Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Chemistry, Biotechnology and Food Sciences

Unbiased Analysis of Metabolite Exchanges in Metabolic Models

Ylva Katarina Wedmark
MSc in Chemistry and Biotechnology

Acknowledgements

The final semester of my master's degree has truly been the most interesting and experience rich semester of my academic endeavor. The wisdom I have gained, and all my efforts have cumulated in this thesis. I want to thank my supervisor Ove Øyås for guiding me and for believing in me. Your advice always left me excited about my work and motivated to improve it. Furthermore, I want to thank Teshome Dagne Mulugeta at Orion support for providing excellent help with setting up the required software on the Orion HPC environment at Norwegian University of Life Sciences. I also want to thank the staff at Systems Biology Lab, especially Bas Teusink, Frank Bruggeman, Bob Planque and Pranas Grigaitis for invaluable insight and interesting discussions about metabolic modeling. I would not have come far without the help from developers of *ecmtool* Tom Clement and Daan de Groot, who troubleshot and subsequently fixed a program bug. Thank you so much for helping me with this. Finally, I thank all family and friends for all their support and love.

Ylva Katarina Wedmark, May 2022

Summary

A metabolic network is the system of biochemical reactions that sustain the life of an organism. The genetics of a given organism is highly involved in the reaction fluxes in the network. In recent years, large amounts of genomic data have been obtained from a variety of organisms. This data can be integrated in mathematical models of a network, enabling both prediction and understanding of diverse metabolic networks under a wide variety of circumstances. The metabolic models can be mathematically analyzed more easily by imposing constraints on the reaction rates of the network, based on empirical data. Two types of constraint-based analysis have emerged: Biased analyses, based on maximizing or minimizing some objective of the metabolic model, for example maximizing cellular growth, and unbiased analyses, attempting to describe all possible flux combinations though a modeled metabolic network that are valid under the constraints of the model. While biased methods have evolved to analyze comprehensive genome-scale and multi-cellular models, the unbiased methods lag behind because the number of computations grow combinatorically with model size. Multiple approaches aim to improve scalability of unbiased analysis. In this thesis, I apply two unbiased methods: Enumeration of elementary conversion modes (ECMs) and enumeration of minimal pathways (MPs) on models of different sizes. I aim to compare the methods, reproduce results from earlier work, and gain insight on how to make unbiased analysis more scalable. The results highlight the importance of reproducing previous findings and the scalability of some unbiased methods. Initially the results were not successfully reproduced, but after the cause was discovered, the results were reproduced correctly. MP analysis scaled to larger models than ECMs, by allowing more constraints on the model and by implementing random sampling of MPs. However, this approach can be challenging, as the number of samples required is not known beforehand. Furthermore, the additional constraints could be a source of bias. Although challenges of unbiased analysis still persist, this research clarifies which strategies that could be implemented for improving scalability of unbiased analysis.

Sammendrag

Et metabolsk nettverk er det biokjemiske systemet som opprettholder en organisme. Genetikken til en gitt organisme er i høyest grad involvert i hvilke reaksjoner som forekommer i nettverket. I de siste årene har store mengder genetisk data blitt hentet fra et mangfoldig utvalg av organismer. Dataen kan broderes inn i matematiske modeller av metabolske nettverk, noe som tillater prediksjon og forståelse av varierte metabolske nettverk, utsatt for varierte betingelser. Disse modellene kan analyseres mye lettere ved å innføre begrensinger av hastigheten til reaksjonene i modellen, basert på empirisk data. Begrensings-basert modellering kan inndeles i to typer analyse: Partisk og upartisk analyse. Partisk analyse er basert på optimering av et egen-definert formål, for eksempel maksimering av cellevekst. Ved upartisk analyse er målet å beskrive alle mulige kombinasjoner av reaksjonshastigheter, som er mulige ved gitte begrensninger, gjennom det modellerte metabolske nettverket. Mens partiske metoder brukes på store metabolske modeller og flercellede modeller, har ikke de upartiske metodene utviklet seg til dette nivået ennå. En viktig grunn til dette er at antall kalkulasjoner som trengs for å finne alle hastighetskombinasjonene øker kombinatorisk med modellstørrelse. Flere strategier har blitt benyttet for å gjøre upartisk analyse mer skalerbart. I denne oppgaven anvender jeg to upartiske analysemetoder på metabolske modeller av ulik størrelse: Opptelling av elementary conversion modes (ECMs) og opptelling av minimal pathways (MPs). Formålet er å sammenligne metodene, reprodusere tidligere funn og forstå hvordan man kan gjøre upartisk analyse mer skalerbart. Resultatene viser viktigheten i å reprodusere tidligere funn og skalerbarheten til noen upartiske metoder. I første omgang ble ikke tidligere funn reprodusert, men etter en feil i programmet ble oppdaget og fikset ble de reprodusert. MP analyse skalerte bedre til større modeller enn ECMs ved å tillate flere begrensinger på modellen og ved å implementere tilfeldig utvalg av MPs. Likevel kan det være utfordrende å bruke denne tilnærmingen fordi man ikke kan vite utvalgsstørrelsen på forhånd. Dette arbeidet oppklarer hvilke strategier som kan implementeres for å forbedre skalerbarheten til upartiske analyser.

