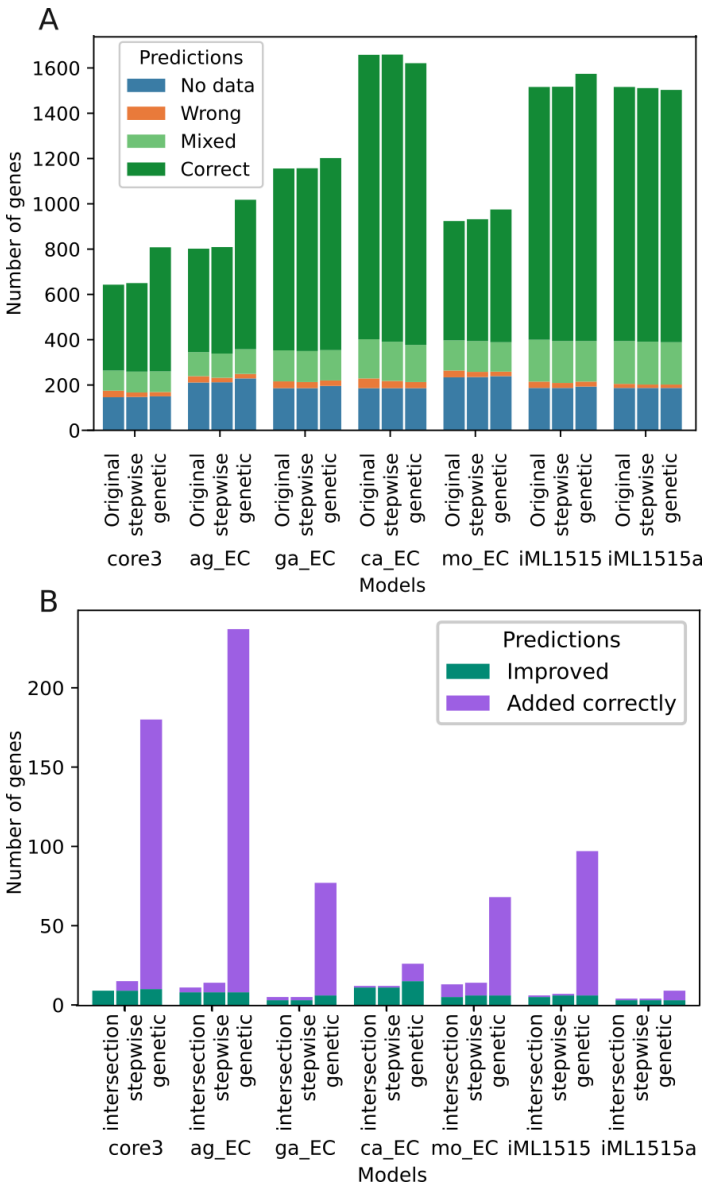


**Figure 2.17.** Genes in *E. coli* models: essentiality predictions and improvements. (A) Number of genes separated by prediction quality status for original and SA/GA-modified models. (B) Number of genes whose essentiality predictions improved after SA or GA, genes newly added with at least partially correct predictions, and the overlap between genes improved by SA and GA. This figure was published in Matveishina et. al., mSystems 2025 [81]; it was made fully by myself.

To understand how performance gains arose in the adjusted iML1515a, the strongest curated model to date, I examined which genes and GPR changes were introduced. Both GA and SA improved predictions for three existing genes (b0131, b0134, b0778) and added one correctly predicted gene (b1593). Each of the three improved genes maps to a single reaction in iML1515a (ASP1DC, MOHMT, and DBTS, respectively).

In the original model these genes were predicted essential, but after GPR revision they became non-essential, matching the experimental dataset. The mechanism was the same in all cases: a single-gene GPR was replaced with a different rule that includes

additional isoenzymes (two or three genes), offering alternative routes. For example, for the dethiobiotin synthase reaction (DBTS), the GPR changed from "b0778" to "b0778 or b1593". Notably, b1593 is annotated in Ensembl [110]as a putative dethiobiotin synthetase and is supported by recent functional experimental assays [111]; it is also correctly predicted as non-essential, consistent with experimental evidence.

Since it is important to test whether the model improvement can be generalised, I tested it on an different set of experimental data. I performed gene essentiality predictions for the same mutants but grown on alternative nitrogen sources (Fig. 2.15). Using the input models and the GPR-modified variants (tuned on the carbon-source data), both the improvement trends and the overall performance levels were recapitulated on the nitrogen-source dataset (Fig. 2.15). This suggests that, although experimental data are needed to guide automatic improvements, the resulting models can transfer those gains to conditions not used during curation.

Overall, GEMsembler's comparison and consensus framework enables merging model features through targeted edits or automated pipelines, improving the predictive performance of both original reconstructions and consensus models.

## 2.4 Discussion

GEMsembler provides a unified, flexible framework for cross-tool comparison of GEMs and for assembling consensus reconstructions at various agreement levels, depending on the required confidence. Cross-tool evaluation is crucial, because the choice of reconstruction pipeline can shape network structure and downstream predictions as much as, or more than, the biological input from the genome itself [24]. By exposing tool-driven variability, offering utilities to navigate large networks, and deriving consensus models from areas of agreement, GEMsembler helps mitigate tool-specific biases.

Beyond comparison, consensus building proves practically useful: models can complement one another and improve predictive performance, as seen for auxotrophy and essentiality benchmarks. Although experimental data remain the arbiter for refinement, GEMsembler provides semi-automatic routes to combine features that best align with observations, thereby accelerating design–build–test–learn cycles. Systematic contrasts of alternative reactions and GPR rules highlight uncertainties and knowledge gaps, guiding experimental prioritisation. The proposed, transparent curation workflow yields consensus models that match or even surpass gold-standard reconstructions for *L. plantarum* and *E. coli*, streamlining the otherwise time-intensive process of model reconstruction and curation while meeting quality standards.

GEMsembler, however, comes with certain limitations. First, it relies on BiGG [33] as the primary biochemical database, which is smaller than broader resources such as

ModelSEED [58]. There are still good reasons to choose BiGG: it is model-oriented, widely adopted, and uses readable identifiers. Nevertheless, the GEMsembler framework is flexible enough to adopt a different primary database if the scientific community needs such change. The second limitation arises from standard conversion issues: some metabolites, reactions, and sometimes even genes cannot be converted. By design, the default supermodel assembly omits non-converted elements to remain conservative about conversion results and for the sake of standardisation, in contrast to the mergem tool [80], which mixes nomenclatures. GEMsembler still mitigates conversion loss issue with a "mixed" supermodel option that retains non-converted elements when they are relevant to the analysis. Finally, GEMsembler requires prior knowledge of the input GEM format, which must be provided by the user. Although six input model types are currently supported, models may follow other formats, some of which may be difficult or impossible to anticipate. The current solution is to allow users to add new formats by adapting the existing conversion classes and pipeline parameters.

Regarding consensus assembly, a central challenge is how to combine specific model attributes, particularly GPR rules. Consider GPR rule "(geneA and geneB) or geneC" in one model versus GPR rule "(geneA and geneD) or geneE" in another. The consensus can be interpreted as "geneA," retaining the shared subclause present in both models, or as empty, requiring intact AND groups because the models disagree on the enzymatically active unit. GEMsembler adopts the former interpretation by default, but provides an "and_as_solid" parameter in the supermodel assembly and comparison functions that handle GPR combination; setting it enforces the stricter interpretation.

A broader issue is that there are many legitimate ways to combine models, and the most appropriate choice depends on the question being asked and the analysis goals; anticipating every use case is challenging. As an example, the current GEMsembler comparison combines model structural elements (metabolites, reactions, genes) at a certain agreement level and evaluates their attributes at the same level. This is sensible in most scenarios, yet it can discard useful information: for instance a reaction may be kept because three models include it, while its GPR is dropped because no GPR overlaps across the same three models. In such cases one might wish to preserve a high-confidence network topology but also retain as much genetic evidence as possible. A more dynamic scheme would decouple structural inclusion from attribute evaluation, for instance keeping the reaction at the chosen agreement level while attaching the best available GPR, defined as the highest non-empty agreement across sources. Exposing this as an option would let users prioritise either strict consistency or richer genetic context, depending on their aims.

Pathway transfer between consensus models raises a related issue. When a route from a lower-agreement consensus is added to a higher-agreement model, reactions are currently

selected via pFBA, following a minimal-length principle for pathway composition. As an alternative, reaction selection could maximise the proportion of reactions in the path at the highest available agreement level, even at the expense of path length, reflecting a most-confidence principle. Further improvements on the GEMsembler roadmap include more nuanced merging of duplicate reactions, exporting richer metabolite and reaction annotations that bridge additional namespaces, and optional user guidance during conversion to resolve ambiguous mappings.

Although the case studies focus on single-species bacterial GEMs, GEMsembler is not limited to this setting. By aligning alternative pathway realisations across models, it can highlight substitute routes that inform strain-engineering strategies to increase target metabolite production or optimise growth [48]. Beyond single strain engineering, GEMsembler can compare models from closely related species or isolates to pinpoint metabolic divergences at the level of pathways, reactions, and genes that may underlie phenotypic differences [23, 49–52]. GEMsembler can also compare distantly related organisms that share a similar metabolic phenotype to reveal recurrent pathway motifs that may explain convergent behaviour [53, 54]. The GEMsembler workflow also extends to microbial communities [10, 55, 56]. Community-level assembly that connects models of its members can reveal overlapping and complementary pathways, suggesting potential cross-feeding, competition, and division-of-labour patterns, and thereby generating hypotheses on interspecies interactions and priorities for experimental design [57, 58].

# Chapter III
# Reconstruction of *B. uniformis* and *P. vulgatus* GEMs with consensus approach and evaluation of their metabolic behaviour

## 3.1 Introduction

The human gut microbiota plays a crucial role in host health and physiology, especially through its involvement in nutrient processing and metabolic regulation [2, 5]. Due to the complexity of the microbiota community, the essential step of clarifying this role is examining the dominant members of the microbiota. In the majority of healthy "western" human intestine communities, the most abundant genus is *Bacteroides*, and the two most abundant species are *Bacteroides uniformis* and *Phocaeicola* (former *Bacteroides*) vulgatus. These species have been shown to play a central role in carbohydrate breakdown, various metabolic and inflammatory responses, drug effects, etc. [5, 112–116].

With respect to carbon source utilisation, *Bacteroides* are able to degrade long-chain polysaccharides, which are intractable for both human enzymes and many other bacteria, due to their unique gene clusters called polysaccharide utilisation loci (PULs) [112, 117]. For example, *B. uniformis* PULs show distinct glycan-degrading functions in the xyloglucan or type-II mucin media and therefore additionally mediate glycan-dependent interactions with butyrate producers in the bacterial community [117]. Another example demonstrates that *B. uniformis* and *P. vulgatus* contain fructan-utilisation loci that allow them to grow on fructose-based carbohydrates in quite minimal media, which contains only L-cystein and hematin/L-histidine as potential additional nutritional compounds [118]. For glucose, fructose and sucrose, the reported doubling time for *B. uniformis* is quite stable: 2 h, 2.1 h, 2.1 h, respectively, while for *P. vulgatus* it varies more: 2.8 h, 3.2 h, 2.4 h, respectively [118].

Depending on the available carbon source, *Bacteroides* demonstrate various metabolic capabilities. *B. uniformis* can utilise gum arabic, mucin, pectin, WBE (wheat bran extract), inulin and has different expression patterns in neurotransmitter GABA (gamma-aminobutyric acid), butyrate and fatty acids biosynthesis pathways in these media [119]. It is also proven to produce more GABA when grown on mucin or pectin instead of glucose [119].

Many *Bacteroides*, including *B. uniformis* and *P. vulgatus*, are associated with different metabolic compounds, crucial for various aspects of host physiology, including bile acid [114], acetate [116], propionate [120], folate [121], butyric- and caproic-acid [115]. They can even share the bile acid biosynthesis pathway with other members of the microbial community, by utilising taurocholic acid with bile salt hydrolase and producing cholic acid, which can be further utilised by *Clostridium* [122].

The abundance and metabolic capabilities of *B. uniformis* and *P. vulgatus* make them key representatives of the gut microbiome, but current knowledge about these bacteria is quite limited compared to model organisms such as *E. coli*. Even within the genus *Bacteroides*, the most studied species is *B. thetaiotaomicron* [112, 123].

The limitation of knowledge about *B. uniformis* and *P. vulgatus* starts with functional annotation. It is striking compared to the most extensively studied bacteria *E. coli* and quite substantial compared to adequately characterised *Bacteroides*, *B. thetaiotaomicron*. Both *B. uniformis* and *P. vulgatus* have reference proteomes in UniProt [124]; the total number of protein entries for *B. uniformis* falls between that of *E. coli* and *B. thetaiotaomicron*, whereas the number of proteins in *P. vulgatus* is lower (Table 3.1). Annotation score of the corresponding entries, which goes from 1 (basic, low confidence) to 5 (high confidence, with experimental evidence), demonstrates how far *B. uniformis*, *P. vulgatus* and *Bacteroides* in general are from the model organism in terms of available annotation, with the vast majority of proteins having only the most basic annotation [125]. *P. vulgatus* does not have any protein in the best annotation category [29], and *B. uniformis* has only one entry in the category above basic [126], while *B. thetaiotaomicron* annotation scores are better (Table 3.1). In regard to genome size, the corresponding NCBI RefSeq assemblies [127] show that *B. uniformis* and *P. vulgatus* genome length is in the range of ~5 Mb, which places them between *E. coli* and *B. thetaiotaomicron*, but their number of annotated genes is lower (Table 3.2).

| Entries annotation score Species ( Ref. proteome ID) | Total entries | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| *B. uniformis* (UP000004110) | 4618 | 4617 | 1 | 0 | 0 | 0 |
| *P. vulgatus* (UP000002861) | 3982 | 3107 | 617 | 241 | 17 | 0 |
| *B. thetaiotaomicron* (UP000001414) | 4782 | 3735 | 725 | 286 | 25 | 11 |
| *E. coli* (UP000000625) | 4402 | 538 | 847 | 679 | 718 | 1620 |

**Table 3.1.** Characteristics of protein annotation quality in *B. uniformis*, *P. vulgatus*, *B. thetaiotaomicron* and *E. coli* according to UniProt [124]. The annotation score goes from 1 (basic, low confidence) to 5 (high confidence).

| Species | Type Strain | NCBI RefSeq assembly | Genome size | Genes (RefSeq) | Protein-coding (RefSeq) |
|---|---|---|---|---|---|
| *B. uniformis* | ATCC8492 | GCF_000154205.1 | 4.7 Mb | 3,872 | 3,692 |
| *P. vulgatus* | ATCC9482 | GCF_000012825.1 | 5.2 Mb | 4205 | 3982 |
| *B. thetaiotaomicron* | VPI-5482 | GCF_000011065.1 | 6.3 Mb | 4888 | 4687 |
| *E. coli* | K-12. MG1655 | GCF_000005845.2 | 4.6 Mb | 4651 | 4290 |

**Table 3.2.** *B. uniformis* and *P. vulgatus* genome information and comparison with model organism *E. coli* and the most studied *Bacteroides* species *B. thetaiotaomicron* according to their NCBI assembly [127].

From a metabolic modelling perspective, there are no genome-scale metabolic models of *B. uniformis* or *P. vulgatus* published to date. For *B. thetaiotaomicron*, there is a GEM from 2013 [98], reconstructed similarly to the AGORA procedure [68]: originally built with ModelSEED and then manually curated based on the literature [29]. This GEM was expanded later in 2021 [45], and includes comparison with *B. thetaiotaomicron* GEM reported as part of the AGORA database [68], as well as experimental biomass, GAM and NGAM quantifications. Though the latest *B. thetaiotaomicron* model is available only in the form of tables, and not in the SBML format [45].

Given the importance of *B. uniformis* and *P. vulgatus*, which remain poorly characterised, a Flagship project (as part of the Microbial Ecosystems Transversal Theme EMBL program) aims to establish these species as new model organisms for the human gut microbiome. As part of this large-scale effort, I carry out the *de novo* reconstruction and curation of genome-scale metabolic models of *B. uniformis* and *P. vulgatus* type strains. To this aim, I use the GEMsembler framework, which I previously developed, to assemble several draft models of these bacteria and to curate them using consensus approach together with the previously reported GEM for *B. thetaiotaomicron* as a phylogenetically close reference point. The available experimental data, collected either by collaborators or by myself, were then used to further curate the models for growth and secretion phenotypes and to assess the metabolic behaviour of *B. uniformis* and *P. vulgatus*.

## 3.2 Materials and Methods

### 3.2.1 Reconstructing draft genome-scale metabolic models (GEMs), assembling supermodels and consensus models for *B. uniformis* and *P. vulgatus*

Genomes of *B. uniformis* (BU) and *P. vulgatus* (PV) type strains were provided by my collaborators from the Michael Zimmermann and Robert Finn groups at EMBL and EMBL-EBI. Genomes underwent PacBio long-read sequencing, and afterwards were assembled with Flye [128] and annotated using mettannotator [129]. The resulting genomes are of similar size to previous assemblies 4.7 Mb (4688977 nt and 22713 nt plasmid) for *B. uniformis* and 5.2 Mb (5163190 nt) for *P. vulgatus*. The mettannotator pipeline increased the number of annotated proteins to 3846 for *B. uniformis* and 4200 for *P. vulgatus* compared to the RefSeq annotation (Table 3.2).

For each of the species, I downloaded AGORA models, built three draft CarveMe, gapseq, modelSEED models, and later, a master student in my lab, Valerio Boccolini, built draft Kbase models, making it a total of five draft GEMs per species. CarveMe command-line tool and modelSEED web service were used with the protein sequence files produced by mettannotator, while gapseq command-line tool and Kbase web platform were used with full genome sequences. In all cases, reconstruction was performed with a gram-negative template for the models and only default gap-filling without specifying any media, and for CarveMe and gapseq, I specified an anaerobic environment. AGORA models and the corresponding genomes were downloaded from the AGORA2 collection [68] https://www.vmh.life/files/reconstructions/AGORA2/version2.01/sbml_files/individual_reconstructions/. As genomes were not available for the AGORA2 collection, I used the ones provided in the AGORA1 version [67] (https://www.vmh.life/files/reconstructions/AGORA/genomes/AGORA-Genomes.zip ).

I used all five draft GEMs per species and the corresponding genome/protein sequences as input to GEMsembler, which I used to assemble supermodels of *B. uniformis* and *P. vulgatus*. Due to the challenges in extracting Kbase internal annotation (performed by RAST [67]), I decided to use Kbase models without genes. Additionally, biomass from KBase models, defined in the form of several artificial reactions and metabolites representing different parts of biomass (e.g. for proteins or lipids), was transformed into a lump biomass reaction in order for these models to be processed by GEMsembler. Supermodel and five consensus models for each species were assembled with default parameters and genes were converted to the custom protein sequence file with genes IDs as locus tags from the mettannotator annotation.

### 3.2.2 Growth in the minimal media

*B. uniformis* and *P. vulgatus* minimal media requirements were investigated with the help from postdoc Mariia Beliaeva and lab technician Matthias Gross from the Michael Zimmermann group. Since the Varel-Bryant minimal medium proposed for *Bacteroides* contains cysteine and methionine (Table 3.3) [130], ), I wanted to replace these compounds to make sure that they do not serve as additional carbon sources with B12 and $Na_2S$ respectively, but due to issues with $Na_2S$ solubility, I had to switch the media base to Eric Marten's minimal medium (Table 3.4) [131].

| Ingredients | Quantity per litre |
|---|---|
| Mineral 3B solution | 50 ml |
| Glucose (20% in $H_2O$, filter sterilised) | 25 ml (0.5% 28 mM) |
| L-cysteine (free base) | 1g |
| Hemin (100 mg hemin in 2ml 1M NaOH in 200ml $dH_2O$) | 10 ml |
| L-methionine (0.2% in $H_2O$, filter sterilised) | 10 ml |
| $FeSO_4$ (0.278g $FeSO_4$*$7H_2O$ in 100ml $H_2O$+2 drops HCl) | 1.5 ml |
| $NaHCO_3$ (10% sterile solution) | 20 ml |
| $H_2O$ | 885 ml |

**Table 3.3.** Composition of the Varel-Bryant minimal medium (VB) [130].

| Ingredients | Quantity per litre |
|---|---|
| 10X Bact salts (pH 6.9-7.1) | 100 ml |
| 100X Trace elements | 10 ml |
| Glucose (20% in $H_2O$, filter sterilised) | 25 ml (0.5% 28 mM) |
| $Na_2S$ solution (nonahydrate, 0.05 g mL-1) (instead of Cys) | 10 ml |
| Vitamin B12 (0.3 g mL-1) | 20 ul |
| Hemin (0.5 mg mL-1) (instead of hematin/histidine) | 4 ml |
| $CaCl_2$ (0.1 M) | 0.75 ml |
| $MgCl_2$ (0.1 M) | 1 ml |
| $FeSO_4$ (0.4 mg mL-1) | 1 ml |
| Vitamin K3 (1 mg mL-1) | 1 ml |
| $H_2O$ | 885 ml |

**Table 3.4.** Composition of the Eric Marten's minimal medium [131] with $Na_2S$ and hemin.

Growth experiments were performed according to the following protocol.

Day 1:

1) Streak out the bacteria from the glycerol stock on the BHI (brain heart infusion) blood plate and incubate at 37°C anaerobically for 2 days. Afterwards keep at RT, wrapped in airtight foil until further usage.

2) Prepare 500 mL of 1X Eric Marten's minimal medium with $Na_2S$, check the final pH (to be between 6.9 - 7.1), sterile filter, and place into the anaerobic chamber.

3) Place 7 bottles for cultures (100 mL) into the chamber (for triplicate of each strain + control).

Day 3:

4) Pick up several colonies (one 'full' 1 µL inoculation loop of biomass) from the BHI plate and inoculate into 5 mL of the preculture media (in triplicate), let it grow at 37°C for two days.

Day 5:

5) Check $OD_{600}$ of all cultures (diluted 1:5 solution in the media was measured with a spectrophotometer, then the value was multiplied by 5 to calculate the actual $OD_{600}$) and start 50 mL cultures in each culture from the corresponding preculture to achieve a starting $OD_{600}$ of 0.05 (start in the morning).

6) Measure $OD_{600}$ at different time points across 3 days: 0 h, 8 h, 24 h, 28 h, 32 h, 48 h, 52 h, and collect samples for metabolomics (metabolomics measurements were planned originally, but not measured due the bacterial cuture growth inconsistency).

The OD-to-gDW (optical density to grams of dry weight) coefficient was estimated for *P. vulgatus* at 52 h as follows:

Collect 10 mL of culture in 15 mL falcon tubes (once from each triplicate tube), spin for 30 min at max speed at 4°C to collect the pellet, and discard the supernatant. Then add 0.5 mL of 0.9% NaCl, resuspend, and transfer carefully to weighted Eppendorf tubes (weighted on 5-digit scales (up to decile mg), and write the mass and label on the tube (no tape label)). Spin for 20 min at 4°C at max speed, remove supernatant, and measure the wet weight. Dry on speedvac for 1 h at 45°C, check the mass, and put on speedvac for 0.5 h at 45°C, check the mass again. If the mass is not changing anymore, then report the final dry weight.

The final minimal medium was developed by another PhD student in my group Daniel Benjamin Torka as VBDBT minimal medium with glucose as the sole carbon source (see Results).

### 3.2.3 Curating biomass composition

Preliminary growth and biomass component production analysis for *B. uniformis* and *P. vulgatus* was performed using the draft (converted to BiGG nomenclature) and consensus models with GEMsembler "run_growth_full_flux_analysis" function and agreement score of potential biomass components was calculated by the GEMsembler "biomass" function.