List of figures

List of tables

List of Acronyms

CBMM	Constraint-based metabolic modeling
COBRA	Constraint-based reconstruction and analysis
GEM	Genome-scale metabolic model
QSSA	Quasi-steady-state assumption
EFM	Elementary flux mode
FC	Flux cone
EFV	Elementary flux vector
FP	Flux polyhedron
ECM	Elementary conversion mode
FVA	Flux variability analysis
MP	Minimal pathway

Table of contents

1 Introduction

1.1 Metabolic pathways and networks

Metabolic pathways are a series of enzyme-catalyzed reactions that occur in a cell or organism. The molecules partaking in the reactions are called metabolites (Nelson & Cox, 2017). Pathways often fulfill a biological objective, for example glycolysis is a pathway that breaks down sugar molecules and transforms it to energy that can be used by the cell. All metabolic pathways of an organism, along with the interplays between pathways form the organism's metabolic network. The network comprises hundreds to thousands of reactions which converts sources of energy from the environment to consumable energy, and synthesizes molecules that are needed for growth (Wagner, 2012). The cell interacts with its environment by transferring nutrients into the cell or secreting waste products out of the cell. Through billions of years of evolution, the metabolic network of a cell has adapted to a wide range of environmental changes, for example conserving energy when few nutrients are available. Thus, metabolic networks need robust regulatory mechanisms. Genes transcribe the proteins that catalyze the reactions of metabolic pathways. The genes determine the metabolic phenotypes of the cell and are highly regulated to account for adaptations in environment. Cell metabolism is complex, and the understanding of it is central to questions about life on earth.

1.2 Constraint-based metabolic modeling

The study of metabolism is often based on the order of enzyme discovery or pathway usage in one model organism. Although much knowledge can be gained from studying pathways in isolation in model organisms, it may not reflect the variation in network flow between organisms (Lewis et al., 2012). Due to advancements in genome sequencing, a huge amount of data on a variety of organisms is now readily available and allow for diverse studies of metabolism. The sequencing data has been incorporated into mathematical models, which have enabled meaningful ways of analyzing metabolic networks. Constraint-based metabolic modelling (CBMM) is one of the fields that have emerged for this type of metabolic network analysis. In this framework, a genome scale metabolic model (GEM) represents the metabolism of a cell. GEMs describe all known metabolic reactions of an organism and are further defined by a handful of genetic processes related to the organism. The link between genetic information and the metabolic network is established by gene-protein-reaction (GPR) associations. They are logical rules that determine which proteins are synthesized, which in turn define a set of metabolic reactions that can be catalyzed (de Mas et al., 2019; Gu et al., 2019). Therefore, the GEMs only integrate genes that are directly linked to metabolism. A cell's metabolism is also governed by environmental factors and physical and chemical laws (Lewis et al., 2012). By imposing constraints on the reaction rates in the network reflecting these factors, the CBMs can accurately predict metabolic capabilities of organisms (Gottstein et al., 2016).

Exactly how are metabolic networks analyzed and mathematically formulated? As previously mentioned, a metabolic pathway or network can be defined as a collection of reactions. Each reaction can be represented as an equation where each metabolite is mass balanced. All the reaction equations of a metabolic pathway or network can be reframed into a matrix, known as the stoichiometric matrix. Each column of the matrix describes a reaction, while each row contains the network stoichiometry of one metabolite. The matrix is multiplied by a vector that contains the fluxes (rates) of the reactions. The stoichiometric matrix and the fluxes describe the flow of the network or change in concentration over time:

$$\frac{dC(t)}{dt} = Nr(t) \tag{1}$$

Where $\frac{dC(t)}{dt}$ is the change in concentration with respect to time, usually measured in mmol $gDW^{-1}h^{-1}$ (Aurich et al., 2016), *N* is the stoichiometric matrix, and *r* is the flux vector.

Additional pseudo reactions can be added to the set of equations. These are reactions that are not based on biochemistry but fulfill a purpose for modeling. For example, exchange reactions import or export metabolites into or out of a network (Thiele & Palsson, 2010). Another important type of pseudo reactions is the biomass reaction. This reaction contains all basic metabolite building blocks for all cellular components, thereby serving as an indicator of cell growth (Lewis et al., 2012).

To be able to predict the fluxes of all reactions in the model, constraints are imposed on them. The most important constraint is the quasi-steady-state assumption (QSSA). It is assumed that all metabolites partaking in reactions internally in the cell must be cancelled out when all reactions are considered, i.e that no metabolites are allowed to accumulate or disappear inside the cell. The QSSA is justified by the fact that metabolism happens very fast compared to other biological processes. Thus, it is assumed that the metabolites already have converged to a steady state at every measured time point (Reimers & Reimers, 2016). Although analysis of dynamic models like equation (1) would yield more accurate predictions, they require extensive experimental data that is generally not available. The lacking experimental data can be approximated by parameter estimation, but at the expense of computational cost of the overall process (Kim et al., 2018). By making the QSSA, the model can be written as:

$$Nr = 0 \tag{2}$$

The reaction rates can be further constrained by setting upper and lower bounds on fluxes. This limits the feasible fluxes of reactions, which can make it easier to analyze the network:

$$r_i^{lb} \leq r \leq r_i^{ub} \tag{3}$$

Where r_i^{lb} and r_i^{ub} are the lower and upper flux bounds of reaction *i*, respectively.

Some methods of metabolic network analysis require constraints with only zeroes on the right-hand side, called homogeneous constraints, to simplify calculations. Thus, many models assume that many reactions of the network are irreversible. This condition is formulated as homogeneous constraints and can be incorporated as a constraint to (2).

$$r_i \geq 0 \; for \; i \in Irr \tag{4}$$

Where *Irr* denotes the set of all reactions that are classified as irreversible (all other reactions are reversible).

The system of equations and inequalities described by equation (2)-(4) forms the basis of CBMM. Combinations of fluxes that satisfy the imposed constraints are solutions to the system, known as flux distributions. They represent combinations of possible metabolic pathways through the network. Together, all valid flux distributions form the solution space or feasible flux space.

1.3 Unbiased pathway analysis

A wide set of analyses in the CBMM field are formulated as optimization problems that find the optimal flux distributions for an objective function (Lewis et al., 2012). In many cases, we can assume that the "objective" of a cell is to grow as fast as possible. Thus, the fluxes in the model could be predicted by maximizing the biomass function. However, the assumption of optimality introduces bias, and the real objective of a cell is unknown. For example, when cells are exposed to short-term environmental changes, the cells do not necessarily optimize growth (Herrmann et al., 2019). A family of unbiased analyses circumvent this bias by attempting to characterize the entire solution space of the system of equations and inequalities shown above. In other words, they attempt to define and comprehensively describe all metabolic routes that are feasible under given constraints (Schuster et al., 2000).

The methods of unbiased analysis can be categorized in two groups: 1) sampling methods and 2) pathway analysis. For flux sampling, a certain number of random feasible solutions are obtained (ideally until most of the solution space is represented). The resulting samples provide a probability distribution over all fluxes. However, analysis of large networks by random sampling is challenging. (Herrmann et al., 2019). Pathway analysis identifies combinations of reactions, or pathways, that reflect all possible flux routes through the metabolic network in steady-state (Lewis et al., 2012). The challenge with this approach is that the number of possible pathways increases combinatorically with network size (Klamt & Stelling, 2002). Therefore, these analyses quickly become incomputable. Models only become larger and more comprehensive with time, and other CBMM approaches have moved to analysis of microbial consortia (Theorell & Stelling, 2022). Pathway analysis could provide further insight into larger networks as well, but the methods are not yet scalable to this degree. In this thesis I will focus on pathway analysis, and handful of key approaches will be introduced in the following sections.