Biomass reaction curation was based on the previously reported *B. thetaiotaomicron* model [45]. I incorporated most of its compounds, excluding metabolites created

specifically for the *B. thetaiotaomicron* model and absent from the BiGG database. Another exception is acyl-CoA, which is a modified form of coenzyme A, already included in the biomass. I also omitted biotin, since it was not a biomass component in any of the draft models and reported as a conditionally essential cofactor [132]. Chorismate was not included either, because it was also not part of the biomass in any of the draft models, and it serves as a major branch-point intermediate in the aromatic amino acid biosynthesis pathway [133]. Finally, vitamin B12 (adenosylcobalamin, adocbl_c) was removed from the biomass despite being present in three *B. uniformis* or four *P. vulgatus* draft models, since it is required for methionine biosynthesis but is not essential for *B. thetaiotaomicron* growth when methionine is available [134, 135].

As a general rule, cellular biomass components not reported in *B. thetaiotaomicron* model were still added to the final biomass if they were confirmed by three or more draft models in both *B. uniformis* and *P. vulgatus*. Additionally, I included KDO(2)-lipid IV(A) with laurate (kdo2lipid4L_c), which can be synthesized by two draft models (AGORA and gapseq) in both species, to enrich the Lipid A class of lipopolysaccharides (LPS), which is less abundant compared to the glycerophospholipid class in the biomass equation. In contrast, I removed spermidine (spmd_c) because, in *B. uniformis*, although it is included in the biomass reaction of three models, it can be synthesized only by the CarveMe model. Similarly, acyl carrier protein (ACP_c), which is present in the biomass reaction of all models except CarveMe, cannot be synthesized at all. ACP is also not included in the biomass of the BiGG models [98], and therefore it was removed from the final biomass.

Biomass reaction stoichiometry was based on the curated *B. thetaiotaomicron* reaction. Initial coefficients were taken from *B. thetaiotaomicron* whenever available; otherwise, the median values from the draft models were used. GAM and NGAM were adopted unchanged from the *B. thetaiotaomicron* model. To normalise the non-ATP biomass stoichiometric coefficients to 1 g of cell dry weight [29, 74], the growth-associated ATP hydrolysis reaction (with GAM-derived coefficients) was temporarily removed from the lump biomass reaction; the remaining coefficients were normalised; the ATP hydrolysis term was then restored.

The final biomass reaction was saved as a separate SBML model with one reaction, so that for the FBA simulations it can be easily swapped with the draft biomass from any model.

### 3.2.4 Curation for growth on the minimal media

Biomass components requiring external curation were identified by intersecting the final biomass components with those not produced by the assembly models of either *B. uniformis* or *P. vulgatus* in the preliminary growth analysis with GEMsembler described in the previous section. Potential biosynthesis pathways for the identified metabolites (three lipids) were determined using GEMsembler pFBA functionality on the universal

BiGG model [33], as the curated *B. thetaiotaomicron* model is available only in tables and not in SBML format, therefore it could not be used as input for FBA simulations or to GEMsembler for direct comparison of synthesis paths. One reaction common to all pathways was verified in the *B. thetaiotaomicron* tables and then added to the consensus models from the universal BiGG. For the modified models, biomass component production analysis and identification of the corresponding pathways were performed again using the GEMsembler "run_growth_full_flux_analysis" functionality.

For all simulations the VBDBT minimal media was used as the following dictionary: vbdt_media = "glc__D_e": 10, "pheme_e": 10, "hemeA_e":10, "b12_e": 10, "adocbl_e": 10, "tsul_e": 10, "k_e": 10, "pi_e": 10, "na1_e": 10, "cl_e": 10, "nh4_e": 10, "so4_e": 10, "mg2_e": 10, "fe2_e": 10, "fe3_e": 10, "ca2_e": 10, "zn2_e": 10, "mn2_e": 10, "cu2_e": 10, "cobalt2_e": 10, "H2O_e": 10, "h_e": 10, "ni2_e": 10, "hco3_e":10

Curation of the core3 *B. uniformis* model was performed by adding missing reactions from other models and systematically ensuring biomass component production one by one. For each biomass compound not produced by core3, I examined the potential pathway maps (automatically generated by GEMsembler using the "run_growth_full_flux_analysis" functionality), calculated either for the core2 model if it could produce the corresponding metabolite, or for the assembly and draft models otherwise. From each map, I identified reactions with lower agreement than core3 that were essential based on pathway topology, and ensured their addition to the core3 model in GEMsembler. The resulting ability to produce each biomass component was verified using FBA optimization of its demand reaction.

Since missing reactions were added sequentially and one biomass component could depend on the production of another, curation was performed starting with the most basic metabolites (e.g., ATP, NAD, etc.). If the pathway map for a certain biomass component showed that it depended on another metabolite, the latter was curated first. Three lipids (pg140_c, pe140_c, clpn140_c), which none of the *B. uniformis* models could produce, were first curated in the *P. vulgatus* core3 model using the assembly model. The identified reactions were then added to the *B. uniformis* model from the *P. vulgatus* assembly model using COBRApy, and demand FBA simulations for these compounds confirmed that production was successfully restored in *B. uniformis* as well.

Curation of the *P. vulgatus* core3 model was initiated by ensuring the addition of reactions curated for *B. uniformis* using GEMsembler. If reactions were already present in the *P. vulgatus* core3 model, they were removed from the curation list; if certain reactions were absent from the *P. vulgatus* supermodel or failed to restore growth, the corresponding biomass components were curated additionally with *P. vulgatus* maps. Glycerophospholipids were curated separately for *B. uniformis* and *P. vulgatus*. For

phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1; pg181_c), I chose the *B. uniformis* pathway for both species, as the pathway predicted for *P. vulgatus* was substantially longer and predicted only by the CarveMe model, without any GPRs for its reactions.

Additionally, I added all available genes to the curated core3 models of *B. uniformis* and *P. vulgatus* by incorporating GPRs for all reactions lacking a core3 GPR, using the corresponding GPR from core2 if available, or otherwise from the assembly model. BiGG and, if available, ModelSEED IDs for metabolites and reactions were added as annotations to their corresponding databases. The final models were then saved as SBML files.

### 3.2.5 Assessing models with experimental growth in the Varel–Bryant minimal medium and extracellular metabolite dynamics

To curate models based on their metabolic capabilities, I utilised the results of targeted GC-MS/MS time-course analysis of extracellular metabolites, performed by PhD student Nikita Denisov from the Michael Zimmermann group. He cultured four biological replicates of *B. uniformis* and *P. vulgatus* (wild-type strains) in Varel–Bryant minimal medium for 24 hours, measuring OD every hour. Supernatants for metabolomic analysis were collected at the start of the experiment (0 h), and then hourly from 8 h to 19 h, with the final samples taken at 24 h. Metabolites were extracted and analyzed with GC-MS/MS using the targeted method developed by Nikita Denisov [136].

For the growth analysis, I normalised OD values so that the mean OD of the three no-bacteria control replicates at each time point was fixed at 0.001. This adjustment corrected for a slight increase in OD in the control samples over time, likely caused by evaporation. Growth rates were estimated by fitting a logistic growth function with the 'curve_fit' method from the SciPy library, using initial estimates ($p_0$ = 0.01, 0.2, 0.4) and bounds ((0.0, 0.001, 0.005), (0.1, 1.5, 1.5)) for three parameters $y_0$ (initial population), r (growth rate) and K (carrying capacity), correspondingly. Four biological replicates for each species were fitted simultaneously using nonlinear least-squares optimization with the trust region reflective algorithm. Dry cell weight for each time point was calculated by multiplying normalised OD values by the OD-to-biomass conversion coefficient determined in Section 2.2.2 and was later used to normalize extracellular metabolite fluxes determined by the metabolomics analysis.

Of the 90 compounds profiled by the GC-MS/MS, 66 have corresponding entries in the BiGG database. I therefore used the absolute concentrations (mM) of these 66 compounds in subsequent analyses. To determine which of these metabolites were unambiguously consumed or produced by *B. uniformis* and *P. vulgatus*, I compared the slopes of linear fits to their concentration profiles with the slope of the control sample over the 8–18 h interval. For slope comparison, bacterial replicate profiles were shifted so that their

initial time point matched the control value. Slope differences were tested using analysis of covariance (ANCOVA), implemented with smf.ols, by fitting a single model to both the control and four biological replicates per species and testing whether the interaction coefficient between time and group (species/control) was significantly different from zero via a t-test on that coefficient. P-values were adjusted for multiple testing using the default Benjamini–Hochberg procedure from stats.false_discovery_control in SciPy.

Metabolites were classified as consumed if they met all three criteria: adjusted p-value < 0.05 (significance), slope < -0.01 (effect size threshold), and $R^2 > 0.5$ (linearity assumption). Metabolites classified as consumed additionally to cysteine and methionine: ectoine in *P. vulgatus*, hydroxycinnamate in *B. uniformis*, and 4-aminobenzoate in both species were excluded from downstream analyses, as these compounds are not part of the defined medium; their presence at certain time points may indicate carryover from the bacterial precultures and is to be investigated separately. Metabolites classified as produced met the criteria: padj < 0.05, slope > 0.01, and $R^2 > 0.5$.

Exchange rates were computed for each replicate as the change in metabolite concentration between the initial and final time points within the 8-18 h interval, normalised by the corresponding change in dry weight biomass over the same period (mmol $gDW^{-1}$ $h^{-1}$). The mean and standard deviation of these consumption and production rates, as well as of the underlying concentration changes and biomass differences, were calculated across four biological replicates.

Consumption rates for cysteine and methionine were incorporated into the *P. vulgatus* model by constraining uptake to at most the mean rate plus one standard deviation, while no uptake was allowed in the *B. uniformis* model since these compounds were not consumed according to the experimental data. Production rates for secreted metabolites were implemented by enforcing secretion at least at the mean rate minus one standard deviation.

All flux balance analysis (FBA) simulations were performed using COBRApy. Network exploration was conducted with GEMsembler pathway and metabolite neighborhood maps. All additional plots were generated using the seaborn and matplotlib libraries.

## 3.3 Results

With GEMsembler workflow which I developed to compare metabolic models and aid the curation process, I set out to reconstruct and curate *B. uniformis* and *P. vulgatus* genome-scale metabolic models by using consensus modelling and experimental data on bacterial growth and metabolites secretion.

### 3.3.1 Reconstructing draft genome-scale metabolic models (GEMs), assembling supermodels and consensus models for *B. uniformis* and *P. vulgatus*

For both *B. uniformis* and *P. vulgatus*, I compiled five different draft GEMs (CarveMe, gapseq, ModelSEED, AGORA, KBase)) and subsequently used them to assemble the supermodel. The majority of metabolites and reactions from the five original models, as well as genes from the four models (excluding Kbase, for which genes were not available, see Methods) were successfully converted and included in the supermodels (Fig. 3.1, 3.2).

From these supermodels I then generated consensus GEMs and characterised both the draft and consensus models in terms of the number of reactions, metabolites, and genes (Fig. 3.1, 3.2). The draft models contained a substantial number of reactions, ranging from 1131 to 2418 for *B. uniformis* and from 1131 to 2474 for *P. vulgatus*, which were assembled into 2869 and 3110 reactions in the respective supermodels. Approximately half of the metabolites (1007/2123 and 1110/2268 for *B. uniformis* and *P. vulgatus* respectively) and a bit more than half of the reactions (1670/2869 and 1867/3110 for *B. uniformis* and *P. vulgatus* respectively) were identified only by one model. Complete agreement between all five models was observed for no more than 20% of metabolites (488/2123 and 499/2268 for *B. uniformis* and *P. vulgatus* respectively) and 10% of reactions (383/2869 and 401/3110 for *B. uniformis* and *P. vulgatus* respectively). Core3 agreement level of the three models balances network agreement and size, demonstrating the number of included reactions within the range for published curated models [44, 99]: 842 for *B. uniformis* and 872 for *P. vulgatus* (Fig. 3.1, 3.2).

Four draft models with genes available, ranging from 707 to 894 for *B. uniformis* and from 710 to 872 for *P. vulgatus*, assemble into 1221 and 1286 genes in respective supermodels (Fig. 3.1, 3.2). There were some losses in gene conversion for gapseq models due the possible difference in gene annotation, since gapseq performs the annotation itself. Similar issue is possible for AGORA models, but in this case some genes can not be converted due to the technical limitations of AGORA. More specifically, the fact that not all gene IDs from AGORA model can be found in AGORA genome and therefore can not be used for conversion. In the case of *B. uniformis* only 248 AGORA genes had sequences, while for *P. vulgatus* the number is 777 and closer to the total number of genes in the model. Therefore *B. uniformis* demonstrated lower agreement for genes than *P. vulgatus* with 573/1221 and 483/1286 genes being predicted only by one model, and 108/1221 and 251/1286 genes having the highest possible agreement of four models, respectively. For core3 agreement level the number of genes is higher (348 for *B. uniformis* and 515 for *P. vulgatus*), but still substantially lower than the number of reactions. The reason for such

difference is that GPR agreement (maximum number of models that agree at least on some part of the predicted GPR for a given reaction) often does not correspond to the agreement score of the reaction itself. From distributions of GPR agreement for reactions with a certain agreement level, it is clear that often GPR agreement is one level lower than agreement of reaction itself, but the difference can be bigger and for any levels of reaction agreement there are reactions without any GPR assigned (Fig. 3.1, 3.2).

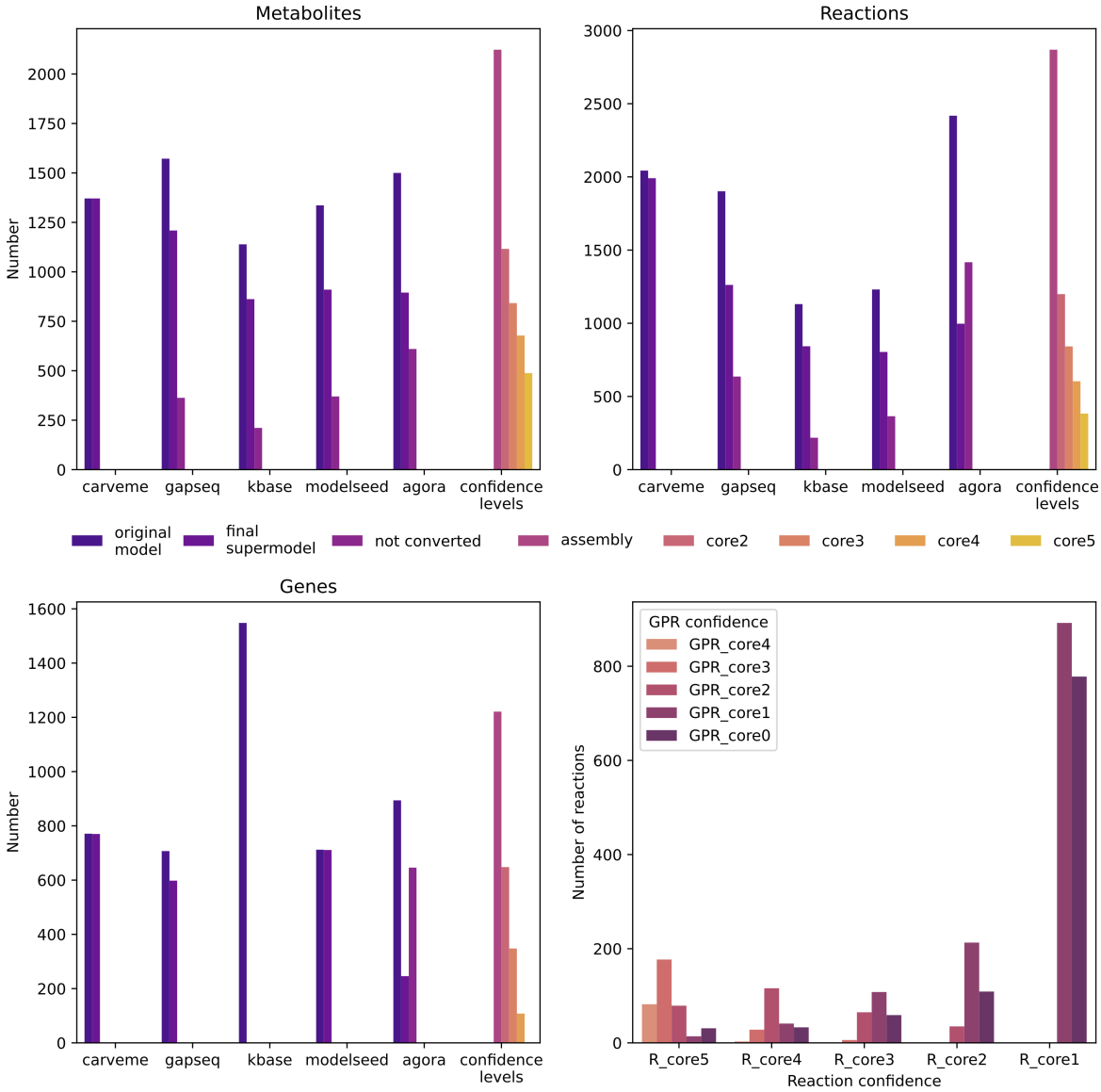


**Figure 3.1.** The number of metabolites, reactions and genes in the draft models, their conversion status to supermodel and models' agreement on metabolites, reactions, genes and GPRs for *B. uniformis*. CoreX corresponds to the agreement of X models (assembly = core1). GPR_core0 means no GPR in any model.

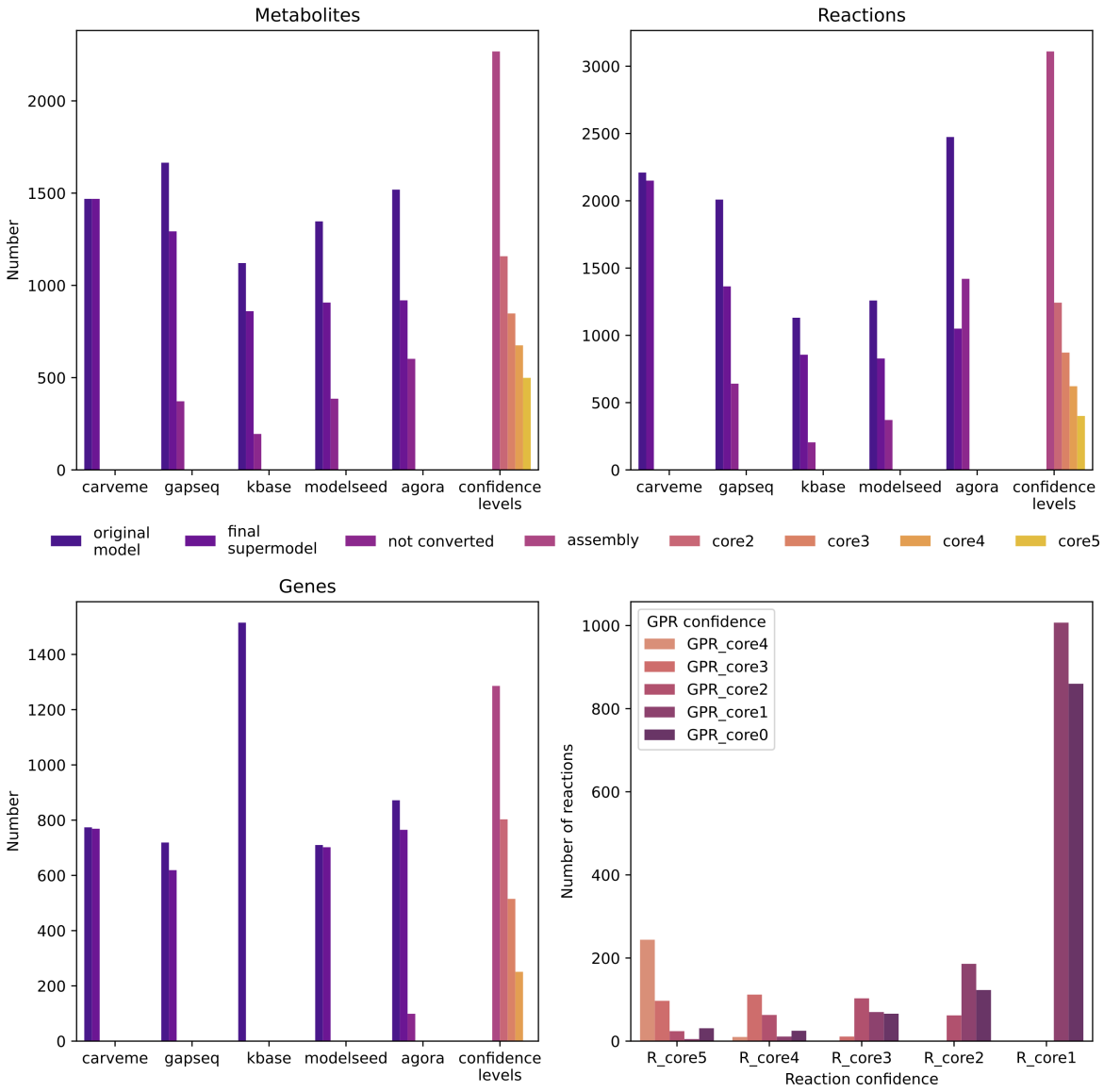


**Figure 3.2.** The number of metabolites, reactions and genes in the draft models, their conversion status to supermodel and models' agreement on metabolites, reactions, genes and GPRs for *P. vulgatus*. CoreX corresponds to the agreement of X models (assembly = core1). GPR_core0 means no GPR in any model.

In general, the agreement between *B. uniformis* models is slightly higher than between *P. vulgatus* models with respect to metabolites and reactions. While in the case of genes, the trend is in favour of *P. vulgatus*. From the perspective of network size, the core3 agreement level provides a reasonable balance between confidence and overall network coverage but still requires additional curation of GPRs to increase the number of genes included in the model.

### 3.3.2 Growth in the minimal media

Before I could start curating *B. uniformis* and *P. vulgatus* genome-scale metabolic models I needed to know their basic nutritional requirements. Given the availability of the well-established Varel–Bryant minimal medium for *Bacteroides* (Table 3.3) [130] or

Eric Marten's minimal medium (Table 3.4) [131], which contains only glucose, cysteine and methionine, or glucose, cysteine and histidine respectively as potential carbon sources, means that *B. uniformis* and *P. vulgatus* are not auxotrophic to any other compounds. The auxotrophy to cysteine and the requirements for histidine and methionine still needed to be assessed. No auxotrophies to L-methionine and L-histidine are likely, because these two media already provide substitutions for these compounds: unlike Eric Marten's medium, which contains hematin and L-histidine, the Varel–Bryant minimal medium includes hemin; whereas in Eric Marten's formulation L-methionine is substituted with vitamin B12, which is required for methionine biosynthesis in *Bacteroides* [134, 135]. L-cysteine serves two roles in these minimal media: reducing agent and sulfur source, since sulfate from the solution can not be utilised by *Bacteroides* [130]. That is why, with the guidance of postdoc Mariia Beliaeva, I set out to replace cysteine with $Na_2S$, as sulfide can overtake both roles [130, 137].

I initially added $Na_2S$ to the Varel–Bryant minimal medium, but due to its poor solubility and precipitation in the buffer, I switched to Eric Marten's minimal medium, which contains a different buffer that allowed $Na_2S$ to dissolve (Table 3.4). With this medium, $Na_2S$ was fully dissolved, and growth of *B. uniformis* and *P. vulgatus* was observed in some culturing experiments (data not shown). However, the bacterial culture growth was inconsistent, since sometimes cultures grew to $OD_{600}$ of more than 1, while in other repetitions of the experiment no growth was observed.

Later, this experiment was continued by another PhD student in my group Daniel Benjamin Torka, who managed to substitute L-cysteine in the Varel–Bryant minimal medium with $Na_2S_2O_3$ as sulfur source and dithiothreitol as a reducing agent. With B12 substitution of L-methionine, it resulted in a minimal medium not containing any amino acids (Table 3.5), with only glucose as a nutritional source. Both *B. uniformis* and *P. vulgatus* can grow reproducibly well in this medium.

| Ingredients | Quantity per litre |
|---|---|
| Mineral 3B solution | 50 ml |
| Glucose (20% in $H_2O$, filter sterilised) | 25 ml (0.5% 28 mM) |
| 160mM Dithiothreitol (instead of Cys) | 25ml (4 mM) |
| 160 $Na_2S_2O_3$ (instead of Cys) | 25ml (4 mM) |
| Hemin solution | 10 ml |
| Vitamin B12 0.05% (instead of Met) | 1 ml |
| $FeSO_4$ solution | 1.5 ml |
| $NaHCO_3$ (10% sterile solution) | 20 ml |
| $H_2O$ | 885 ml |

**Table 3.5.** Final minimal medium with only glucose as a nutritional source (VBDBT)

In order to use experimental results as constraints for modelling, I estimated a growth-related parameter: the conversion ratio between optical density (OD), measured experimentally, and gram dry weight (gDW), used in modelling. I measured the $OD_{600}$ of *P. vulgatus* in one of my growth experiments and dried and weighed the biomass from three corresponding cultures. The resulting OD-to-gDW conversion coefficient k equals 0.4896 ± 0.0137, which I extrapolated on *B. uniformis* as well.

| Measurement | P. vulgatus 1 | P. vulgatus 2 | P. vulgatus 3 |
|---|---|---|---|
| $OD_{600}$ at 52 h | 1.260 | 1.436 | 1.260 |
| tube weight (g) | 1.0035 | 1.0042 | 1.0074 |
| wet weight (g) 10ml | 1.1137 | 1.1214 | 1.1372 |
| dry weight -1h (g) | 1.0122 | 1.0162 | 1.0171 |
| dry weight -1.5h (g) | 1.0093 | 1.0121 | 1.0136 |
| dry weight - 2h (g) | 1.0096 | 1.0115 | 1.0134 |
| biomass (min - tube) (g) | 0.0061 | 0.0073 | 0.0060 |
| ConversionConvertion 1L k=biomass/OD*100 | 0.4841 | 0.5084 | 0.4762 |

**Table 3.6.** OD to gDW (dry weight) conversion. The three *P. vulgatus* columns correspond to three biological replicates.

Results above demonstrate that neither *B. uniformis* nor *P. vulgatus* are auxotrophic to any metabolite and can generate all required biomass components from glucose. This information together with thecalculation of their dry weight I will use further for curation of their genome-scale metabolic models.

### 3.3.3 Curating biomass composition

With the availability of VBDBT minimal media and no auxotrophies identified, *B. uniformis* and *P. vulgatus* were expected to synthesize all biomass components from glucose. Therefore, the first crucial curation task was to define the composition of their biomass reaction.

Initially, I used GEMsembler to verify that the draft and consensus models of *B. uniformis* and *P. vulgatus* could not grow in the VBDBT minimal medium without curation, and to identify biomass components that could not be produced, as well as the differences in their representation in the biomass reaction. Predicted synthesis of biomass components varied considerably across the draft models (Fig. 3.3), and this information was later incorporated into the curation process (see Materials and Methods).

Despite the differences between draft models, potential biomass components do not differ between *B. uniformis* and *P. vulgatus* except cardiolipin tetraoctadecanoyl, n-C18:0 (clpn180_c), which is only present in the biomass of AGORA *P. vulgatus* model, leading to 74 and 75 assembly reactants respectively (Table 3.7). Given this similarity and the common practice of inheriting biomass reactions from closely related, well-curated bacterial species, I decided to assemble a single biomass reaction for both *B. uniformis* and *P. vulgatus*, based on the reported curated biomass reaction of *B. thetaiotaomicron* [45] (Table 3.7).

The majority of biomass components from *B. thetaiotaomicron* [45] were included in the final biomass, regardless of how many draft GEMs of *B. uniformis* and *P. vulgatus* included them (Table 3.7). With this approach, 54 biomass components predicted in some *B. uniformis* and *P. vulgatus* models were confirmed by the curated model of *B. thetaiotaomicron*. Adenosylcobalamin (adocbl_c) was removed from the biomass (see Materials and Methods). Thirteen biomass components that were absent from the biomass reactions of both *B. uniformis* and *P. vulgatus*, as well as clpn180_c, which was not detected in *B. uniformis*, were also inherited from the *B. thetaiotaomicron* biomass reaction. Among the remaining 19 assembly reactants not present in the *B. thetaiotaomicron* biomass reaction, eight were included based on the agreement among draft GEMs and their ability to synthesise these metabolites. The products of the biomass reaction were kept in the minimal BiGG form, with products from the ATP hydrolysis and diphosphate (ppi_c) derived from nucleotide and protein biosynthesis. As a result, the biomass reaction of *B. uniformis* and *P. vulgatus* contained 76 reactant metabolites and 4 product metabolites (Table 3.7).

Biomass stoichiometry was inherited from *B. thetaiotaomicron* wherever possible, with the remaining coefficients derived from the draft models. Growth-associated maintenance (GAM) was incorporated into the biomass reaction as an ATP hydrolysis term, while non-growth-associated maintenance (NGAM) was represented as the lower bound of the ATP maintenance (ATPM) reaction. The GAM and NGAM values (18.526 and 5.54, respectively) were adopted from the curated model of *B. thetaiotaomicron* as well [45]. To complete the curation, I normalised the biomass reaction coefficients to 1 gDW of biomass.

The curated *B. uniformis* and *P. vulgatus* biomass reaction was finalized as a standalone model with single reaction and subsequently incorporated into the consensus models, replacing the draft biomass formulations.

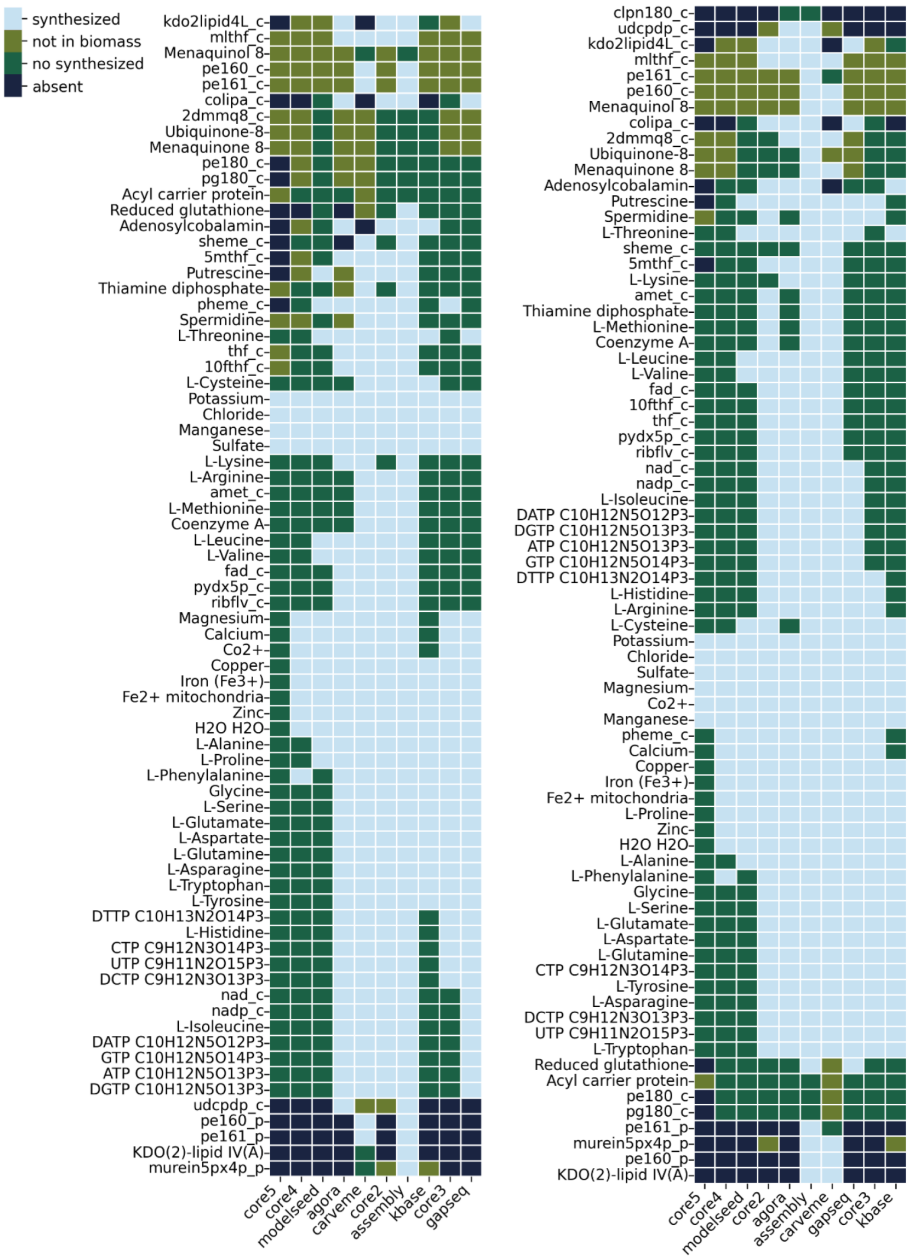


**Figure 3.3.** Preliminary production of potential biomass components in *B. uniformis* (left) and *P. vulgatus* (right) calculated with GEMsembler. The list of biomass components is taken from the assembly model.

| Name | Metabolite | B.theta | Type | Met_bu | Biomass_bu | Met_pv | Biomass_pv | Decision |
|---|---|---|---|---|---|---|---|---|
| Biotin | btn_c | 2.00E-06 | r | | | | | no |
| Co2+ | cobalt2_c | 2.50E-05 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Malonyl CoA $C_{24}H_{33}N_7O_{19}P_3S$ | malcoa_c | 3.13E-05 | r | | | | | yes |
| Nicotinamide adenine dinucleotide - reduced | nadh_c | 4.47E-05 | r | | | | | yes |
| Undecaprenyl diphosphate | udcpdp_c | 5.54E-05 | r | Core2 | Core1 | Core2 | Core1 | yes |
| Nicotinamide adenine dinucleotide phosphate | nadp_c | 1.12E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Coenzyme A | coa_c | 1.68E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| 5,10-Methylenetetrahydrofolate | mlthf_c | 2.23E-04 | r | Core5 | Core1 | Core5 | Core1 | yes |
| 5-Methyltetrahydrofolate | 5mthf_c | 2.23E-04 | r | Core5 | Core3 | Core5 | Core4 | yes |
| Adenosylcobalamin | adocbl_c | 2.23E-04 | r | Core4 | Core3 | Core4 | Core4 | no |
| 10-Formyltetrahydrofolate | 10fthf_c | 2.23E-04 | r | Core5 | Core4 | Core5 | Core5 | yes |
| 5,6,7,8-Tetrahydrofolate | thf_c | 2.23E-04 | r | Core5 | Core4 | Core5 | Core5 | yes |
| Thiamine diphosphate | thmpp_c | 2.23E-04 | r | Core5 | Core4 | Core5 | Core5 | yes |
| S-Adenosyl-L-methionine | amet_c | 2.23E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Flavin adenine dinucleotide oxidized | fad_c | 2.23E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Pyridoxal 5'-phosphate | pydx5p_c | 2.23E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Riboflavin $C_{17}H_{20}N_4O_6$ | ribflv_c | 2.23E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Chorismate | chor_c | 2.23E-04 | r | | | | | no |
| Acetyl-CoA | accoa_c | 2.79E-04 | r | | | | | no |
| Zinc | zn2_c | 3.33E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Nicotinamide adenine dinucleotide phosphate - reduced | nadph_c | 3.35E-04 | r | | | | | yes |
| Manganese | mn2_c | 6.77E-04 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Nicotinamide adenine dinucleotide | nad_c | 1.79E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Sulfate | so4_c | 4.25E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Chloride | cl_c | 5.10E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Magnesium | mg2_c | 8.50E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| DTTP $C_{10}H_{13}N_2O_{14}P_3$ | dttp_c | 9.24E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| DCTP $C_9H_{12}N_3O_{13}P_3$ | dctp_c | 9.31E-03 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Phosphatidylglycerol (dihexadecanoyl, n-C16:0) | pg160_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylethanolamine (dihexadecanoyl, n-C16:0) | pe160_c | 1.06E-02 | r | Core5 | Core1 | Core5 | Core1 | yes |
| Phosphatidylethanolamine (dioctadecanoyl, n-C18:0) | pe180_c | 1.06E-02 | r | Core5 | Core3 | Core5 | Core4 | yes |
| Phosphatidylglycerol (dioctadecanoyl, n-C18:0) | pg180_c | 1.06E-02 | r | Core5 | Core3 | Core5 | Core4 | yes |
| Cardiolipin (tetraoctadecanoyl, n-C18:0) | clpn180_c | 1.06E-02 | r | | | Core1 | Core1 | yes |
| Cardiolipin (tetratetradecanoyl, n-C14:0) | clpn140_c | 1.06E-02 | r | | | | | yes |

| Name | Metabolite | B.theta | Type | Met_bu | Biomass_bu | Met_pv | Biomass_pv | Decision |
|---|---|---|---|---|---|---|---|---|
| Cardiolipin (tetrahexadecanoyl, n-C16:0) | clpn160_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylethanolamine (didodecanoyl, n-C12:0) | pe120_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylethanolamine (ditetradecanoyl, n-C14:0) | pe140_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1) | pe181_c | 1.06E-02 | r | | | | | yes |
| | peai15_c | 1.06E-02 | r | | | | | no |
| | peai17_c | 1.06E-02 | r | | | | | no |
| | pei14_c | 1.06E-02 | r | | | | | no |
| | pei15_c | 1.06E-02 | r | | | | | no |
| | pei16_c | 1.06E-02 | r | | | | | no |
| | pei17_c | 1.06E-02 | r | | | | | no |
| Phosphatidylglycerol (didodecanoyl, n-C12:0) | pg120_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylglycerol (ditetradecanoyl, n-C14:0) | pg140_c | 1.06E-02 | r | | | | | yes |
| Phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1) | pg181_c | 1.06E-02 | r | | | | | yes |
| | pgai17_c | 1.06E-02 | r | | | | | no |
| | pgi17_c | 1.06E-02 | r | | | | | no |
| DATP $C_{10}H_{12}N_5O_{12}P_3$ | datp_c | 1.23E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| DGTP $C_{10}H_{12}N_5O_{13}P_3$ | dgtp_c | 1.24E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Ammonium | nh4_c | 1.27E-02 | r | | | | | yes |
| $Fe_2$+ mitochondria | fe2_c | 1.38E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Tryptophan | trp__L_c | 3.12E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| GTP $C_{10}H_{12}N_5O_{14}P_3$ | gtp_c | 3.78E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| UTP $C_9H_{11}N_2O_{15}P_3$ | utp_c | 4.23E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| CTP $C_9H_{12}N_3O_{14}P_3$ | ctp_c | 4.74E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Histidine | his__L_c | 6.42E-02 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Cysteine | cys__L_c | 1.01E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Calcium | ca2_c | 1.03E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Methionine | met__L_c | 1.21E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Arginine | arg__L_c | 1.30E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Threonine | thr__L_c | 1.65E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Asparagine | asn__L_c | 1.87E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Aspartate | asp__L_c | 1.87E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Potassium | k_c | 1.91E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Isoleucine | ile__L_c | 2.16E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Phenylalanine | phe__L_c | 2.19E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Proline | pro__L_c | 2.25E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| Glycine | gly_c | 2.30E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Valine | val__L_c | 2.51E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Tyrosine | tyr__L_c | 2.54E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Glutamine | gln__L_c | 2.62E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |

| Name | Metabolite | B.theta | Type | Met_bu | Biomass_bu | Met_pv | Biomass_pv | Decision |
|---|---|---|---|---|---|---|---|---|
| L-Glutamate | glu__L_c | 2.62E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Alanine | ala__L_c | 2.63E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Serine | ser__L_c | 2.79E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Leucine | leu__L_c | 3.58E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| L-Lysine | lys__L_c | 3.69E-01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| $H_2O$ $H_2O$ | H2O_c | 1.44E+01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| ATP $C_{10}H_{12}N_5O_{13}P_3$ | atp_c | 1.86E+01 | r | Core5 | Core5 | Core5 | Core5 | yes |
| KDO(2)-lipid IV(A) with laurate | kdo2lipid4L_c | | r | Core4 | Core1 | Core4 | Core1 | yes |
| KDO(2)-lipid IV(A) | kdo2lipid4_p | | r | Core1 | Core1 | Core1 | Core1 | no |
| Menaquinol 8 | mql8_c | | r | Core5 | Core1 | Core5 | Core1 | no |
| Two disacharide linked murein units, pentapeptide crosslinked tetrapeptide (A2pm->D-ala) (middle of chain) | murein5px4p_p | | r | Core2 | Core1 | Core2 | Core1 | no |
| Phosphatidylethanolamine (dihexadecanoyl, n-C16:0) | pe160_p | | r | Core1 | Core1 | Core1 | Core1 | no |
| Phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1) | pe161_c | | r | Core5 | Core1 | Core5 | Core1 | no |
| Phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1) | pe161_p | | r | Core1 | Core1 | Core1 | Core1 | no |
| 2-Demethylmenaquinone 8 | 2dmmq8_c | | r | Core5 | Core2 | Core5 | Core3 | no |
| Menaquinone 8 | mqn8_c | | r | Core5 | Core2 | Core5 | Core3 | no |
| Ubiquinone-8 | q8_c | | r | Core5 | Core2 | Core5 | Core3 | no |
| Core oligosaccharide lipid A | colipa_c | | r | Core3 | Core3 | Core3 | Core3 | yes |
| Reduced glutathione | gthrd_c | | r | Core4 | Core3 | Core5 | Core4 | yes |
| Putrescine | ptrc_c | | r | Core5 | Core3 | Core5 | Core4 | yes |
| Spermidine | spmd_c | | r | Core5 | Core3 | Core5 | Core4 | no |
| Acyl carrier protein | ACP_c | | r | Core5 | Core4 | Core5 | Core4 | no |
| Protoheme $C_{34}H_{30}FeN_4O_4$ | pheme_c | | r | Core5 | Core4 | Core5 | Core5 | yes |
| Siroheme $C_{42}H_{36}FeN_4O_{16}$ | sheme_c | | r | Core4 | Core4 | Core5 | Core5 | yes |
| Copper | cu2_c | | r | Core5 | Core5 | Core5 | Core5 | yes |
| Iron (Fe3+) | fe3_c | | r | Core5 | Core5 | Core5 | Core5 | yes |
| Diphosphate | ppi_c | 2.03E-01 | p | Core5 | Core5 | Core5 | Core5 | yes |
| ADP $C_{10}H_{12}N_5O_{10}P_2$ | adp_c | 1.85E+01 | p | Core5 | Core5 | Core5 | Core5 | yes |
| H+ | h_c | 1.85E+01 | p | Core5 | Core5 | Core5 | Core5 | yes |
| Phosphate | pi_c | 1.85E+01 | p | Core5 | Core5 | Core5 | Core5 | yes |
| Biomass: total | biomass_c | | p | Core1 | Core1 | Core1 | Core1 | no |
| Cobinamide | cbi_c | | p | Core4 | Core3 | Core4 | Core4 | no |
| 5,6-Dimethylbenzimidazole | dmbzid_c | | p | Core4 | Core3 | Core4 | Core4 | no |
| Apoprotein [acyl carrier protein] | apoACP_c | | p | Core5 | Core4 | Core5 | Core4 | no |

**Table 3.7.** Potential biomass components of *B. uniformis* and *P. vulgatus* originating from the assembly model, agreement among draft models on these metabolites, comparison with the curated *B. thetaiotaomicron* biomass model, and final decision on their inclusion in the biomass reaction. Metabolites without names are not present in the BiGG database.

### 3.3.4 Curation for growth on the minimal media

With the VBDBT minimal medium and curated list of biomass components, I set out to ensure their synthesis and *B. uniformis* and *P. vulgatus* growth.

Among the final curated biomass components three, phosphatidylethanolamine (pe180_c), phosphatidylglycerol (pg180_c), and cardiolipin (clpn180_c), could not be produced by any of consensus models in either *B. uniformis* or *P. vulgatus* (Fig. 3.3). Therefore, I searched for a potential biosynthesis pathway in the universal BiGG model [33] with pFBA. All three pathways included a common reaction, glycerol-3-phosphate acyltransferase (C18:0) (G3PAT180), which is also present in the curated *B. thetaiotaomicron* model. Adding this reaction to the *B. uniformis* and *P. vulgatus* models and reassessing curated biomass components production (Fig. 3.4, 3.5) demonstrated that the addition is sufficient to enable synthesis of these three lipids at least in the assembly models.

After introducing the G3PAT180 reaction from the *B. thetaiotaomicron* model, the largest consensus assembly model of *B. uniformis* was still unable to produce three glycerophospholipids (pg140_c, pe140_c, clpn140_c), whereas the *P. vulgatus* assembly model could produce all biomass components, including these metabolites (Fig. 3.4, 3.5). Therefore curating reactions across both species provides reactions in the *B. uniformis* and *P. vulgatus* assembly models sufficient to produce all required biomass components.

For both bacteria, curation is based on the core3 consensus model due to its balance between draft model agreement and functionality, specifically its ability to produce biomass components. The complete-agreement core5 model barely produces any biomass components, while core4 is also quite limited. In contrast, core3 matches the performance of the draft models, producing more biomass components than the ModelSEED and KBase GEMs and only falling behind gapseq by a few metabolites. Core2 reaches a similarly high level of biomass component production as the CarveMe and AGORA models, but confidence in this network is lower, as it may reflect agreement between only two draft models, which are already similar due to the features of the reconstruction tools (Fig. 3.4, 3.5).

Thus, biomass component production and growth in the VBDBT minimal medium were achieved by incorporating 93 reactions into the *B. uniformis* core3 model: 45 reactions were added from core2, 35 from one of the draft *B. uniformis* models, 12 from *P. vulgatus* models, and 1 from *B. thetaiotaomicron*, as described above (Table 3.8). The core3 model of *P. vulgatus* also incorporated 93 additional reactions: 37 confirmed by core2, 53 predicted by one of the draft *P. vulgatus* models, the same 1 reaction from *B. thetaiotaomicron*, and 2 reactions from *B. uniformis* models (Table 3.9). These latter two reactions were selected for phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1; pg181_c) biosynthesis because the

*P. vulgatus* pathway is considerably less confidently reconstructed than that of *B. uniformis*. These reactions are priority candidates for further validation.

Additionally, to avoid losing information on potential enzymes when reaction or GPR agreement was low, while still retaining GPRs with high core3 agreement whenever possible, I added GPRs from the core2 or assembly consensus models for all reactions lacking core3 GPRs. This increased the number of genes in the models by approximately twofold.


**Figure 3.4.** Production of curated biomass components in the VBDBT minimal medium and summary of corresponding biosynthesis pathway agreement in *B. uniformis* calculated with GEMsembler. Light blue indicates metabolites that are synthesized, while green indicates metabolites that are not. The intensity of orange represents the reaction agreement score, with the brightest colour corresponding to core5 reactions and the lightest to core1. The intensity of yellow represents the agreement score of the corresponding GPR, with the brightest colour indicating core5 and white indicating no GPR.