1.3.1 Elementary flux modes and elementary flux vectors

The steady-state assumption, equation (2), in addition to the constraint that all reactions must be non-negative, similarly as in equation (4), defines the solution space of elementary flux modes (EFMs). The solutions to the set of equations and inequalities lie in the intersection of the nullspace, and the non-negative half-spaces associated with the irreversible reactions. The null space (light and dark grey area in Figure 1.1) consists of all vectors *r* (flux distributions) that satisfy equation (2). The nullspace is divided into half-spaces by planes, which in the case of the irreversibility restrictions, cut each orthant in half at 0. This solution space is called the flux cone (*FC*) of the network (dark grey area in Figure 1.1). The flux cone is mathematically expressed as:

$$FC = \{r \in \mathbb{R}^n \mid Nr = 0, r_i \geq 0 \; for \; i \in Irr\} \tag{5}$$

EFMs are defined as the support-minimal vectors of this flux cone (Klamt et al., 2017). This means that there exists no flux vector that is a proper subset (in terms of reactions with nonzero indices) of another flux vector in that set. EFMs are also linearly independent, meaning that no EFM can be expressed as a linear combination of the other EFMs in the set. Every flux distribution in the flux cone can be described as a non-negative linear combination of EFMs without cancelling out a reaction (Clement et al., 2021). In addition, the fluxes of every subnetwork of the original network i.e., a network where some of the reactions in the original network are excluded, can be expressed as parts of EFMs that contain the reactions included in the sub-network (Klamt et al., 2017). By summarizing the flux cone in a set of vectors, EFMs provide a useful way of characterizing all possible fluxes in a metabolic network.

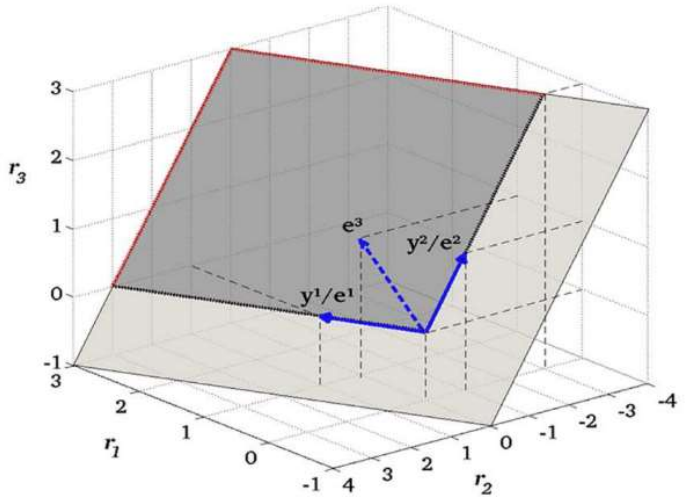

Figure 1.1 The flux cone of an example network. Irreversible constraints are imposed on flux r_1 and r_2. The flux cone (colored in dark grey) is outlined by removing the halfspaces associated with negative fluxes of r_1 and r_2 from the nullspace (light grey). The red dotted lines indicate unbounded edges of the cone. The y^1 and y^2 vectors represents the edges of the cone, and along with the additional e^3 vector, they make up the set of EFMs (e^1, e^2, e^3) that describe this flux cone. The figure is taken from Klamt et al. (2017).

EFMs are limited to homogeneous constraints. As discussed earlier, some network analyses benefit from imposing inhomogeneous constraints, as they can reduce the solution space. For multiple flux constraints, the constraints can be expressed in matrix form:

$$Gr \geq h \tag{6}$$

where each flux constraint occupies a row, and each reaction occupies a column in matrix *G*, and *h* represents the boundaries of the constraints. These boundaries create additional planes that intersect the nullspace and further restricts it (green and yellow areas in Figure 1.2). When incorporating these constraints, the resulting flux polyhedron (*FP*) becomes a subset of the flux cone:

$$FP = \{r \in \mathbb{R}^n \mid Nr = 0, r_i \geq 0 \; for \; i \in Irr, \; Gr \leq h\} \tag{7}$$

To describe the flux polyhedron, a generalization of EFMs has been defined: E lementary flux vectors (EFVs). Most of the properties applied to EFMs are conserved for EFVs, in that every element of the polyhedron can be expressed as a linear combination of EFVs without cancellation, and that subnetworks can be expressed by subsets of the EFVs. They are, however, not support-minimal (Klamt et al., 2017).