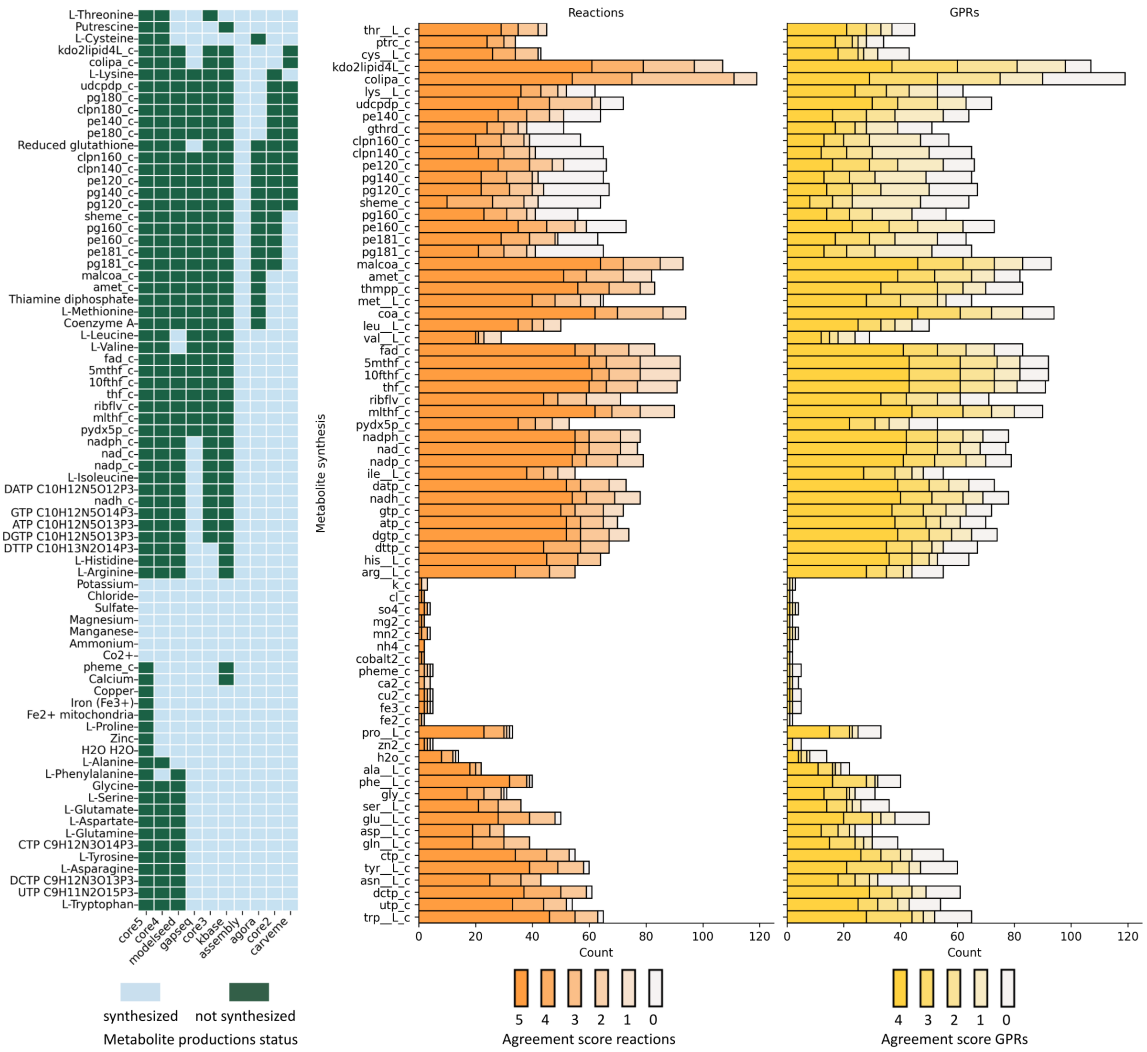


**Figure 3.5.** Production of curated biomass components in the VBDBT minimal medium and summary of corresponding biosynthesis pathway agreement in *P. vulgatus* calculated with GEMsembler. Light blue indicates metabolites that are synthesized, while green indicates metabolites that are not. The intensity of orange represents the reaction agreement score, with the brightest colour corresponding to core5 reactions and the lightest to core1. The intensity of yellow represents the agreement score of the corresponding GPR, with the brightest colour indicating core5 and white indicating no GPR.

The resulting core3 model of *B. uniformis* contains 984 reactions and 608 genes, while the core3 model of *P. vulgatus* contains 1,020 reactions and 651 genes. With an NGAM value of 5.54 set as the lower boundary for the ATP maintenance reaction, and VBDBT minimal medium uptake constrained to 10 mmol $gDW^{-1}$ $h^{-1}$, these *B. uniformis* and *P. vulgatus* models demonstrated substantial growth, reaching the rate of approximately 1.2 $h^{-1}$.

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| FE2t | 2 | carveme kbase | 1 |
| RHCYS | 1 | gapseq | 1 |
| GTHS | 1 | gapseq | 0 |
| GLUCYS | 1 | gapseq | 0 |
| 3HAD120 | 2 | agora kbase | 0 |
| G1PACT | 2 | agora modelseed | 0 |
| UAGDP | 2 | agora carveme | 1 |
| C120SN | 1 | gapseq | 0 |
| KAS15 | 2 | gapseq kbase | 1 |
| ACGAM6PS | 1 | gapseq | 0 |
| UDCPDPS | 1 | agora | 0 |
| HDECH | 1 | carveme | 0 |
| KAS7 | 1 | carveme | 1 |
| KAS8 | 1 | carveme | 1 |
| AGPATCOA_PALM_PALM_c | 1 | agora | 0 |
| DHDPRx_r | 1 | carveme | 1 |
| THDPS | 1 | carveme | 1 |
| SDPTA | 2 | carveme gapseq | 1 |
| SDPDS | 1 | carveme | 1 |
| MCTP1App | 1 | carveme | 1 |
| MPTG | 1 | carveme | 1 |
| DALAabcpp | 1 | carveme | 0 |
| AIRC1 | 2 | carveme gapseq | 1 |
| PAPSR | 2 | carveme kbase | 1 |
| SULR | 2 | agora carveme | 1 |
| AKGDH | 2 | carveme gapseq | 2 |
| DHORD6 | 2 | gapseq kbase | 1 |
| RBFSa | 2 | agora carveme | 1 |
| DHPPDA | 2 | agora carveme | 1 |
| GTPCII | 2 | agora carveme | 1 |
| AGTi | 2 | carveme gapseq | 1 |
| GCALDD | 2 | agora gapseq | 0 |
| DHNPA_1 | 2 | agora carveme | 0 |
| HPPK | 2 | agora carveme | 0 |
| FOLD3 | 2 | agora carveme | 1 |
| DNTPPA | 1 | carveme | 1 |
| DNMPPA | 1 | carveme | 1 |
| HPPK2 | 2 | agora carveme | 2 |
| DHPS2 | 2 | agora carveme | 2 |
| GCALDt | 2 | agora carveme | 0 |
| EX_gcald_e | 2 | agora carveme | 0 |
| ACLS | 2 | carveme kbase | 1 |
| METGL | 2 | carveme gapseq | 1 |
| AHSERL | 2 | carveme gapseq | 2 |
| METS | 2 | agora carveme | 1 |
| AHSERL2 | 2 | agora carveme | 2 |

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| PDX5PS | 2 | agora carveme | 2 |
| PDX5PO2 | 2 | agora gapseq | 2 |
| AHMMPS_1 | 2 | agora carveme | 1 |
| DM_4hba_c | 1 | agora | 0 |
| HSK | 2 | agora gapseq | 1 |
| OCBT | 2 | carveme gapseq | 1 |
| ORNDC | 2 | gapseq modelseed | 1 |
| ALAS | 2 | carveme gapseq | 0 |
| UAGDP | 2 | agora carveme | 1 |
| 3OAR40 | 2 | agora carveme | 1 |
| ACGAMPM | 1 | gapseq | 0 |
| UDPACGLP | 1 | carveme | 0 |
| GALUi | 2 | agora gapseq | 0 |
| MEPCT | 2 | agora carveme | 1 |
| UDCPDP | 2 | agora carveme | 1 |
| HPYRDC | 1 | modelseed | 1 |
| HPYRRx | 2 | carveme gapseq | 2 |
| AGPAT180 | 1 | agora | 0 |
| CLPNS180 | 1 | agora | 0 |
| 3OAR180 | 1 | agora | 0 |
| 3OAS180 | 1 | agora | 0 |
| 3HAD180 | 2 | agora kbase | 0 |
| AACPS3 | 1 | carveme | 1 |
| AACPS5 | 1 | carveme | 1 |
| KAS17 | 1 | carveme | 1 |
| G3PAT181 | 1 | carveme | 1 |
| PGSA181 | 2 | agora carveme | 0 |
| CLPNS160 | 1 | agora | 0 |
| PGSA160 | 2 | agora carveme | 0 |
| PItex | 1 | carveme | 1 |
| $H_2O$tex | 1 | carveme | 1 |
| PGSA120 | 2 | agora carveme | 0 |
| G3PAT120 | 1 | kbase | 0 |
| AGPAT120 | 1 | agora | 0 |
| 2AGPG140tipp | 0 | P.vulgatus | 0 |
| 2AGPGAT140 | 0 | P.vulgatus | 0 |
| AACPS1 | 0 | P.vulgatus | 0 |
| AGPATCOA_MYRS_MYRS_c | 0 | P.vulgatus | 0 |
| CLPNS140 | 0 | P.vulgatus | 0 |
| FACOAL140t2pp | 0 | P.vulgatus | 0 |
| G3PAT140 | 0 | P.vulgatus | 0 |
| G3PAT180 | 0 | B.thetaiotaomicron | 0 |
| KAS2 | 0 | P.vulgatus | 0 |
| PGP140abcpp | 0 | P.vulgatus | 0 |
| PGPP140pp | 0 | P.vulgatus | 0 |
| PGSA140 | 0 | P.vulgatus | 0 |

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| PLIPA1G140pp | 0 | P.vulgatus | 0 |

**Table 3.8.** Reactions incorporated into the *B. uniformis* core3 consensus model to enable biomass component synthesis and, consequently, growth in the VBDBT minimal medium. Number of draft *B. uniformis* models agreeing on each reaction, corresponding source of each reaction, and number of draft models agreeing on a non-empty GPR for each reaction.

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| FE2t | 2 | carveme kbase | 1 |
| HMBS | 2 | agora gapseq | 0 |
| RHCYS | 1 | gapseq | 1 |
| GTHS | 1 | gapseq | 0 |
| GLUCYS | 1 | gapseq | 0 |
| G1PACT | 2 | agora carveme | 1 |
| UAGDP | 2 | agora carveme | 1 |
| KAS15 | 2 | gapseq kbase | 1 |
| ACGAM6PS | 1 | gapseq | 0 |
| HDECH | 1 | carveme | 0 |
| KAS7 | 1 | carveme | 1 |
| KAS8 | 1 | carveme | 1 |
| AGPATCOA_PALM_PALM_c | 1 | agora | 1 |
| DHDPRx_r | 1 | carveme | 1 |
| THDPS | 1 | carveme | 1 |
| SDPTA | 2 | carveme gapseq | 1 |
| SDPDS | 1 | carveme | 0 |
| MCTP1App | 1 | carveme | 1 |
| MPTG | 1 | carveme | 1 |
| UDPDPS3 | 1 | agora | 1 |
| AIRC1 | 2 | carveme gapseq | 1 |
| SULR | 2 | carveme modelseed | 1 |
| AKGDH | 2 | carveme gapseq | 2 |
| RBFSa | 2 | agora carveme | 2 |
| DHPPDA | 2 | agora carveme | 2 |
| GTPCII | 2 | agora carveme | 2 |
| AGTi | 1 | gapseq | 1 |
| DHNPA_1 | 2 | agora carveme | 2 |
| HPPK | 2 | agora carveme | 1 |
| FOLD3 | 2 | agora carveme | 2 |
| DNTPPA | 1 | carveme | 1 |
| DNMPPA | 1 | carveme | 1 |
| HPPK2 | 2 | agora carveme | 1 |
| DHPS2 | 2 | agora carveme | 1 |
| GCALDt | 2 | agora carveme | 0 |
| EX_gcald_e | 2 | agora carveme | 0 |
| ACLS | 2 | carveme kbase | 1 |

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| METGL | 2 | carveme gapseq | 1 |
| AHSERL | 2 | carveme gapseq | 2 |
| METS | 2 | agora carveme | 2 |
| AHSERL2 | 2 | agora carveme | 2 |
| PDX5PS | 2 | agora carveme | 1 |
| AHMMPS_1 | 1 | carveme | 1 |
| DM_4hba_c | 2 | agora carveme | 0 |
| HSK | 2 | agora gapseq | 1 |
| ALAS | 1 | gapseq | 0 |
| UAGDP | 2 | agora carveme | 1 |
| 3OAR40 | 2 | agora carveme | 2 |
| ACGAMPM | 2 | gapseq modelseed | 0 |
| UDPACGLP | 1 | carveme | 0 |
| MEPCT | 2 | agora carveme | 2 |
| UDCPDP | 2 | agora carveme | 2 |
| HPYRDC | 1 | modelseed | 1 |
| AGPAT180 | 1 | agora | 1 |
| CLPNS180 | 1 | agora | 1 |
| PGPP180 | 2 | agora gapseq | 0 |
| AACPS6 | 1 | agora | 0 |
| KAS13 | 2 | carveme gapseq | 1 |
| AACPS5 | 1 | carveme | 1 |
| KAS17 | 1 | carveme | 1 |
| G3PAT181 | 1 | carveme | 1 |
| CLPNS160 | 1 | agora | 1 |
| PGSA160 | 1 | carveme | 0 |
| PGPP160 | 1 | carveme | 1 |
| CLPNS140 | 1 | agora | 1 |
| 2AGPGAT140 | 1 | carveme | 0 |
| KAS2 | 1 | carveme | 1 |
| AACPS1 | 2 | carveme gapseq | 1 |
| G3PAT140 | 1 | agora | 0 |
| PGSA140 | 1 | carveme | 0 |
| AGPATCOA_MYRS_MYRS_c | 1 | agora | 1 |
| PGP140abcpp | 1 | carveme | 1 |
| PGPP140pp | 1 | carveme | 0 |
| FACOAL140t2pp | 1 | carveme | 1 |
| PLIPA1G140pp | 1 | carveme | 0 |
| 2AGPG140tipp | 1 | carveme | 0 |
| PItex | 1 | carveme | 1 |
| $H_2Otex$ | 1 | carveme | 1 |
| 2AGPGAT120 | 1 | carveme | 0 |
| 2AGPG120tipp | 1 | carveme | 0 |
| DDCAt2pp | 1 | carveme | 0 |
| PLIPA1G120pp | 1 | carveme | 0 |
| PGPP120pp | 1 | carveme | 0 |

| Reaction | R_agreement | R_models | GPR_agreement |
|---|---|---|---|
| PGP120abcpp | 1 | carveme | 1 |
| PGSA120 | 1 | carveme | 0 |
| G3PAT120 | 1 | kbase | 0 |
| AGPAT120 | 1 | agora | 1 |
| EX_cys__L_e | 2 | carveme gapseq | 0 |
| CYSabc | 2 | carveme gapseq | 1 |
| CYSt2r | 1 | gapseq | 1 |
| G3PAT180 | 0 | B.thetaiotaomicron | 0 |
| PGSA181 | 0 | B.uniformis | 0 |
| PGPP181 | 0 | B.uniformis | 0 |

**Table 3.9.** Reactions incorporated into the *P. vulgatus* core3 consensus model to enable biomass component synthesis and, consequently, growth in the VBDBT minimal medium. Number of draft *P. vulgatus* models agreeing on each reaction, corresponding source of each reaction, and number of draft models agreeing on a non-empty GPR for each reaction.

### 3.3.5 Assessing quantitative model performance with experimental data on growth and extracellular metabolite fluxes in minimal medium

After ensuring that the models of *B. uniformis* and *P. vulgatus* can grow in glucose minimal medium, I accessed coupled growth and extracellular metabolic fluxes quantified with time-course targeted metabolomics data, generated by PhD student Nikita Denisov in the Varel–Bryant minimal medium.

*B. uniformis* began growing at the start of the 24-hour observation period but showed only moderate growth, reaching approximately 0.1 OD without clearly defined lag, exponential, or saturation phases (Fig. 3.6). In contrast, *P. vulgatus* exhibited a long lag phase, followed by rapid exponential growth and a plateau at around 0.4-0.5 OD by the end of the 24-hour observation. *P. vulgatus* also demonstrated higher variability between replicates at the second part of its growth (Fig. 3.6).

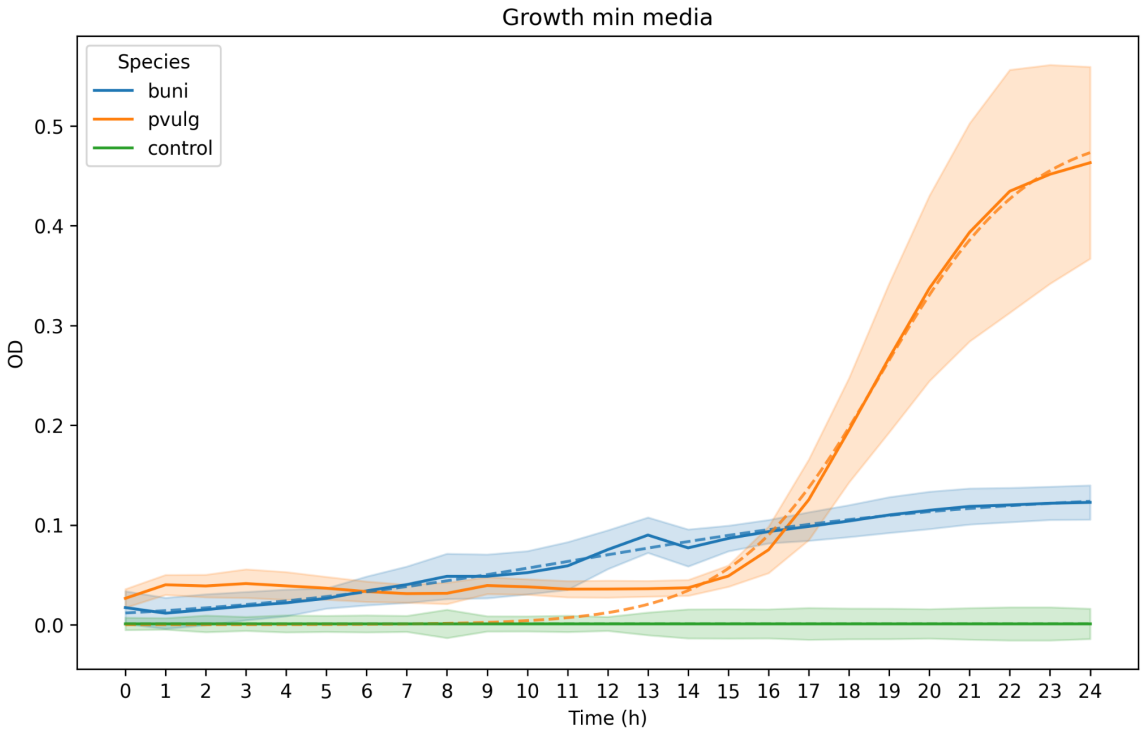


**Figure 3.6.** Growth curves of *B. uniformis* and *P. vulgatus* in the Varel-Bryant minimal medium. Growth curves are normalised and plotted as mean OD values of four biological replicates; shaded areas indicate standard deviation. Dashed curve demonstrates logistic fitting, used for growth rates calculation (0.20 for *B. uniformis* and 0.54 for *P. vulgatus*). The plot was generated by myself based on the experimental data provided by Nikita Denisov.

For the growth rate calculation I used logistic model, that incorporates the environmental carrying capacity K to account for the saturation, eventual plateau of the growth [138] (Fig. 3.6). The estimated model parameters indicate growth rates (r) of 0.20 $h^{-1}$ for *B. uniformis* and 0.54 $h^{-1}$ for *P. vulgatus*, carrying capacities (K) of 0.133 and 0.501, respectively, and initial population sizes ($y_0$) of 0.0118 and 0.000018. The initial population size estimated for *P. vulgatus* is lower than the observed values, likely due to its long lag phase. However, for subsequent analyses only the growth rates were required, and these estimates fall within the range of previously reported growth rates for *B. uniformis* and *P. vulgatus* in similar minimal media [118].

Since bacterial growth depends on substrate consumption, but glucose was not measured in these experiments, I first focused on the remaining nutrients available to *B. uniformis* and *P. vulgatus* in the Varel-Bryant minimal medium: cysteine and methionine. The metabolite concentration profiles showed that *P. vulgatus* consumed both cysteine and methionine, whereas *B. uniformis* did not take up either compound (Fig. 3.7).

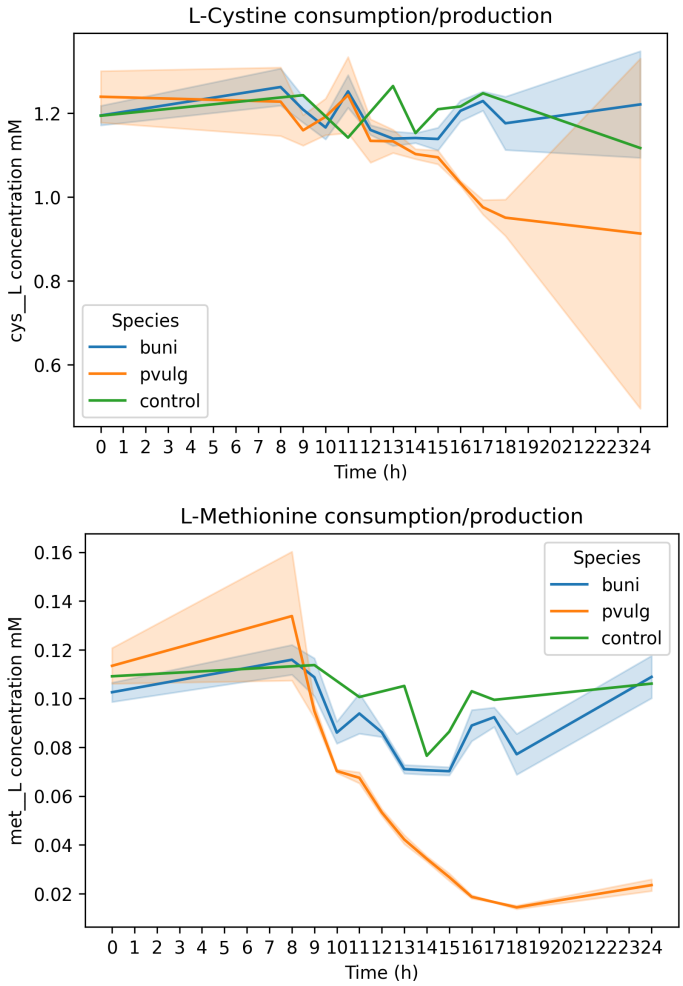


**Figure 3.7.** Cysteine and methionine consumption for *B. uniformis* and *P. vulgatus* in the Varel-Bryant minimal medium. The concentration curves are plotted as mean of four biological replicates; shaded areas indicate standard deviation. The plot was generated by myself based on the experimental data provided by Nikita Denisov.

To formally assess the consumption and secretion of all analyzed metabolites in *B. uniformis* and *P. vulgatus* grown in minimal medium, I decided to assume a linear concentration dynamic over the time interval between 8 and 18 hours, as this period corresponds to the early exponential phase in *P. vulgatus* (Fig. 3.6). This approach makes it possible to capture the exponential growth phase of both species and to identify the main exchange trends. Secreted metabolites were identified by comparing the slopes of linear fits to metabolite concentrations between bacterial and control samples using covariance analysis (ANCOVA). Eight metabolites exhibited a trend following the assumption of linearity and statistically significant increases with slopes of sufficient magnitude, indicating notable production. All eight metabolites were attributed to *P. vulgatus*, with four of these were also produced by *B. uniformis* (Table 3.10, Fig. 3.8). Actual secretion rates were calculated as the change in metabolite concentration in the 8–18 h interval over time, normalised to the corresponding change in dry weight biomass (mmol $gDW^{-1}$ $h^{-1}$)

calculated from the $OD_{600}$ curve. Final values are reported as the mean and standard deviation from four replicates (Table 3.10, Fig. 3.8).

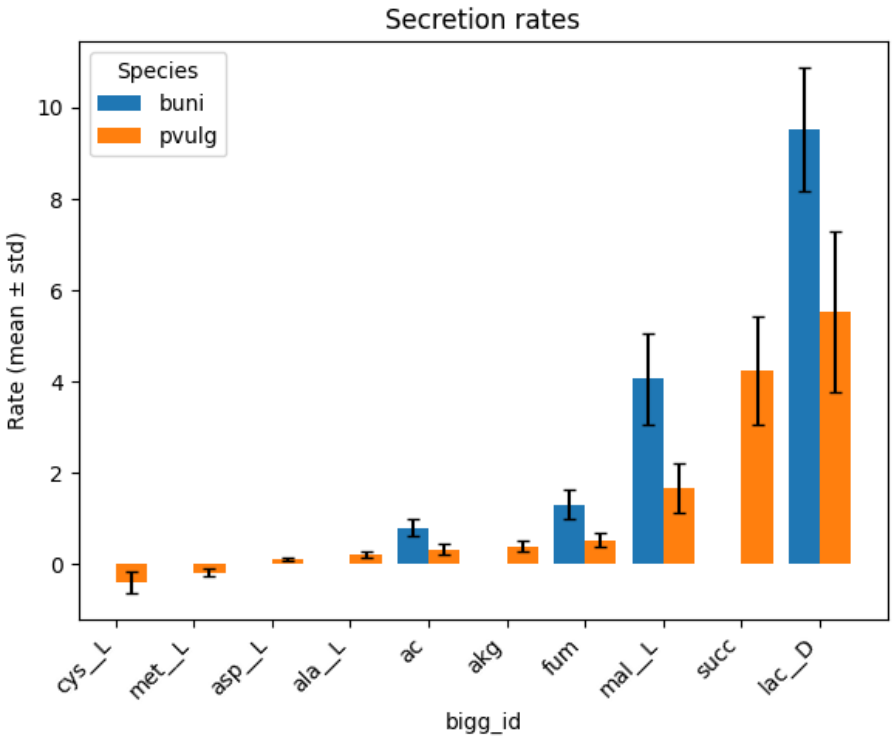


**Figure 3.8.** Secretion/uptake rates (mmol $gDW^{-1}$ $h^{-1}$) for *B. uniformis* and *P. vulgatus* in the Varel-Bryant minimal medium. Secreted metabolites are identified with ANCOVA. Error bars represent standard deviations of four replicates. The rate calculations were performed and the plot was generated by myself based on the experimental data provided by Nikita Denisov.

| bigg_id | Species | padj | rate_mean | rate_std | delta_conc_mean | delta_conc_std | delta_g_bio_mean | delta_g_bio_std |
|---|---|---|---|---|---|---|---|---|
| cys__L | pvulg | 3.13E-03 | -3.95E-01 | 2.28E-01 | -2.77E-01 | 8.79E-02 | 8.00E-02 | 2.79E-02 |
| met__L | pvulg | 7.10E-04 | -1.73E-01 | 9.10E-02 | -1.19E-01 | 2.69E-02 | 8.00E-02 | 2.79E-02 |
| asp__L | pvulg | 7.53E-08 | 1.10E-01 | 3.18E-02 | 8.20E-02 | 9.79E-03 | 8.00E-02 | 2.79E-02 |
| ala__L | pvulg | 2.65E-18 | 2.12E-01 | 5.80E-02 | 1.58E-01 | 1.34E-02 | 8.00E-02 | 2.79E-02 |
| ac | pvulg | 2.75E-04 | 3.30E-01 | 1.14E-01 | 2.40E-01 | 3.63E-03 | 8.00E-02 | 2.79E-02 |
| akg | pvulg | 7.39E-03 | 3.95E-01 | 1.17E-01 | 2.94E-01 | 3.33E-02 | 8.00E-02 | 2.79E-02 |
| fum | pvulg | 2.25E-08 | 5.38E-01 | 1.65E-01 | 3.99E-01 | 4.51E-02 | 8.00E-02 | 2.79E-02 |
| ac | buni | 2.31E-03 | 8.01E-01 | 1.92E-01 | 1.96E-01 | 2.70E-03 | 2.62E-02 | 5.61E-03 |
| fum | buni | 7.01E-05 | 1.31E+00 | 3.27E-01 | 3.22E-01 | 2.21E-02 | 2.62E-02 | 5.61E-03 |
| mal__L | pvulg | 4.99E-11 | 1.67E+00 | 5.40E-01 | 1.22E+00 | 6.68E-02 | 8.00E-02 | 2.79E-02 |
| mal__L | buni | 9.93E-07 | 4.06E+00 | 9.88E-01 | 9.96E-01 | 6.52E-02 | 2.62E-02 | 5.61E-03 |
| succ | pvulg | 4.98E-25 | 4.25E+00 | 1.18E+00 | 3.16E+00 | 2.70E-01 | 8.00E-02 | 2.79E-02 |
| lac__D | pvulg | 1.53E-10 | 5.52E+00 | 1.76E+00 | 4.10E+00 | 6.22E-01 | 8.00E-02 | 2.79E-02 |
| lac__D | buni | 8.97E-04 | 9.52E+00 | 1.35E+00 | 2.37E+00 | 2.25E-01 | 2.62E-02 | 5.61E-03 |

**Table 3.10.** *B. uniformis* and *P. vulgatus* secretion/uptake rates (mmol $gDW^{-1}$ $h^{-1}$) calculation with normalisation per gDW biomass in the 8-18 h growth period. Metabolite production and the corresponding p-values identified with ANCOVA.

Lactate, succinate, and malate were identified as the three main products under these conditions, showing the highest secretion rates and corresponding extracellular concentration profiles (Fig. 3.8, 3.9). The data indicate that *B. uniformis* and *P. vulgatus* differ not only in their growth patterns but also in their metabolic states. First, only *P. vulgatus* produces succinate, while in *B. uniformis* the change in succinate concentration is not significant (Fig. 3.8, 3.9). Second, although lactate is the major product for both species, its concentration profile shows a greater absolute change in *P. vulgatus*. However, because *P. vulgatus* grows much faster, biomass-normalised secretion is higher in *B. uniformis*, reversing the trend (Table 3.10, Fig. 3.9). Third, malate shows a similar change in concentration for both species, but biomass-normalised secretion is substantially higher in *B. uniformis* (Table 3.10, Fig. 3.9).

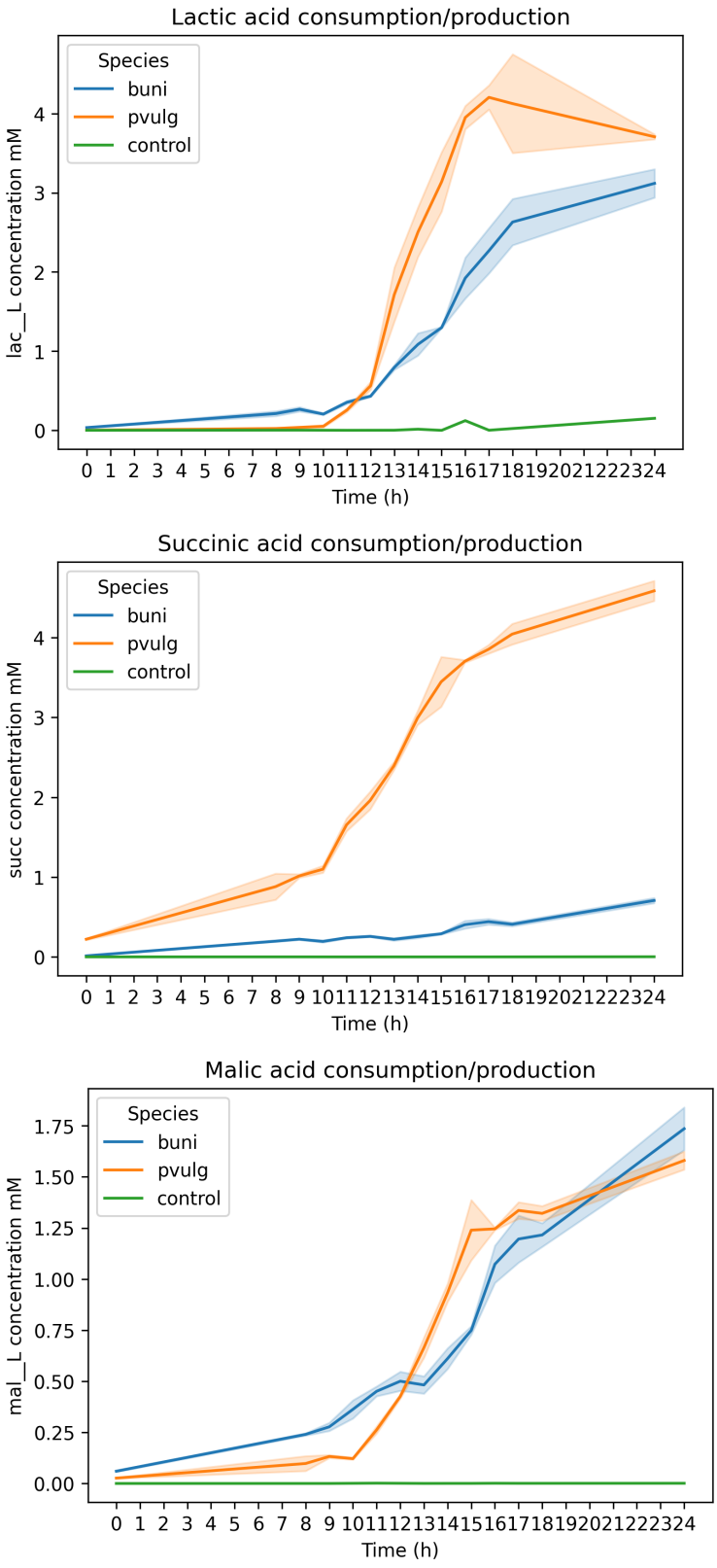


**Figure 3.9.** Profiles of the three top products, lactate, succinate and malate, for *B. uniformis* and *P. vulgatus* in the Varel-Bryant minimal medium. The concentration curves are plotted as mean of four biological replicates; shaded areas indicate standard deviation. The plot was generated by myself based on the experimental data provided by Nikita Denisov.

With experimental growth and metabolite exchange rates calculated, I set out to compare them with quantitative *B. uniformis* and *P. vulgatus* models flux simulation and curate the models accordingly.

To simulate growth in the *B. uniformis* and *P. vulgatus* models, I constrained them with the Varel–Bryant minimal media, allowing cysteine and methionine uptake for *P. vulgatus* according to experimental data, and none for *B. uniformis*. Since glucose uptake was not measured, I explored a plausible range of this boundary (between 0.5 and 10 mmol $gDW^{-1}$ $h^{-1}$) and its effect on the growth rate (Fig. 3.10). The initially curated core3 models of *B. uniformis* and *P. vulgatus* showed similar growth predictions, ranging from 0.06 $h^{-1}$ to 1.20 $h^{-1}$ depending on the glucose uptake (Fig. 3.10, left). Experimentally observed growth rates fell within this predicted range, proving that the models can provide a reasonable approximation of growth behaviour. Glucose uptake rates of 2 mmol $gDW^{-1}$ $h^{-1}$ for *B. uniformis* and 5 mmol $gDW^{-1}$ $h^{-1}$ for *P. vulgatus* resulted in optimized growth rates close to the experimental data (0.24 and 0.60 $h^{-1}$, respectively; Fig. 3.10, 3.11). A summary of the FBA simulations shows that the growth difference is primarily driven by glucose consumption, while *P. vulgatus* uptake of cysteine and methionine remains substantially lower than the allowed limits and contributes little to the overall uptake or secretion fluxes (Fig. 3.11).

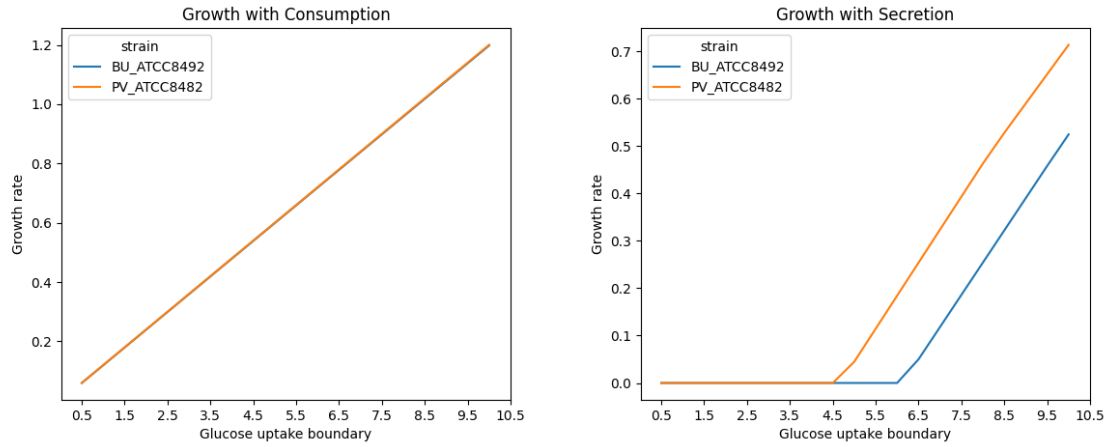


**Figure 3.10.** FBA-simulated growth rates ($h^{-1}$) of *B. uniformis* and *P. vulgatus* models in Varel–Bryant minimal medium with varying glucose uptake boundaries (mmol $gDW^{-1}$ $h^{-1}$), shown before (left) and after (right) curation for the secretion and constraining the secretion of three major products: lactate, malate, and succinate.

### Objective

1.0 Biomass = 0.23961373064365632

#### Uptake

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| ca2_e | EX_ca2_e | 0.02904 | 0 | 0.00% |
| cl_e | EX_cl_e | 0.001438 | 0 | 0.00% |
| cobalt2_e | EX_cobalt2_e | 7.049E-06 | 0 | 0.00% |
| cu2_e | EX_cu2_e | 0.0008731 | 0 | 0.00% |
| fe3_e | EX_fe3_e | 0.005644 | 0 | 0.00% |
| glc__D_e | EX_glc__D_e | 2 | 6 | 99.75% |
| h_e | EX_h_e | 10 | 0 | 0.00% |
| k_e | EX_k_e | 0.05391 | 0 | 0.00% |
| mg2_e | EX_mg2_e | 0.002396 | 0 | 0.00% |
| mn2_e | EX_mn2_e | 0.0001909 | 0 | 0.00% |
| nh4_e | EX_nh4_e | 1.839 | 0 | 0.00% |
| pheme_e | EX_pheme_e | 0.0008731 | 34 | 0.25% |
| pi_e | EX_pi_e | 1.236 | 0 | 0.00% |
| so4_e | EX_so4_e | 0.06465 | 0 | 0.00% |
| zn2_e | EX_zn2_e | 9.39E-05 | 0 | 0.00% |

#### Secretion

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| 4hba_c | DM_4hba_c | -6.297E-05 | 7 | 0.03% |
| co2_e | EX_co2_e | -1.432 | 1 | 99.97% |
| h2o_e | EX_h2o_e | -6.894 | 0 | 0.00% |
| o2_e | EX_o2_e | -0.001224 | 0 | 0.00% |
| ppi_e | EX_ppi_e | -0.5414 | 0 | 0.00% |

### Objective

1.0 Biomass = 0.6000953952449876

#### Uptake

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| ca2_e | EX_ca2_e | 0.07273 | 0 | 0.00% |
| cl_e | EX_cl_e | 0.0036 | 0 | 0.00% |
| cobalt2_e | EX_cobalt2_e | 1.765E-05 | 0 | 0.00% |
| cu2_e | EX_cu2_e | 0.002187 | 0 | 0.00% |
| cys__L_e | EX_cys__L_e | 0.07361 | 3 | 0.72% |
| fe2_e | EX_fe2_e | 0.01195 | 0 | 0.00% |
| fe3_e | EX_fe3_e | 0.002187 | 0 | 0.00% |
| glc__D_e | EX_glc__D_e | 5 | 6 | 97.65% |
| h_e | EX_h_e | 10 | 0 | 0.00% |
| k_e | EX_k_e | 0.135 | 0 | 0.00% |
| met__L_e | EX_met__L_e | 0.0853 | 5 | 1.39% |
| mg2_e | EX_mg2_e | 0.006 | 0 | 0.00% |
| mn2_e | EX_mn2_e | 0.0004781 | 0 | 0.00% |
| nh4_e | EX_nh4_e | 4.447 | 0 | 0.00% |
| pheme_e | EX_pheme_e | 0.002187 | 34 | 0.24% |
| pi_e | EX_pi_e | 3.004 | 0 | 0.00% |
| so4_e | EX_so4_e | 0.003 | 0 | 0.00% |
| zn2_e | EX_zn2_e | 0.0002352 | 0 | 0.00% |

#### Secretion

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| 4hba_c | DM_4hba_c | -0.0001577 | 7 | 0.03% |
| co2_e | EX_co2_e | -3.132 | 1 | 74.89% |
| h2o_e | EX_h2o_e | -16.71 | 0 | 0.00% |
| ppi_e | EX_ppi_e | -1.31 | 0 | 0.00% |
| succ_e | EX_succ_e | -0.2622 | 4 | 25.08% |

**Figure 3.11.** COBRApy summary of the FBA simulations for the initially curated *B. uniformis* (left) and *P. vulgatus* (right) models, with consumption constraints based on the Varel–Bryant minimal medium and experimental data on cysteine and methionine uptake, and without secretion constraints. Glucose uptake boundaries were set to 2 mmol $gDW^{-1}$ $h^{-1}$ for *B. uniformis* and 5 mmol $gDW^{-1}$ $h^{-1}$ for *P. vulgatus*, chosen as the nearest round values yielding biomass reaction fluxes close to the experimental data.

Even without any constraints on metabolite production, the *B. uniformis* and *P. vulgatus* models showed a difference in the secretion of one metabolite, succinate (Fig. 3.11). This aligns with experimental data, where succinate is the second major product of *P. vulgatus*, whereas it is not significantly produced by *B. uniformis*, and suggests that the models already capture differences in network topology that lead to distinct succinate secretion fluxes.

As succinate is an intermediate in the TCA cycle, I examined this key pathway in *B. uniformis* and *P. vulgatus* using GEMsembler pathway maps (Fig. 3.12, 3.13). These maps indicate that most TCA reactions are supported by at least three draft models and

some genetic evidence. In *B. uniformis*, the main uncertainty concerns the AKGDH (2-oxoglutarate dehydrogenase) reaction (Fig. 3.12). However, this reaction is supported by two draft models with assigned GPRs and was already incorporated into the core3 model during curation, resulting in a fully functional TCA cycle for *B. uniformis*. In contrast, *P. vulgatus* shows greater uncertainty at the SUCOAS (succinyl-CoA synthetase) reaction, which was gap-filled by two models (gapseq and AGORA) but lacks genetic evidence (Fig. 3.13).

Therefore, the succinyl-CoA synthetase reaction, which directly involves succinate, is absent in the curated core3 *P. vulgatus* model. Meanwhile, the *B. uniformis* model includes this reaction, as it is part of core3, and the CarveMe model identified a candidate gene (BU_ATCC8492_00941) for its enzyme (Fig. 3.13).

The absence of SUCOAS in *P. vulgatus* explains the observed differences in succinate secretion between the two models, which align with experimental data. Indeed, in *B. uniformis*, the TCA flux through the reductive branch flows from succinate to succinyl-CoA and 2-oxoglutarate, whereas in *P. vulgatus* these reactions are missing, therefore succinate needs to be secreted. However, the underlying biological reasons require further investigation, including examining the presence of unusual succinyl-CoA synthetase variants, other pathway differences, or enzyme activity regulation.

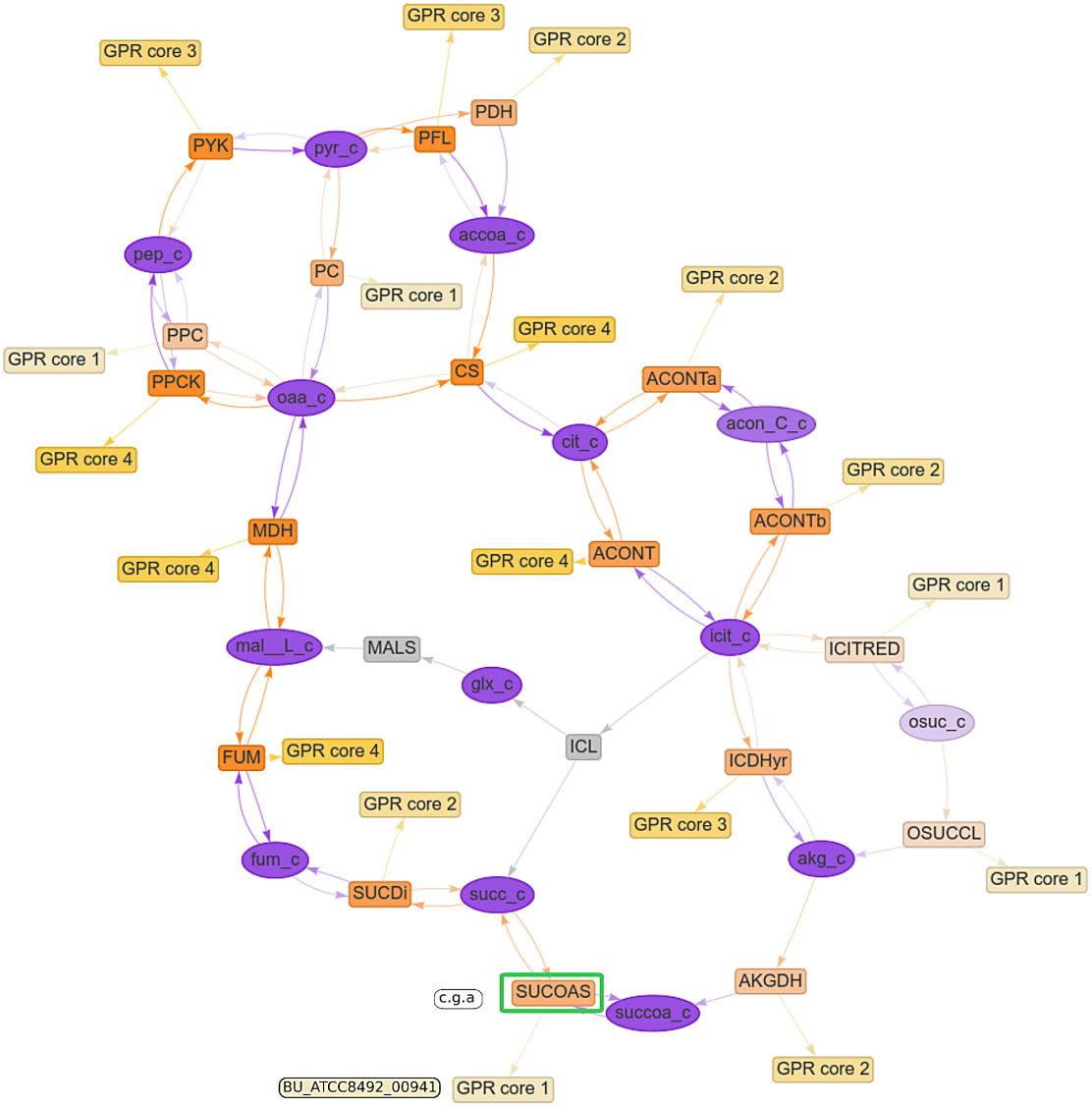


**Figure 3.12.** TCA pathway map in the *B. uniformis* supermodel showing agreement among three draft models (CarveMe, gapseq, and AGORA) on the reaction SUCOAS (succinyl-CoA synthetase, GPR: BU_ATCC8492_00941), which is included in the curated core3 model. This reaction is proposed as a hypothesis to explain the observed difference in succinate secretion (*B. uniformis* does not produce succinate).