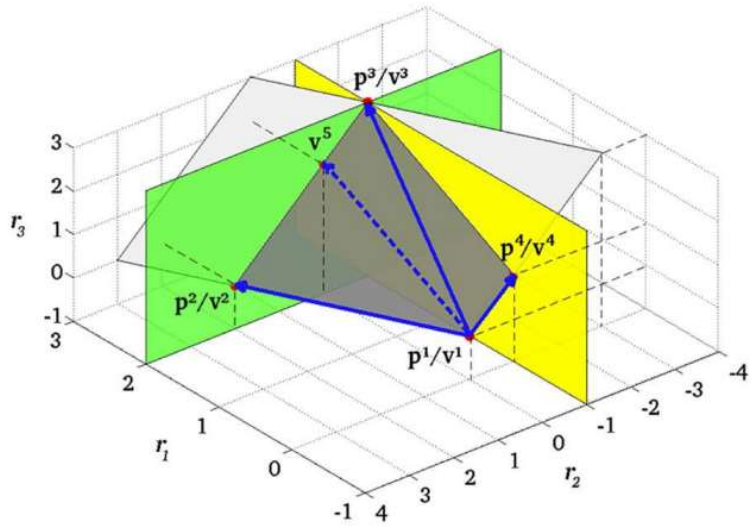

Figure 1.2 The flux polyhedron of the example network provided in Figure 1.1. Additional inhomogeneous flux boundaries, $0 \leq r_1 \leq 2$ and $r_2 \geq -1$, give rise to two planes (green and yellow respectively) that cut into the flux cone (light grey). The remaining space (dark grey) forms the flux polyhedron. The resulting polyhedron has four vectors on the edges of the polyhedron (p^1, p^2, p^3 and p^4), and five elementary flux modes (v^1, v^2, v^3, v^4, v^5). The figure is taken from Klamt et al. (2017).

1.3.2 Elementary conversion modes

Elementary conversion modes (ECMs) are defined similarly as EFMs; the main difference is the solution space. EFMs describe the flux cone defined by the reaction rates of the network, while ECMs describe the conversion cone defined by the metabolites in the network. Like the flux cone, the conversion cone is also constrained by the steady-state assumption in equation (2), which is imposed on all "internal metabolites". The internal metabolites are defined as all metabolites only partaking in reactions inside the cell. In addition, all reactions are assumed to be irreversible i.e., reverse reactions are split into separate forward and reverse reactions. This makes the calculations easier, as the solution space exclusively exists in one orthant (Clement et al., 2021). These constraints together form the conversion cone:

$$C = \{c = Nr \mid c_i = 0 \; for \; i \in Int, r_j \geq 0 \; \forall \; j\} \tag{8}$$

where *C* denotes the conversion cone, *c* denotes the change in concentration of a given metabolite (conversion), and *Int* denotes all internal metabolites. ECMs are the minimal set of vectors of the conversion cone such that each steady-state conversion can be written as a positive sum of ECMs, without the production of any external metabolite being cancelled in that sum (Clement et al., 2021). The ECMs are a set of vectors, where each vector contains stoichiometric coefficients of net conversions of the network's external metabolites. By scaling and adding the ECMs, they can describe every possible conversion of the network.

There are many flux combinations that lead to the same overall conversion. This means that multiple EFMs can be directly linked to one ECM. By shifting the focus to unique metabolite conversions instead of reaction fluxes, the number of vectors significantly decreases, making this approach more scalable than EFM enumeration. Enumeration of a network's ECMs is implemented in *ecmtool* developed by Clement et al. (2021), which has a handful of features that improves scalability. By only describing inputs and outputs of a network, ECM analysis gains an important advantage which significantly speeds up enumeration of ECMs compared to EFMs. As reaction rates are not reported

by ECMs, many reactions can be merged or even deleted (Clement et al., 2021). This reduces the network size, thereby accelerating the enumeration of ECMs, without affecting the metabolic conversions of the network.