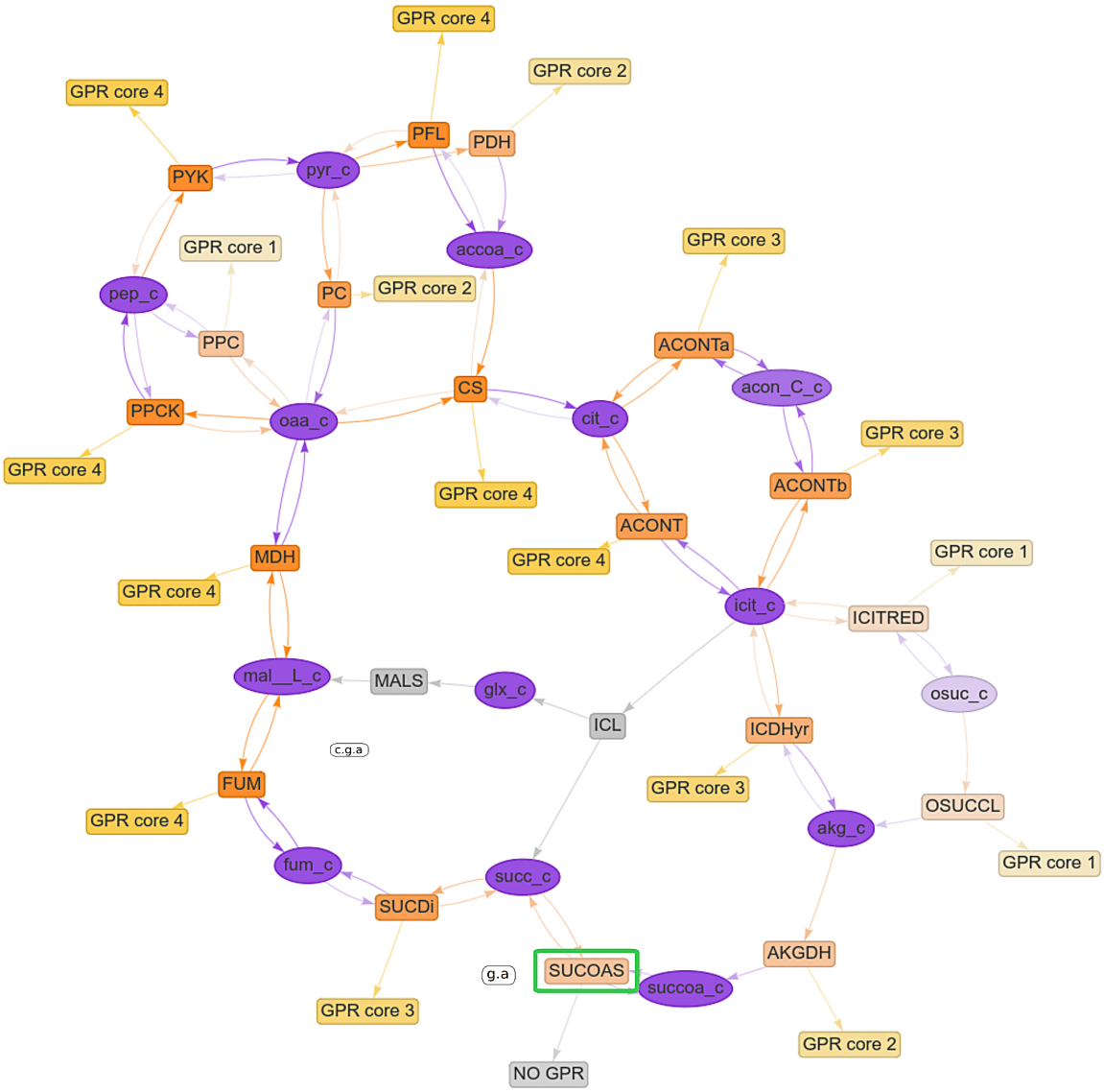


**Figure 3.13.** TCA pathway map in the *P. vulgatus* supermodel showing agreement among two draft models (gapseq, and AGORA) on the reaction SUCOAS (succinyl-CoA synthetase) without any GPR assigned, which is excluded in the curated core3 model. This reaction is proposed as a hypothesis to explain the observed difference in succinate secretion (*P. vulgatus* produces succinate).

Next, I investigated the remaining two major products that were not produced by the models initially. D-lactate could be secreted by both *B. uniformis* and *P. vulgatus* core3 models, so this metabolite did not require curation of network topology. In contrast, L-malate could not be secreted, even though intracellular L-malate could be produced by both species, indicating that transport curation was necessary. The GEMsembler-generated map of the extracellular L-malate neighborhood revealed three potential transport routes, all predicted by a single draft model, involving the same putative transporters (Fig. 3.14). Therefore, I selected the simplest transport reaction, without additional ions or compartment transitions, to add to the core3 models and restore L-malate secretion.

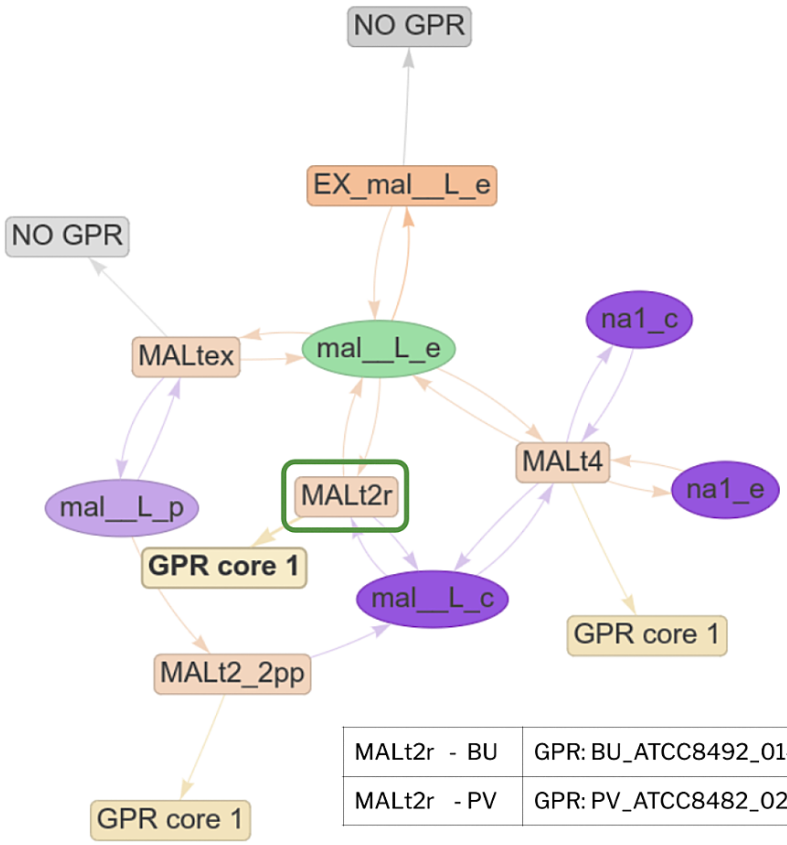


**Figure 3.14.** Neighborhood map for extracellular malate from GEMsembler, showing the curated transport reaction (MALt2r), which, together with the exchange reaction, restores malate secretion in *B. uniformis* and *P. vulgatus*.

With the network topology of the curated *B. uniformis* and *P. vulgatus* models aligned with the experimentally identified exchange metabolites, I next evaluated whether the observed extracellular metabolite fluxes, together with the growth patterns, could be captured by FBA simulations. After introducing secretion constraints, the glucose uptake boundary shifted (Fig. 3.10, right), requiring at least 6.5 mmol $gDW^{-1}$ $h^{-1}$ for *B. uniformis* and 5 mmol $gDW^{-1}$ $h^{-1}$ for *P. vulgatus* to obtain maximal FBA solutions for biomass flux comparable to the experimentally observed values. Under both uptake and secretion constraints, the models began to display differences in predicted growth rates, with *B. uniformis* growing more slowly at the same glucose uptake rate (Fig. 3.10, right). Consequently assuming an equal glucose uptake boundary of 8 mmol $gDW^{-1}$ $h^{-1}$ for both species resulted in optimized biomass production rates close to the experimental growth data: 0.25 $h^{-1}$ for *B. uniformis* and 0.46 $h^{-1}$ for *P. vulgatus* (Fig. 3.15).

## Objective

1.0 Biomass = 0.2540729289491626

### Uptake

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| ca2_e | EX_ca2_e | 0.03079 | 0 | 0.00% |
| cl_e | EX_cl_e | 0.001524 | 0 | 0.00% |
| cobalt2_e | EX_cobalt2_e | 7.475E-06 | 0 | 0.00% |
| cu2_e | EX_cu2_e | 0.0009258 | 0 | 0.00% |
| fe3_e | EX_fe3_e | 3.194 | 0 | 0.00% |
| glc__D_e | EX_glc__D_e | 8 | 6 | 99.93% |
| k_e | EX_k_e | 0.05716 | 0 | 0.00% |
| mg2_e | EX_mg2_e | 0.00254 | 0 | 0.00% |
| mn2_e | EX_mn2_e | 0.0002024 | 0 | 0.00% |
| nh4_e | EX_nh4_e | 1.95 | 0 | 0.00% |
| pheme_e | EX_pheme_e | 0.0009258 | 34 | 0.07% |
| pi_e | EX_pi_e | 2.066 | 0 | 0.00% |
| so4_e | EX_so4_e | 0.4462 | 0 | 0.00% |
| zn2_e | EX_zn2_e | 9.956E-05 | 0 | 0.00% |

### Secretion

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| 4hba_c | DM_4hba_c | -6.677E-05 | 7 | 0.00% |
| fe2_e | EX_fe2_e | -3.188 | 0 | 0.00% |
| h2o_e | EX_h2o_e | -7.647 | 0 | 0.00% |
| h2s_e | EX_h2s_e | -0.3776 | 0 | 0.00% |
| lac__D_e | EX_lac__D_e | -8.169 | 3 | 66.60% |
| mal__L_e | EX_mal__L_e | -3.072 | 4 | 33.40% |
| ppi_e | EX_ppi_e | -0.9517 | 0 | 0.00% |

## Objective

1.0 Biomass = 0.4631316710626314

### Uptake

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| ca2_e | EX_ca2_e | 0.05613 | 0 | 0.00% |
| cl_e | EX_cl_e | 0.002778 | 0 | 0.00% |
| cobalt2_e | EX_cobalt2_e | 1.363E-05 | 0 | 0.00% |
| cu2_e | EX_cu2_e | 0.001688 | 0 | 0.00% |
| cys__L_e | EX_cys__L_e | 0.05681 | 3 | 0.35% |
| fe2_e | EX_fe2_e | 0.009221 | 0 | 0.00% |
| fe3_e | EX_fe3_e | 0.001688 | 0 | 0.00% |
| glc__D_e | EX_glc__D_e | 8 | 6 | 98.85% |
| h_e | EX_h_e | 38.56 | 0 | 0.00% |
| k_e | EX_k_e | 0.1042 | 0 | 0.00% |
| met__L_e | EX_met__L_e | 0.06583 | 5 | 0.68% |
| mg2_e | EX_mg2_e | 0.00463 | 0 | 0.00% |
| mn2_e | EX_mn2_e | 0.000369 | 0 | 0.00% |
| nh4_e | EX_nh4_e | 3.432 | 0 | 0.00% |
| pheme_e | EX_pheme_e | 0.001688 | 34 | 0.12% |
| pi_e | EX_pi_e | 2.548 | 0 | 0.00% |
| so4_e | EX_so4_e | 0.07452 | 0 | 0.00% |
| zn2_e | EX_zn2_e | 0.0001815 | 0 | 0.00% |

### Secretion

| Metabolite | Reaction | Flux | C-Number | C-Flux |
|---|---|---|---|---|
| 4hba_c | DM_4hba_c | -0.0001217 | 7 | 0.00% |
| h2o_e | EX_h2o_e | -14.59 | 0 | 0.00% |
| h2s_e | EX_h2s_e | -0.0722 | 0 | 0.00% |
| lac__D_e | EX_lac__D_e | -3.76 | 3 | 40.18% |
| mal__L_e | EX_mal__L_e | -1.127 | 4 | 16.06% |
| ppi_e | EX_ppi_e | -1.126 | 0 | 0.00% |
| succ_e | EX_succ_e | -3.071 | 4 | 43.76% |

**Figure 3.15.** COBRApy summary of the FBA simulations for the finally curated *B. uniformis* (left) and *P. vulgatus* (right) models, with consumption constraints based on the Varel–Bryant minimal medium and experimental data on cysteine and methionine uptake, as well as on lactate, succinate and malate secretion. Glucose uptake boundaries were set to 8 mmol $gDW^{-1}$ $h^{-1}$ for both *B. uniformis* and *P. vulgatus*, chosen as the single nearest round value yielding biomass reaction fluxes close to the experimental data.

Growth reduction under all secretion constraints occurs primarily because enforcing the secretion of the major product, lactate, redirects a substantial portion of carbon derived from glucose and glycolysis into the D-lactate dehydrogenase (LDH_D) reaction, leaving less carbon available for central carbon metabolism and, consequently, growth. At a glucose uptake rate of 8 mmol $gDW^{-1}$ $h^{-1}$, 56.95% of pyruvate is consumed for lactate synthesis

in *B. uniformis* (38.73% enters the TCA cycle), while in *P. vulgatus* this fraction is roughly half of that in *B. uniformis*, at 27.76% (62.62% enters the TCA cycle).

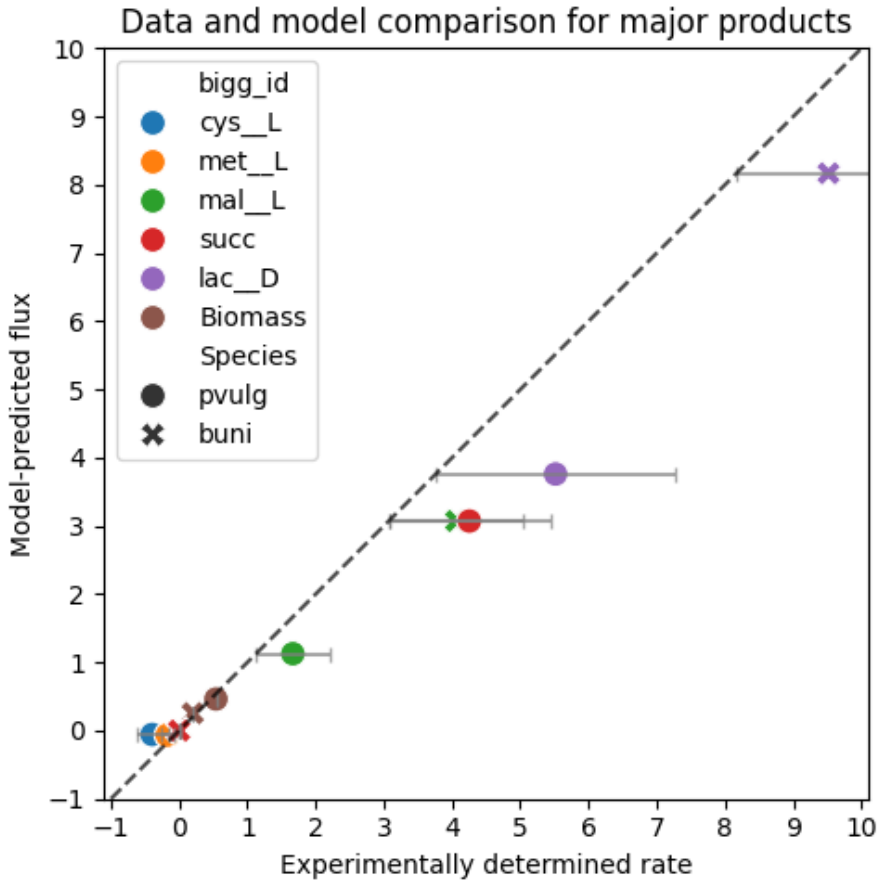


**Figure 3.16.** Comparison of experimentally determined uptake and secretion rates of major products, obtained from time-course extracellular metabolomics data (x-axis), with corresponding flux predictions in the final curated models of *B. uniformis* and *P. vulgatus* (y-axis). Error bars represent the standard deviations of four replicates. Glucose uptake was limited to 8 mmol $gDW^{-1}$ $h^{-1}$. for both species, and the corresponding predicted biomass flux (y-axis) is shown together with the experimentally measured growth rate (x-axis).

In summary, the final genome-scale metabolic models of *B. uniformis* and *P. vulgatus*, curated using a consensus-based approach, reproduce all experimental observations taken together (Fig. 3.16), including major metabolite uptake and secretion fluxes within the standard deviation range of the data and the experimentally determined growth rates.

## 3.4 Discussion

Genome-scale metabolic model reconstruction and refinement is a highly iterative process, with each step building on previous discoveries and earlier versions of the model [29, 31]. In this project, I initiated the *de novo* reconstruction of genome-scale metabolic models for two key human gut bacteria and model organisms to be, *B. uniformis* and *P. vulgatus*.

This *de novo* reconstruction was based on a comparison of draft models, automatically derived from bacterial genomes using different algorithms. Applying a consensus-based approach using the GEMsembler workflow that I previously developed, enabled me to make initial decisions about the overall model structure while simultaneously identifying nodes of uncertainty in the metabolic networks and prioritizing reactions for future validation.

Validation of uncertain reactions in future stages of model curation can take many paths, ranging from additional computational approaches to identify enzymes and their functions to experimental confirmation of predicted reactions. One promising computational strategy is to move beyond sequence-based annotation and incorporate protein structure-based predictions, which have been made significantly more reliable with neural-network based models, such as AlphaFold2 [139].

Building on computational predictions, experimental validation provides the strongest evidence for confirming model reactions. A particularly effective approach, directly linking genes to reactions in the model, is based on the gain-of-function strategy [140]. In this method, a reaction of interest is first identified in a model organism such as *E. coli*, along with a mutant strain lacking the enzyme required for this reaction, which results in loss of viability under specific conditions. Introducing a candidate gene from the organism of interest (ex. *B. uniformis* or *P. vulgatus*) into the mutant and observing restoration of the growth provides evidence that the gene encodes the predicted function. If no clear candidate is available, a gain-of-function screen can be performed. In this case a gain-of-function library, composed of plasmids carrying random genomic fragments, can be used to identify positive hits restoring the phenotype of the recipient strain and thus pinpoint potential genes for further investigation.

However, despite its value, the gain-of-function approach has a fundamental limitation: only a small subset of reactions can feasibly be tested. This constraint arises from differences between the metabolic networks of the model organism and the target species, ambiguity in gene-to-reaction relationships, like isoenzymes or protein complexes, and variability in mutant phenotypes, which can range from complete growth loss to no detectable deviation from the wild type.

In addition to uncertain reactions, the biomass reaction composition is a crucial element of genome-scale metabolic models that requires validation. For the *B. uniformis* and *P. vulgatus* models, a consensus-based biomass reaction derived from draft models was used to supplement a curated formulation from the related species *B. thetaiotaomicron*, resulting in a biomass composition that incorporates both high-confidence genus-level and species-specific elements. While inheriting biomass reactions from better-studied organisms is a common practice [132], achieving model-organism-level accuracy for *B. uniformis* and *P. vulgatus* will require experimental validation, including verification

of uncertain precursors, determination of the stoichiometry of biomass components, and calculation of growth-associated (GAM) and non-growth-associated maintenance (NGAM) parameters for these species [141].

Since glucose consumption was not quantified in the growth experiments available for this curation, the growth rate predictions of the *B. uniformis* and *P. vulgatus* models cannot yet be fully evaluated. If the experimentally measured glucose uptake substantially deviates from the model-derived value of 8 mmol $gDW^{-1}$ $h^{-1}$, reassessment of the biomass reaction parameters, including GAM, NGAM, and stoichiometric composition, will be a priority. In particular, the GAM parameter requires confirmation, as it has a strong influence on biomass yield [142]. Another reason for careful re-evaluation is that the GAM value reported for *B. thetaiotaomicron* (18.526 mmol $gDW^{-1}$) [45] is markedly lower than the typical value included in biomass templates (~55 mmol $gDW^{-1}$) and falls below the lower bound for the GAM demand of *E. coli* (22.363 mmol $gDW^{-1}$) [143]. Finally, it is essential to consider that the steady-state assumption of flux balance analysis is most appropriate for chemostat conditions, and that the use of batch culture experiments may introduce additional discrepancies in growth rate estimates.

As this work represents the first stage of reconstructing genome-scale metabolic models for *B. uniformis* and *P. vulgatus*, the reconstruction was based on data from highly minimal and defined media, allowing precise determination of nutritional constraints and identification of major metabolic trends. However, this work can be further expanded, particularly in the comparative analysis of the two species, by exploring *P. vulgatus*'s superior growth and broader metabolic activity, as well as *B. uniformis*'s greater production capacity for major metabolites. Modeling approaches such as flux variability or sampling analysis could be applied to investigate differences in metabolic fluxes and pathway utilisation between the two species. In addition, studies of enzymatic activity and regulatory mechanisms may provide further insight into the observed phenotypes.

A common approach to extend minimal media conditions is to investigate the ability of bacteria to utilise a variety of carbon sources beyond glucose. For *B. uniformis* and *P. vulgatus*, this is an important direction, although large variation is not expected, as *Bacteroides* are known for their broad saccharide utilisation capabilities [112]. In the future, such metabolic capacities should also be incorporated and investigated in the models.

In addition to reaction validation, the genes included in the genome-scale metabolic models of *B. uniformis* and *P. vulgatus* should also be curated with experimental data. The most common and high-throughput source for this is gene essentiality datasets, which can be directly compared with model predictions of essential genes. It would be particularly useful to apply the same approach I used in Section 2.3.6 for *E. coli* to refine the existing

gene–reaction rules of *B. uniformis* and *P. vulgatus*, taking into account their variations in the corresponding supermodels.

Overall, while minimal media provided a controlled starting point for model development, their restrictive nature inevitably limits the range of metabolic functions that can be assessed. Expanding to more favorable growth conditions and incorporating strain-level diversity in *B. uniformis* and *P. vulgatus* will be essential for advancing from initial curated reconstructions to comprehensive, biologically representative genome-scale metabolic models, thereby bringing these species closer to established model organisms.

# Conclusion

My PhD began as an effort to reconstruct genome-scale metabolic models (GEMs) for several non-model human gut bacteria in order to probe community metabolic interactions. Early on, it became clear that *de novo* reconstruction of GEMs for non-model organisms presents a substantive challenge: different pipelines yield substantially distinct networks, and the available means of comparison were limited and fragmented, with no tool available that treated models as coherent systems rather than isolated components. Motivated by this gap, I embraced the challenge and focused on filling the niche created by the lack of holistic approaches to GEM comparison. The guiding idea is that comparison is a cornerstone of scientific inquiry, and implementing model-level comparison in GEM research enables the formulation of hypotheses about metabolic differences between models, the identification of complementary functions, and systematic reasoning about confidence in network elements. In the context of GEM reconstruction, this idea underpins a consensus approach, turning reconstruction-tool disagreement from a source of noise into a source of evidence and testable alternatives.

I implemented GEM comparison and a consensus model reconstruction approach in the GEMsembler Python package, which keeps track of all network elements and their origin via supermodel structure that enables the various comparisons and combinations. Moreover, GEMsembler also provides tools to explore metabolic networks: from systematic assessment of biomass-component biosynthesis to enumeration of the neighbourhood of a metabolite of interest. I believe that GEMsembler's systematic approach to building consensus models and its flexibility and accessibility will help researchers at any stage, even with minimal programming skills, who need to build, analyse, or curate GEMs for their organisms of interest. GEMsembler's functionality can also be generalised and plugged into other computational pipelines and research settings that address different aspects of bacterial metabolism. If the systems and computational biology community adopts GEMsembler widely, it will aid in assessing and resolving uncertainties in metabolic models and support the reconstruction of models that are more comprehensive, concise, and biologically informed.

Case study to demonstrate GEMsembler functionality includes curation of models for *Lactiplantibacillus plantarum* and *Escherichia coli*, based on a consensus of four automatically reconstructed GEMs and begins with a systematic assessment of biomass-component biosynthesis. The curation workflow pinpointed low-confidence reactions missing from high-confidence networks, and adding reactions originating from multiple reconstruction sources restored coherent pathways. For *L. plantarum*, this process yielded a consensus model with more accurate auxotrophy predictions than the reference gold-standard model and made it possible to trace the improvement to specific biosynthetic routes. For *E. coli*, evaluation against gene-essentiality data also showed that the consensus-curated model outperformed the reference gold standard. In addition, strategies to combine alternative gene–protein–reaction (GPR) rules drawn from the input and consensus models improved gene-essentiality predictions across all models, including the curated gold-standard ones.

With the consensus approach established, I next applied it to the *de novo* reconstruction of two of the most abundant yet underexplored human gut bacteria, *Bacteroides uniformis* and *Phocaeicola vulgatus*. I built and curated their models by balancing confidence in network features against identified biosynthetic capability. The resulting first-iteration curated models reproduced characteristic growth patterns and biosynthetic product profiles. With experimentally measured secretion-rate constraints enforced for major products (lactate and malate for *B. uniformis*; lactate, malate, and succinate for *P. vulgatus*), the models' topology supported the experimentally observed growth behaviour, with *B. uniformis* growing more slowly. Although a more complete understanding of *B. uniformis* and *P. vulgatus* metabolic capacities will require further investigation, these results already show that the approach works well for relatively poorly characterised gut bacteria.

Overall, this thesis establishes a consensus approach for assessing and resolving uncertainty in genome-scale metabolic models. By comparing and integrating alternative reconstructions into agreement-driven consensus models, it converts inter-tool disagreement into evidence and testable biological alternatives, prioritises targeted curation, increases confidence in the network content and improves the models' predictive reliability and interpretability. In doing so, it advances systems-level understanding of microbial metabolism and provides a practical foundation for robust, concise, and biologically informed GEMs.

# Acknowledgements

It was a challenging period for different reasons, and it is not the best time to write acknowledgements when everything already feels blurred at the end of the thesis. Still, I am truly grateful for the opportunity to do this PhD, in this place with these people.

One person I especially want to thank is my PI, Maria Zimmermann-Kogadeeva, for being consistently supportive throughout these years. I appreciate having a supervisor who is truly present, someone I can talk to, who makes time for meetings, listens, and is willing to think through ideas together. Having the trust and space to shape my own work made a real difference, and I'm very thankful for that.

I also want to thank the members of my TAC, Ursula Kummer, Mikhail Savitski and Georg Zeller, for their guidance and the scientific discussions.

I am thankful to my lab group for the atmosphere we shared. I especially want to mention Bartosz, who kept everything running when things became technically complicated. To all members of the group, thank you for being part of this work.

I would also like to thank my collaborators, from the Michael Zimmermann group, with whom I shared the experimental part, and from the Flagship Project, for the biological discussions and perspectives.

I am also grateful to my friends and family, near and extended, for their encouragement and support beyond the academic side of this journey. They helped, in one way or another, and I am thankful for their presence throughout this time.

A special thanks goes to my husband, Gleb. Throughout these challenging times, he supported me in every possible way, far beyond the academic side, and even within it when it was needed. It is difficult to imagine any part of this journey where he was not present or willing to help, and I am deeply grateful for that.

I could have added many more names here, but at some point, this has to come to an end. Those I did not mention are no less important, and I remain grateful for their part in this long process. Thank you all!

# Appendices

## Appendix I

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_LP | ca_LP | ga_LP | mo_LP | Decision |
|---|---|---|---|---|---|---|---|---|---|
| ACP_c | r | Core4 | Core3 | yes | + | - | + | + | no |
| ala__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| amet_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| arg__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| asn__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| asp__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| atp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ca2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cl_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| coa_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cobalt2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ctp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cu2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cys__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| datp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dctp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dgtp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dttp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fad_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fe2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fe3_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gln__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| glu__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gly_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gtp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| h2o_c | r | Core4 | Core3 | yes | + | + | + | - | yes |
| his__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ile__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| k_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| leu__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| lys__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| met__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| mg2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_LP | ca_LP | ga_LP | mo_LP | Decision |
|---|---|---|---|---|---|---|---|---|---|
| mn2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| nad_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| nadp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| phe__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| pro__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| pydx5p_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ribflv_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ser__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| so4_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| tcam_c | r | Core3 | Core3 | yes | + | - | + | + | yes |
| thr__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| trp__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| tyr__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| udcpdp_c | r | Core2 | Core1 | no | + | - | - | - | yes |
| utp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| val__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| zn2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| 10fthf_c | r | Core4 | Core3 | yes | - | + | + | + | yes |
| gtca1_45_BS_c | r | Core1 | Core1 | no | - | + | - | - | yes |
| gtca2_45_BS_c | r | Core1 | Core1 | no | - | + | - | - | yes |
| gtca3_45_BS_c | r | Core1 | Core1 | no | - | + | - | - | yes |
| lipo1_24_BS_c | r | Core1 | Core1 | no | - | + | - | - | no |
| lipo2_24_BS_c | r | Core1 | Core1 | no | - | + | - | - | no |
| lipo3_24_BS_c | r | Core1 | Core1 | no | - | + | - | - | no |
| lipo4_24_BS_c | r | Core1 | Core1 | no | - | + | - | - | no |
| mlthf_c | r | Core4 | Core1 | no | - | + | - | - | no |
| mql8_c | r | Core4 | Core1 | no | - | + | - | - | no |
| peptido_BS_c | r | Core1 | Core1 | no | - | + | - | - | yes |
| thf_c | r | Core4 | Core3 | yes | - | + | + | + | yes |
| thmpp_c | r | Core4 | Core3 | yes | - | + | + | + | yes |
| gthrd_c | r | Core4 | Core2 | q | - | - | + | + | no |
| 5mthf_c | r | Core4 | Core2 | q | - | - | + | + | no |
| pheme_c | r | Core3 | Core2 | q | - | - | + | + | no |
| ptrc_c | r | Core4 | Core2 | q | - | - | + | + | yes |
| spmd_c | r | Core4 | Core2 | q | - | - | + | + | yes |
| adocbl_c | r | Core2 | Core2 | q | - | - | + | + | no |
| sheme_c | r | Core2 | Core2 | q | - | - | + | + | no |
| pg180_c | r | Core4 | Core2 | q | - | - | + | + | no |
| pe180_c | r | Core3 | Core2 | q | - | - | + | + | no |
| q8_c | r | Core4 | Core1 | no | - | - | - | + | no |
| mqn8_c | r | Core4 | Core1 | no | - | - | - | + | no |
| h_c | r | Core4 | Core1 | no | - | - | - | + | no |
| 2dmmq8_c | r | Core4 | Core1 | no | - | - | - | + | no |
| adp_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| apoACP_c | p | Core4 | Core3 | yes | + | - | + | + | no |
| biomass_c | p | Core1 | Core1 | no | + | - | - | - | no |

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_LP | ca_LP | ga_LP | mo_LP | Decision |
|---|---|---|---|---|---|---|---|---|---|
| h_c | p | Core4 | Core3 | yes | + | + | + | - | yes |
| pi_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| ppi_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| dmbzid_c | p | Core2 | Core2 | q | - | - | + | + | no |
| cbi_c | p | Core2 | Core2 | q | - | - | + | + | no |
| h2o_c | p | Core4 | Core1 | no | - | - | - | + | no |

**Table 5.1.** Curating the Biomass reaction for *L. plantarum*.

# Appendix II

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_EC | ca_EC | ga_EC | mo_EC | Decision |
|---|---|---|---|---|---|---|---|---|---|
| 10fthf_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ala__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| amet_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| arg__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| asn__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| asp__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| atp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ca2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cl_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| coa_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cobalt2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ctp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cu2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| cys__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| datp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dctp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dgtp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| dttp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fad_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fe2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| fe3_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gln__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| glu__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gly_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| gtp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| h2o_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| his__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ile__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| k_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| leu__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| lys__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| met__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_EC | ca_EC | ga_EC | mo_EC | Decision |
|---|---|---|---|---|---|---|---|---|---|
| mg2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| mn2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| nad_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| nadp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| phe__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| pheme_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| pro__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| pydx5p_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ribflv_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| ser__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| sheme_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| so4_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| thf_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| thmpp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| thr__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| trp__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| tyr__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| utp_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| val__L_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| zn2_c | r | Core4 | Core4 | yes | + | + | + | + | yes |
| 2dmmq8_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| 5mthf_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| ACP_c | r | Core4 | Core3 | yes | + | - | + | + | no |
| adocbl_c | r | Core3 | Core3 | yes | + | - | + | + | no |
| colipa_c | r | Core3 | Core3 | yes | + | - | + | + | no |
| gthrd_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| mqn8_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| pe180_c | r | Core4 | Core3 | yes | + | - | + | + | no |
| ptrc_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| spmd_c | r | Core4 | Core3 | yes | + | - | + | + | yes |
| pg180_c | r | Core3 | Core2 | q | + | - | - | + | no |
| q8_c | r | Core4 | Core2 | q | + | - | - | + | no |
| clpn180_c | r | Core1 | Core1 | no | + | - | - | - | no |
| dtdprmn_c | r | Core3 | Core1 | no | + | - | - | - | no |
| kdo2lipid4L_c | r | Core3 | Core1 | no | + | - | - | - | no |
| udcpdp_c | r | Core2 | Core1 | no | + | - | - | - | yes |
| kdo2lipid4_p | r | Core1 | Core1 | no | - | + | - | - | no |
| mlthf_c | r | Core4 | Core1 | no | - | + | - | - | no |
| mql8_c | r | Core4 | Core1 | no | - | + | - | - | no |
| murein5px4p_p | r | Core1 | Core1 | no | - | + | - | - | no |
| pe160_c | r | Core4 | Core1 | no | - | + | - | - | yes |
| pe160_p | r | Core1 | Core1 | no | - | + | - | - | no |
| pe161_c | r | Core4 | Core1 | no | - | + | - | - | no |
| pe161_p | r | Core1 | Core1 | no | - | + | - | - | no |
| adp_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| h_c | p | Core4 | Core4 | yes | + | + | + | + | yes |

| Metabolite | Type | Met_conf. | Biomass_conf. | Status | ag_EC | ca_EC | ga_EC | mo_EC | Decision |
|---|---|---|---|---|---|---|---|---|---|
| pi_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| ppi_c | p | Core4 | Core4 | yes | + | + | + | + | yes |
| apoACP_c | p | Core4 | Core3 | yes | + | - | + | + | no |
| cbi_c | p | Core3 | Core3 | yes | + | - | + | + | no |
| dmbzid_c | p | Core3 | Core3 | yes | + | - | + | + | no |
| biomass_c | p | Core1 | Core1 | no | + | - | - | - | no |

**Table 5.2.** Curating the Biomass reaction for *E. coli*.

# Appendix III

Resourcers for GEMsembler framework.

GEMsembler source code is available at: https://github.com/zimmermann-kogadeeva-group/GEMsembler

GEMsembler tutorials and example notebooks are available at: https://grp-zimmermann-kogadeeva.embl-community.io/gemsembler/, and https://git.embl.org/grp-zimmermann-kogadeeva/GEMsembler/-/blob/master/docs/tutorial.ipynb.

Code and data required to reproduce results from Chapter II GEMsembler, as well as tables and interactive maps used, are available at: https://git.embl.de/grp-zimmermann-kogadeeva/GEMsembler_paper and on Zenodo: https://doi.org/10.5281/zenodo.16529342.

# Appendix IV

Authorship declaration and use of AI.

I confirm that this dissertation is my own work and has not been submitted, in whole or in part, for a degree at any other institution. Contributions from collaborators are explicitly acknowledged in the relevant parts of the text.

The writing, editing, and formatting of this dissertation were partially supported by AI tools, specifically OpenAI's ChatGPT, Google's Gemini, and Perplexity.ai. These tools were used solely for rephrasing, improving clarity, summarising notes, and maintaining stylistic consistency throughout the text. For the visual component, AI assistance was also used to suggest LaTeX formatting and code for generating plots, without involvement in any aspect of figure design. All scientific ideas, data analyses, interpretations, and conclusions are my own.

# Bibliography


[1] Ruth E. Ley, Daniel A. Peterson, and Jeffrey I. Gordon. Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell*, 124(4):837–848, February 2006.

[2] Monika Yadav, Manoj Kumar Verma, and Nar Singh Chauhan. A review of metabolic potential of human gut microbiome in human nutrition. *Arch Microbiol*, 200(2):203–217, March 2018.

[3] Francesca Turroni, Angela Ribbera, Elena Foroni, Douwe Van Sinderen, and Marco Ventura. Human gut microbiota and bifidobacteria: from composition to functionality. *Antonie van Leeuwenhoek*, 94(1):35–50, June 2008.

[4] Juliana Durack and Susan V. Lynch. The gut microbiome: Relationships with disease and opportunities for therapy. *Journal of Experimental Medicine*, 216(1):20–40, January 2019. Publisher: Rockefeller University Press.

[5] Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol*, 19(1):55–71, January 2021. Publisher: Nature Publishing Group.

[6] Piero Portincasa, Leonilde Bonfrate, Mirco Vacca, Maria De Angelis, Ilaria Farella, Elisa Lanza, Mohamad Khalil, David Q.-H. Wang, Markus Sperandio, and Agostino Di Ciaula. Gut Microbiota and Short Chain Fatty Acids: Implications in Glucose Homeostasis. *IJMS*, 23(3):1105, January 2022.

[7] Harry J. Flint, Edward A. Bayer, Marco T. Rincon, Raphael Lamed, and Bryan A. White. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol*, 6(2):121–131, February 2008.

[8] Jean Guy LeBlanc, Christian Milani, Graciela Savoy De Giori, Fernando Sesma, Douwe Van Sinderen, and Marco Ventura. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current Opinion in Biotechnology*, 24(2):160–168, April 2013.

[9] Kei E Fujimura, Alexandra R Sitarik, Suzanne Havstad, Din L Lin, Sophia Levan, Douglas Fadrosh, Ariane R Panzer, Brandon LaMere, Elze Rackaityte, Nicholas W Lukacs, Ganesa Wegienka, Homer A Boushey, Dennis R Ownby, Edward M Zoratti, Albert M Levin, Christine C Johnson, and Susan V Lynch. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat Med*, 22(10):1187–1191, October 2016.

[10] Denise Kelly, Jamie I Campbell, Timothy P King, George Grant, Emmelie A Jansson, Alistair G P Coutts, Sven Pettersson, and Shaun Conway. Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-γ and RelA. *Nat Immunol*, 5(1):104–112, January 2004.

[11] Julia Montenegro, Anissa M. Armet, Benjamin P. Willing, Edward C. Deehan, Priscila G. Fassini, João F. Mota, Jens Walter, and Carla M. Prado. Exploring the Influence of Gut Microbiome on Energy Metabolism in Humans. *Advances in Nutrition*, 14(4):840–857, July 2023.

[12] Li Fu, Ancha Baranova, Hongbao Cao, and Fuquan Zhang. Gut microbiome links obesity to type 2 diabetes: insights from Mendelian randomization. *BMC Microbiol*, 25(1):253, April 2025.

[13] Zhilin Zhang, Huan Zhang, Tian Chen, Lin Shi, Daorong Wang, and Dong Tang. Regulatory role of short-chain fatty acids in inflammatory bowel disease. *Cell Commun Signal*, 20(1):64, December 2022.

[14] Kathleen Machiels, Marie Joossens, João Sabino, Vicky De Preter, Ingrid Arijs, Venessa Eeckhaut, Vera Ballet, Karolien Claes, Filip Van Immerseel, Kristin Verbeke, Marc Ferrante, Jan Verhaegen, Paul Rutgeerts, and Séverine Vermeire. A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut*, 63(8):1275–1283, August 2014.

[15] Filip Van Immerseel, Richard Ducatelle, Martine De Vos, Nico Boon, Tom Van De Wiele, Kristin Verbeke, Paul Rutgeerts, Benedikt Sas, Petra Louis, and Harry J. Flint. Butyric acid-producing anaerobic bacteria as a novel probiotic treatment approach for inflammatory bowel disease. *Journal of Medical Microbiology*, 59(2):141–143, February 2010.

[16] Silke Rath, Benjamin Heidrich, Dietmar H. Pieper, and Marius Vital. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. *Microbiome*, 5(1):54, December 2017.

[17] Mobina Kouhzad, Friedrich Götz, Tahereh Navidifar, Elahe Taki, Mahsa Ghamari, Roghayeh Mohammadzadeh, Maryam Seyedolmohadesin, and Narjess Bostanghadiri. Carcinogenic and anticancer activities of microbiota-derived secondary bile acids. *Front. Oncol.*, 15:1514872, January 2025.

[18] Jason M. Ridlon, Dae J. Kang, Phillip B. Hylemon, and Jasmohan S. Bajaj. Bile acids and the gut microbiome:. *Current Opinion in Gastroenterology*, 30(3):332–338, May 2014.

[19] Huawei Zeng, Shahid Umar, Bret Rust, Darina Lazarova, and Michael Bordonaro. Secondary Bile Acids and Short Chain Fatty Acids in the Colon: A Focus on Colonic Microbiome, Cell Proliferation, Inflammation, and Cancer. *IJMS*, 20(5):1214, March 2019.

[20] Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, October 2007. Publisher: Nature Publishing Group.

[21] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Songgang Shengting Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Songgang Shengting Shaochuan Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Shengting Shaochuan Li, Nan Qin, Huanming Yang, Jun Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S. Dusko Ehrlich, Jun Jian Wang, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem de Vos, Yohanan Winogradsky, Erwin Zoetendal, Peer Bork, S. Dusko Ehrlich, and Jun Jian Wang. A human gut microbial gene catalogue established by

metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010. Publisher: Nature Publishing Group.

[22] Shuo Han, Will Van Treuren, Curt R. Fischer, Bryan D. Merrill, Brian C. DeFelice, Juan M. Sanchez, Steven K. Higginbottom, Leah Guthrie, Lalla A. Fall, Dylan Dodd, Michael A. Fischbach, and Justin L. Sonnenburg. A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature*, 595(7867):415–420, July 2021. Publisher: Nature Publishing Group.

[23] Jarrod Roach, Rohit Mital, Jacob J. Haffner, Nathan Colwell, Randy Coats, Horvey M. Palacios, Zongyuan Liu, Joseane L.P. Godinho, Monica Ness, Thilini Peramuna, and Laura-Isobel McCall. Microbiome metabolite quantification methods enabling insights into human health and disease. *Methods*, 222:81–99, February 2024.

[24] Liam G Fearnley and Michael Inouye. Metabolomics in epidemiology: from metabolite concentrations to integrative reaction networks. *Int. J. Epidemiol.*, 45(5):1319–1328, October 2016.

[25] Uwe Sauer. Metabolic networks in motion:$^{13}$ C-based flux analysis. *Molecular Systems Biology*, 2(1):62, January 2006.

[26] Bruna De Falco, Francesco Giannino, Fabrizio Carteni, Stefano Mazzoleni, and Dong-Hyun Kim. Metabolic flux analysis: a comprehensive review on sample preparation, analytical techniques, data analysis, computational modelling, and main application areas. *RSC Adv.*, 12(39):25528–25548, 2022.

[27] Michael Zimmermann, Maria Zimmermann-Kogadeeva, Rebekka Wegmann, and Andrew L. Goodman. Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science*, 363(6427):eaat9931, February 2019. Publisher: American Association for the Advancement of Science.

[28] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, March 2010. Publisher: Nature Publishing Group.

[29] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121, January 2010. Number: 1 Publisher: Nature Publishing Group.

[30] Robert Schuetz, Lars Kuepfer, and Uwe Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology*, 3(1):119, January 2007. Publisher: John Wiley & Sons, Ltd.

[31] Edward J. O'Brien, Jonathan M. Monk, and Bernhard O. Palsson. Using Genome-scale Models to Predict Biological Capabilities. *Cell*, 161(5):971–987, May 2015. Publisher: Elsevier.

[32] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard Ø Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 6(9):1290–307, September 2011.

[33] Zachary A. King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson, and Nathan E. Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1):D515–D522, January 2016.

[34] Markus W. Covert and Bernhard Ø. Palsson. Transcriptional Regulation in Constraints-based Metabolic Models of Escherichia coli. *Journal of Biological Chemistry*, 277(31):28058–28064, August 2002.

[35] Nathan E. Lewis, Harish Nagarajan, and Bernhard O. Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol*, 10(4):291–305, April 2012.

[36] R. Mahadevan and C.H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, October 2003.

[37] Sergio Bordel, Rasmus Agren, and Jens Nielsen. Sampling the Solution Space in Genome-Scale Metabolic Networks Reveals Transcriptional Regulation in Key Enzymes. *PLoS Comput Biol*, 6(7):e1000859, July 2010.

[38] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(1):390, January 2010.

[39] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium forrepresentation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, March 2003.

[40] Ali Ebrahim, Joshua A. Lerman, Bernhard O. Palsson, and Daniel R. Hyduke. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, 7(1):74, August 2013.

[41] Axel von Kamp, Sven Thiele, Oliver Hädicke, and Steffen Klamt. Use of *CellNetAnalyzer* in biotechnology and metabolic engineering. *Journal of Biotechnology*, 261:221–228, November 2017.

[42] Maike K. Aurich, Ronan M. T. Fleming, and Ines Thiele. MetaboTools: A Comprehensive Toolbox for Analysis of Genome-Scale Metabolic Models. 7.

[43] Meiyappan Lakshmanan, Geoffrey Koh, Bevan K S Chung, and Dong-Yup Lee. Software applications for flux balance analysis. *Briefings in bioinformatics*, November 2012.

[44] Jennifer Levering, Tomas Fiedler, Antje Sieg, Koen W.A. Van Grinsven, Silvio Hering, Nadine Veith, Brett G. Olivier, Lara Klett, Jeroen Hugenholtz, Bas Teusink, Bernd Kreikemeyer, and Ursula Kummer. Genome-scale reconstruction of the Streptococcus pyogenes M49 metabolic network reveals growth requirements and indicates potential drug targets. *Journal of Biotechnology*, 232:25–37, August 2016.

[45] Kangsan Kim, Donghui Choe, Yoseb Song, Minjeong Kang, Seung-Goo Lee, Dae-Hee Lee, and Byung-Kwan Cho. Engineering Bacteroides thetaiotaomicron to produce non-native butyrate based on a genome-scale metabolic model-guided design. *Metabolic Engineering*, 68:174–186, November 2021.

[46] Enzymatic Determinations of Glucose. In *Advances in Clinical Chemistry*, volume 6, pages 67–96. Elsevier, 1963.

[47] Aarash Bordbar, James T. Yurkovich, Giuseppe Paglia, Ottar Rolfsson, Ólafur E. Sigurjónsson, and Bernhard O. Palsson. Elucidating dynamic metabolic physiology

through network integration of quantitative time-course metabolomics. *Sci Rep*, 7(1):46249, April 2017.

[48] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009.

[49] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, October 1995.

[50] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003.

[51] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*, 10(4):e1003580, April 2014.

[52] Ruth Großeholz, Ching-Chiek Koh, Nadine Veith, Tomas Fiedler, Madlen Strauss, Brett Olivier, Ben C. Collins, Olga T. Schubert, Frank Bergmann, Bernd Kreikemeyer, Ruedi Aebersold, and Ursula Kummer. Integrating highly quantitative proteomics and genome-scale metabolic modeling to study pH adaptation in the human pathogen Enterococcus faecalis. *npj Syst Biol Appl*, 2(1):1–9, September 2016. Number: 1 Publisher: Nature Publishing Group.

[53] Benjamín J Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology*, 13(8):935, August 2017.

[54] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Ø Bernhard. Reconstruction of Biochemical Networks in Microbial Organisms. *Nature reviews. Microbiology*, 7(2):129–143, 2009.

[55] Sebastián N. Mendoza, Brett G. Olivier, Douwe Molenaar, and Bas Teusink. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, 20(1):158, August 2019. Publisher: BioMed Central Ltd.

[56] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, page gkac963, October 2022.

[57] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research*, 48(D1):D445–D453, January 2020.

[58] Samuel M D Seaver, Filipe Liu, Qizhi Zhang, James Jeffryes, José P Faria, Janaka N Edirisinghe, Michael Mundy, Nicholas Chia, Elad Noor, Moritz E Beber, Aaron A Best, Matthew DeJongh, Jeffrey A Kimbrel, Patrik D'haeseleer, Sean R McCorkle, Jay R Bolton, Erik Pearson, Shane Canon, Elisha M Wood-Charlson, Robert W Cottingham, Adam P Arkin, and Christopher S Henry. 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*, 49(D1):D575–D588, January 2021.

[59] Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, September 2018. Publisher: Oxford University Press.

[60] João Capela, Davide Lagoa, Ruben Rodrigues, Emanuel Cunha, Fernando Cruz, Ana Barbosa, José Bastos, Diogo Lima, Eugénio C Ferreira, Miguel Rocha, and Oscar Dias. merlin , an improved framework for the reconstruction of high-quality genome-scale metabolic models. *Nucleic Acids Research*, 50(11):6052–6066, June 2022. Publisher: Oxford University Press (OUP).

[61] Brett Olivier. SystemsBioinformatics/cbmpy-metadraft: MetaDraft is now available, December 2018.

[62] Mark Hanemaaijer, Brett G. Olivier, Wilfred F. M. Röling, Frank J. Bruggeman, and Bas Teusink. Model-based quantification of metabolic interactions from dynamic microbial-community data. *PLoS ONE*, 12(3):e0173183, March 2017.

[63] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12(1):59–60, January 2015.

[64] Johannes Zimmermann, Christoph Kaleta, and Silvio Waschina. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology*, 22(1):81, March 2021.

[65] Christopher S Henry, Matthew DeJongh, Aaron a Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis

of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–82, September 2010. Publisher: Nature Publishing Group.

[66] Adam P Arkin, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, Michael W Sneddon, Matthew L Henderson, William J Riehl, Dan Murphy-Olson, Stephen Y Chan, Roy T Kamimura, Sunita Kumari, Meghan M Drake, Thomas S Brettin, Elizabeth M Glass, Dylan Chivian, Dan Gunter, David J Weston, Benjamin H Allen, Jason Baumohl, Aaron A Best, Ben Bowen, Steven E Brenner, Christopher C Bun, John-Marc Chandonia, Jer-Ming Chia, Ric Colasanti, Neal Conrad, James J Davis, Brian H Davison, Matthew DeJongh, Scott Devoid, Emily Dietrich, Inna Dubchak, Janaka N Edirisinghe, Gang Fang, José P Faria, Paul M Frybarger, Wolfgang Gerlach, Mark Gerstein, Annette Greiner, James Gurtowski, Holly L Haun, Fei He, Rashmi Jain, Marcin P Joachimiak, Kevin P Keegan, Shinnosuke Kondo, Vivek Kumar, Miriam L Land, Folker Meyer, Marissa Mills, Pavel S Novichkov, Taeyun Oh, Gary J Olsen, Robert Olson, Bruce Parrello, Shiran Pasternak, Erik Pearson, Sarah S Poon, Gavin A Price, Srividya Ramakrishnan, Priya Ranjan, Pamela C Ronald, Michael C Schatz, Samuel M D Seaver, Maulik Shukla, Roman A Sutormin, Mustafa H Syed, James Thomason, Nathan L Tintle, Daifeng Wang, Fangfang Xia, Hyunseung Yoo, Shinjae Yoo, and Dantong Yu. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol*, 36(7):566–569, July 2018.

[67] Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross A Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, 9(1):75, December 2008.

[68] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, November 2016. ISBN: 1546-1696 (Electronic)\r1087-0156 (Linking).

[69] Almut Heinken, Johannes Hertel, Geeta Acharya, Dmitry A. Ravcheev, Malgorzata Nyga, Onyedika Emmanuel Okpala, Marcus Hogan, Stefanía Magnúsdóttir, Filippo

Martinelli, Bram Nap, German Preciat, Janaka N. Edirisinghe, Christopher S. Henry, Ronan M. T. Fleming, and Ines Thiele. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat Biotechnol*, 41(9):1320–1331, September 2023. Publisher: Nature Publishing Group.

[70] Alberto Noronha, Jennifer Modamio, Yohan Jarosz, Elisabeth Guerard, Nicolas Sompairac, German Preciat, Anna Dröfn Daníelsdóttir, Max Krecke, Diane Merten, Hulda S Haraldsdóttir, Almut Heinken, Laurent Heirendt, Stefanía Magnúsdóttir, Dmitry A Ravcheev, Swagatika Sahoo, Piotr Gawron, Lucia Friscioni, Beatriz Garcia, Mabel Prendergast, Alberto Puente, Mariana Rodrigues, Akansha Roy, Mouss Rouquaya, Luca Wiltgen, Alise Žagare, Elisabeth John, Maren Krueger, Inna Kuperstein, Andrei Zinovyev, Reinhard Schneider, Ronan M.T. Fleming, and Ines Thiele. The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Research*, 47(D1):D614–D624, January 2018. Publisher: Oxford Academic.

[71] Changdai Gu, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. Current status and applications of genome-scale metabolic models. *Genome Biol*, 20(1):121, June 2019.

[72] David B. Bernstein, Snorre Sulheim, Eivind Almaas, and Daniel Segrè. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology*, 22(1):1–22, December 2021. Publisher: BioMed Central Ltd.

[73] Jacob Tamburro and Nanette R. Boyle. Charting the state of GEMs in microalgae: progress, challenges, and innovations. *Front. Plant Sci.*, 16:1614397, June 2025.

[74] Jean-Christophe Lachance, Colton J. Lloyd, Jonathan M. Monk, Laurence Yang, Anand V. Sastry, Yara Seif, Bernhard O. Palsson, Sébastien Rodrigue, Adam M. Feist, Zachary A. King, and Pierre-Étienne Jacques. BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Comput Biol*, 15(4):e1006971, April 2019.

[75] A. Wegkamp, B. Teusink, W.M. De Vos, and E.J. Smid. Development of a minimal growth medium for Lactobacillus plantarum. *Letters in Applied Microbiology*, 50(1):57–64, January 2010.

[76] David R. Hillyard and Michael J. Redd. Identification of Essential Genes in Bacteria. In *Methods in Enzymology*, volume 421, pages 21–34. Elsevier, 2007.

[77] David B Bernstein, Batu Akkas, Morgan N Price, and Adam P Arkin. Evaluating E. coli genome-scale metabolic model accuracy with high-throughput mutant fitness data. *Molecular Systems Biology*, 19(12):e11566, December 2023. Publisher: John Wiley & Sons, Ltd.

[78] Yunli Eric Hsieh, Kshitij Tandon, Heroen Verbruggen, and Zoran Nikoloski. Comparative analysis of metabolic models of microbial communities reconstructed from automated tools and consensus approaches. *npj Syst Biol Appl*, 10(1):1–10, May 2024. Publisher: Nature Publishing Group.

[79] Philipp Wendering and Zoran Nikoloski. COMMIT: Consideration of metabolite leakage and community composition improves microbial community reconstructions. *PLoS Comput Biol*, 18(3):e1009906, March 2022.

[80] Archana Hari, Arveen Zarrabi, and Daniel Lobo. *mergem* : merging, comparing, and translating genome-scale metabolic models using universal identifiers. *NAR Genomics and Bioinformatics*, 6(1):lqae010, January 2024.

[81] Elena K. Matveishina, Bartosz J. Bartmanski, Sara Benito-Vaquerizo, and Maria Zimmermann-Kogadeeva. GEMsembler: consensus model assembly and structural comparison of genome-scale metabolic models across tools improve functional performance. *mSystems*, pages e00574–25, September 2025.

[82] Julia Isenring, Lea Bircher, Annelies Geirnaert, and Christophe Lacroix. In vitro human gut microbiota fermentation models: Opportunities, challenges, and pitfalls. 2(1):2.

[83] Benjamin D. Anderson and Jordan E. Bisanz. Challenges and opportunities of strain diversity in gut microbiome research. 14:1117122.

[84] Daniel Rios Garza, Didier Gonze, Haris Zafeiropoulos, Bin Liu, and Karoline Faust. Metabolic models of human gut microbiota: Advances and challenges. 14(2):109–121.

[85] Elif Esvap and Kutlu O. Ulgen. Advances in Genome-Scale Metabolic Modeling toward Microbial Community Analysis of the Human Microbiome. 10(9):2121–2137.

[86] Jun Sun, Shelley A Haveman, Olivia Bui, Tom R Fahland, and Derek R Lovley. Constraint-based modeling analysis of the metabolism of two Pelobacter species. 4(1):174.

[87] Eugen Bauer, Cedric Christian Laczny, Stefania Magnusdottir, Paul Wilmes, and Ines Thiele. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. 3(1):55.

[88] Charlotte Ramon and Jörg Stelling. Functional comparison of metabolic networks across species. 14(1):1699.

[89] Kai Blin. ncbi-genome-download, July 2023.

[90] Mathias Ganter, Thomas Bernard, Sébastien Moretti, Joerg Stelling, and Marco Pagni. MetaNetX.org: a website and repository for accessing, analyzing, and manipulating metabolic networks. *Bioinformatics (Oxford, England)*, pages 2–3, January 2013.

[91] Sébastien Moretti, Van Du T Tran, Florence Mehl, Mark Ibberson, and Marco Pagni. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Research*, 49(D1):D570–D574, January 2021.

[92] John T. Sauls and Joerg M. Buescher. Assimilating genome-scale metabolic reconstructions with modelBorgifier. *Bioinformatics*, 30(7):1036–1038, April 2014.

[93] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, December 2009.

[94] Aarthi Ravikrishnan, Meghana Nasre, and Karthik Raman. Enumerating all possible biosynthetic pathways in metabolic networks. *Sci Rep*, 8(1):9932, July 2018. Publisher: Nature Publishing Group.

[95] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. SymPy: symbolic computing in Python. *PeerJ Computer Science*, 3:e103, January 2017.

[96] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Rev.*, 45(2):167–256, January 2003.

[97] Giancarlo Perrone, Jose Unpingco, and Haw-minn Lu. Network visualizations with Pyvis and VisJS, 2020. Version Number: 1.

[98] Almut Heinken, Swagatika Sahoo, Ronan M T Fleming, and Ines Thiele. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*, 4(1):28–40, January 2013. ISBN: 1949-0984 (Electronic) 1949-0976 (Linking).

[99] Bas Teusink, Frank H. J. Van Enckevort, Christof Francke, Anne Wiersma, Arno Wegkamp, Eddy J. Smid, and Roland J. Siezen. In Silico Reconstruction of the Metabolic Pathways of *Lactobacillus plantarum* : Comparing Predictions of Nutrient Requirements with Those from Growth Experiments. *Appl Environ Microbiol*, 71(11):7253–7262, November 2005.

[100] Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, Adam M Feist, and Bernhard O Palsson. iML1515, a knowledgebase that computes Escherichia coli traits. *Nat Biotechnol*, 35(10):904–908, October 2017.

[101] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[102] Kelly M. Wetmore, Morgan N. Price, Robert J. Waters, Jacob S. Lamson, Jennifer He, Cindi A. Hoover, Matthew J. Blow, James Bristow, Gareth Butland, Adam P. Arkin, and Adam Deutschbauer. Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. *mBio*, 6(3):e00306–15, July 2015.

[103] Morgan N. Price, Kelly M. Wetmore, R. Jordan Waters, Mark Callaghan, Jayashree Ray, Hualan Liu, Jennifer V. Kuehl, Ryan A. Melnyk, Jacob S. Lamson, Yumi Suh, Hans K. Carlson, Zuelma Esquivel, Harini Sadeeshkumar, Romy Chakraborty, Grant M. Zane, Benjamin E. Rubin, Judy D. Wall, Axel Visel, James Bristow, Matthew J. Blow, Adam P. Arkin, and Adam M. Deutschbauer. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509, May 2018. Publisher: Nature Publishing Group.

[104] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30, January 2000.

[105] N Sakamoto, A M Kotre, and M A Savageau. Glutamate dehydrogenase from Escherichia coli: purification and properties. *Journal of Bacteriology*, 124(2):775–783, November 1975. Publisher: American Society for Microbiology.

[106] Francesco M. Veronese, Enrico Boccu, and Luciano Conventi. Glutamate dehydrogenase from *Escherichia coli*: Induction, purification and properties of the enzyme. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 377(2):217–228, February 1975.

[107] Sok Ho Kim, Barbara L. Schneider, and Larry Reitzer. Genetics and Regulation of the Major Enzymes of Alanine Synthesis in Escherichia coli. *Journal of Bacteriology*, 192(20):5304–5311, October 2010. Publisher: American Society for Microbiology.

[108] Ronald C. Wek, Craig A. Hauser, and G. Wesley Hatfield. The nucleotide sequence of the ilvBN operon of Escherichia coli: sequence homologies of the acetohydroxy acid synthase isozymes. *Nucleic Acids Research*, 13(11):3995–4010, June 1985.

[109] P. Friden, J. Donegan, J. Mullen, P. Tsui, M. Freundlich, L. Eoyang, R. Weber, and P.M. Silverman. The ilvB locus of Escherichia coli K-12 is an operon encoding both subunits of acetohydroxyacid synthase I. *Nucleic Acids Research*, 13(11):3979–3992, June 1985.

[110] Kevin L Howe, Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, Matthieu Barba, Dan M Bolser, Lahcen Cambell, Manuel Carbajo, Marc Chakiachvili, Mikkel Christensen, Carla Cummins, Alayne Cuzick, Paul Davis, Silvie Fexova, Astrid Gall, Nancy George, Laurent Gil, Parul Gupta, Kim E Hammond-Kosack, Erin Haskell, Sarah E Hunt, Pankaj Jaiswal, Sophie H Janacek, Paul J Kersey, Nick Langridge, Uma Maheswari, Thomas Maurel, Mark D McDowall, Ben Moore, Matthieu Muffato, Guy Naamati, Sushma Naithani, Andrew Olson, Irene Papatheodorou, Mateus Patricio, Michael Paulini, Helder Pedro, Emily Perry, Justin Preece, Marc Rosello, Matthew Russell, Vasily Sitnik, Daniel M Staines, Joshua Stein, Marcela K Tello-Ruiz, Stephen J Trevanion, Martin Urban, Sharon Wei, Doreen Ware, Gary Williams, Andrew D Yates, and Paul Flicek. Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic Acids Research*, 48(D1):D689–D695, January 2020.

[111] Xuejiao Song and John E. Cronan. A conserved and seemingly redundant Escherichia coli biotin biosynthesis gene expressed only during anaerobic growth. *Molecular Microbiology*, 116(5):1315–1327, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mmi.14826.

[112] Aaron G. Wexler and Andrew L. Goodman. An insider's perspective: Bacteroides as a window into the microbiome. *Nat Microbiol*, 2(5):17026, April 2017.

[113] Lisa Maier, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, Keith Conrad Fernandez, Hitomi Dose, Hirotada Mori, Kiran Raosaheb Patil, Peer Bork, and Athanasios Typas. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, March 2018. Publisher: Nature Publishing Group.

[114] Patricia G. Wolf, Saravanan Devendran, Heidi L. Doden, Lindsey K. Ly, Tyler Moore, Hajime Takei, Hiroshi Nittono, Tsuyoshi Murai, Takao Kurosawa, George E. Chlipala, Stefan J. Green, Genta Kakiyama, Purna Kashyap, Vance J. McCracken, H. Rex Gaskins, Patrick M. Gillevet, and Jason M. Ridlon. Berberine alters gut microbial function through modulation of bile acids. *BMC Microbiol*, 21(1):24, December 2021.

[115] Yvonne Konkol, Anniina Keskitalo, Heikki Vuorikoski, Sami Pietilä, Laura L. Elo, Eveliina Munukka, Jenni Bernoulli, and Johanna Tuomela. Chronic nonbacterial prostate inflammation in a rat model is associated with changes of gut microbiota that can be modified with a galactoglucomannan-rich hemicellulose extract in the diet. *BJU International*, 123(5):899–908, May 2019.

[116] Hiroaki Yamada, Yuto Watabe, Yutaka Suzuki, Satoshi Koike, Shu Shimamoto, and Yasuo Kobayashi. Chemical and microbial characterization for fermentation of water-soluble cellulose acetate in human stool cultures. *J Sci Food Agric*, 101(7):2950–2960, May 2021.

[117] Jun Feng, Yili Qian, Zhichao Zhou, Sarah Ertmer, Eugenio I. Vivas, Freeman Lan, Joshua J. Hamilton, Federico E. Rey, Karthik Anantharaman, and Ophelia S. Venturelli. Polysaccharide utilization loci in Bacteroides determine population fitness and community-level interactions. *Cell Host & Microbe*, 30(2):200–215.e12, February 2022.

[118] Erica D. Sonnenburg, Hongjun Zheng, Payal Joglekar, Steven K. Higginbottom, Susan J. Firbank, David N. Bolam, and Justin L. Sonnenburg. Specificity of Polysaccharide Use in Intestinal Bacteroides Species Determines Diet-Induced Microbiota Alterations. *Cell*, 141(7):1241–1252, June 2010.

[119] Alfonso Benítez-Páez, Eva M. Gómez del Pulgar, and Yolanda Sanz. The Glycolytic Versatility of Bacteroides uniformis CECT 7771 and Its Genome Response to Oligo and Polysaccharides. *Frontiers in Cellular and Infection Microbiology*, 7, 2017.

[120] Nicole Reichardt, Sylvia H Duncan, Pauline Young, Alvaro Belenguer, Carol McWilliam Leitch, Karen P Scott, Harry J Flint, and Petra Louis. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *The ISME Journal*, 8(6):1323–1335, June 2014.

[121] Shanshan Qiao, Li Bao, Kai Wang, Shanshan Sun, Mingfang Liao, Chang Liu, Nan Zhou, Ke Ma, Yuwei Zhang, Yihua Chen, Shuang-Jiang Liu, and Hongwei Liu. Activation of a Specific Gut Bacteroides-Folate-Liver Axis Benefits for the Alleviation of Nonalcoholic Hepatic Steatosis. *Cell Reports*, 32(6):108005, August 2020.

[122] Jason M. Ridlon, Saravanan Devendran, João Mp Alves, Heidi Doden, Patricia G. Wolf, Gabriel V. Pereira, Lindsey Ly, Alyssa Volland, Hajime Takei, Hiroshi Nittono, Tsuyoshi Murai, Takao Kurosawa, George E. Chlipala, Stefan J. Green, Alvaro G. Hernandez, Christopher J. Fields, Christy L. Wright, Genta Kakiyama, Isaac Cann, Purna Kashyap, Vance McCracken, and H. Rex Gaskins. The 'in vivo lifestyle' of bile acid 7α-dehydroxylating bacteria: comparative genomics, metatranscriptomic, and bile acid metabolomics analysis of a defined microbial community in gnotobiotic mice. *Gut Microbes*, 11(3):381–404.

[123] M.A. Zocco, M.E. Ainora, G. Gasbarrini, and A. Gasbarrini. Bacteroides thetaiotaomicron in the gut: Molecular aspects of their interaction. *Digestive and Liver Disease*, 39(8):707–712, August 2007.

[124] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, January 2015.

[125] Matthew A Oberhardt, Bernhard Ø Palsson, and Jason A Papin. Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, 5(1):320, January 2009.

[126] Ove Øyås and Jörg Stelling. Genome-scale metabolic networks in time and space. *Current Opinion in Systems Biology*, 8:51–58, April 2018.

[127] Nuala A. O'Leary, Eric Cox, J. Bradley Holmes, W. Ray Anderson, Robert Falk, Vichet Hem, Mirian T. N. Tsuchiya, Gregory D. Schuler, Xuan Zhang, John Torcivia, Anne Ketter, Laurie Breen, Jonathan Cothran, Hena Bajwa, Jovany Tinne, Peter A. Meric, Wratko Hlavina, and Valerie A. Schneider. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data*, 11(1):732, July 2024.

[128] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, 37(5):540–546, May 2019.

[129] Tatiana A Gurbich, Martin Beracochea, Nishadi H De Silva, and Robert D Finn. mettannotator: a comprehensive and scalable Nextflow annotation pipeline for prokaryotic assemblies. *Bioinformatics*, 41(2):btaf037, February 2025.

[130] Vincent H. Varel and Marvin P. Bryant. Nutritional Features of *Bacteroides fragilis* subsp. *fragilis*. *Appl Microbiol*, 28(2):251–257, August 1974.

[131] Eric C. Martens, Herbert C. Chiang, and Jeffrey I. Gordon. Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host & Microbe*, 4(5):447–457, November 2008.

[132] Joana C. Xavier, Kiran Raosaheb Patil, and Isabel Rocha. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. *Metabolic Engineering*, 39:200–208, January 2017.

[133] Vikram V. Shende, Katherine D. Bauman, and Bradley S. Moore. The shikimate pathway: gateway to metabolic diversity. *Nat. Prod. Rep.*, 41(4):604–648, 2024.

[134] E. E. Putnam, J. Abellon-Ruiz, B. J. Killinger, J. J. Rosnow, A. G. Wexler, E. Folta-Stogniew, A. T. Wright, B. Van Den Berg, and A. L. Goodman. Gut Commensal *Bacteroidetes* Encode a Novel Class of Vitamin $B_{12}$ -Binding Proteins. *mBio*, 13(2):e02845–21, April 2022.

[135] Andrew L. Goodman, Nathan P. McNulty, Yue Zhao, Douglas Leip, Robi D. Mitra, Catherine A. Lozupone, Rob Knight, and Jeffrey I. Gordon. Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host & Microbe*, 6(3):279–289, September 2009.

[136] Nikita Denisov, Fabian Springer, Amber Brauer-Nikonow, George Maftei, Georg Zeller, Denise Medeiros Selegato, and Michael Zimmermann. Development of a GC-MS/MS method to quantify 120 gut microbiota-derived metabolites.

[137] David P. Giedroc. A new player in bacterial sulfide-inducible transcriptional regulation. *Molecular Microbiology*, 105(3):347–352, August 2017.

[138] Daniel A. Charlebois and Gábor Balázsi. Modeling cell population dynamics. *In Silico Biology: Journal of Biological Systems Modeling and Multi-Scale Simulation*, 13(1-2):21–39, May 2019.

[139] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Publisher: Nature Research.

[140] Shu Pan, Kiel Nikolakakis, Paul A. Adamczyk, Min Pan, Edward G. Ruby, and Jennifer L. Reed. Model-enabled gene search (MEGS) allows fast and direct discovery of enzymatic and transport gene functions in the marine bacterium Vibrio fischeri. *Journal of Biological Chemistry*, 292(24):10250–10261, June 2017. Publisher: Elsevier.

[141] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, January 2010. Publisher: Nature Publishing Group.

[142] Hoang V. Dinh, Debolina Sarkar, and Costas D. Maranas. Quantifying the propagation of parametric uncertainty on flux balance analysis. *Metabolic Engineering*, 69:26–39, January 2022.

[143] Axel Von Kamp and Steffen Klamt. Balancing biomass reaction stoichiometry and measured fluxes in flux balance analysis. *Bioinformatics*, 39(10):btad600, October 2023.