Clement et al. (2021) demonstrated how enumeration of ECMs can be applied. They have applied the method to a model of the core metabolism of *Escherichia coli* (Orth et al., 2010), to a genome-scale model of *Helicobacter pylori* (Thiele et al., 2005), and to a genome-scale model of *E. coli* (Reed et al., 2003). In the small example of the *E. coli* core model (95 reactions, 72 metabolites), they found 689 ECMs in total. This was contrasted with the total of 100 000 EFMs computed for the same model, showing that ECMs scale better than EFMs. They were also able to enumerate all ECMs of the *H. pylori* model on a minimal medium, "minII" defined by Thiele et al. (2005), and found a total of 874 236 ECMs. The full set of ECMs of the genome scale model of *E. coli* was never calculated, but all conversions (12 in total) involving oxygen and glucose to biomass were found within 14,3 hours. This displays another important feature of *ecmtool*: the "hide" option, which can disregard enumeration of certain metabolites in favor of scaling the analysis to larger models. It works by adding a reverse reaction of the metabolites one wants to hide, then imposing the steady-state assumption on the forward-reverse reaction pair of the hidden metabolite. This way, the metabolite can partake in reactions, but it will not be included in the final set of ECMs. This is true because imposing the steady-state constraint on a metabolite means the net conversion of a given metabolite will always be zero. If the conversion of a metabolite is set to zero, the axis disappears from the conversion cone, and is not included in the ECMs. Therefore, hiding metabolites makes ECM enumeration more scalable, by reducing the number of ECMs.

1.3.3 Minimal pathways

Minimal pathways (MPs) are found in a solution space with inhomogeneous constraints, like EFVs. The polyhedron is then projected to a subnetwork, which includes a subset of reactions of the full network. The remaining fluxes satisfy all constraints on reaction rates of the original network, but also only includes fluxes of the subnetwork. MPs are defined as the support minimal EFVs of this subspace, meaning that an MP consists of the smallest set of reactions in the subnetwork that need to be active (have non-zero flux) to satisfy all constraints on the whole network. Constraints would be violated if one of the reactions in an MP was turned off (Øyås & Stelling, 2020). MPs can be compared to ECMs by using the exchange reactions of a network as a subnetwork. As discussed in Section 1.3.2 ("Elementary conversion modes"), the internal metabolites are assumed to have a net conversion of zero. Therefore, only exchange reactions reflect overall conversions of the network.

MPs can be calculated using the software *mptool*. They are found by iteratively removing reactions from the subnetwork and minimizing subnetwork flux until no further reactions can be removed without affecting the functional requirements on the full network. The software also provides random sampling of MPs. The capabilities of MP analysis were demonstrated in the paper of Øyås and Stelling (2020). They used MP sampling to predict gene importance in a GEM of *Escherichia coli*, iJO1366 (Orth et al., 2011) and MP enumeration to predict metabolic interactions between human and microbial cells in a constrained subnetwork of a model of a microbial community in the human gut. By imposing additional constraints, complete enumeration of MPs representing metabolic interactions was made feasible, although at the cost of introducing bias. Lessening the constraints yields larger sets of MPs, but enumeration of all MPs becomes infeasible (Øyås & Stelling, 2020).

1.4 The link between elementary conversion modes and minimal pathways

The solution space of the MPs is considerably smaller than the solution space of ECMs. Figure 1.3 illustrates the steps required to obtain the solution space for both methods. MPs restrict the solution space by allowing inhomogeneous flux constraints and projecting the space of the full network onto the space of only the exchange reactions. The remaining space has the same number of dimensions as the conversion cone, but unlike the conversion cone can include upper and lower bounds in each dimension, as well as any other linear flux constraints. In addition, only flux vectors that are support minimal are found by *mptool*. The set of support-minimal vectors might describe a solution space with even fewer dimensions than the initial solution space if some fluxes in the solution space do not contribute to satisfying the constraints. It is expected that MPs should always contain fewer or the same number of metabolites as the ECMs, as illustrated in Figure 1.4.

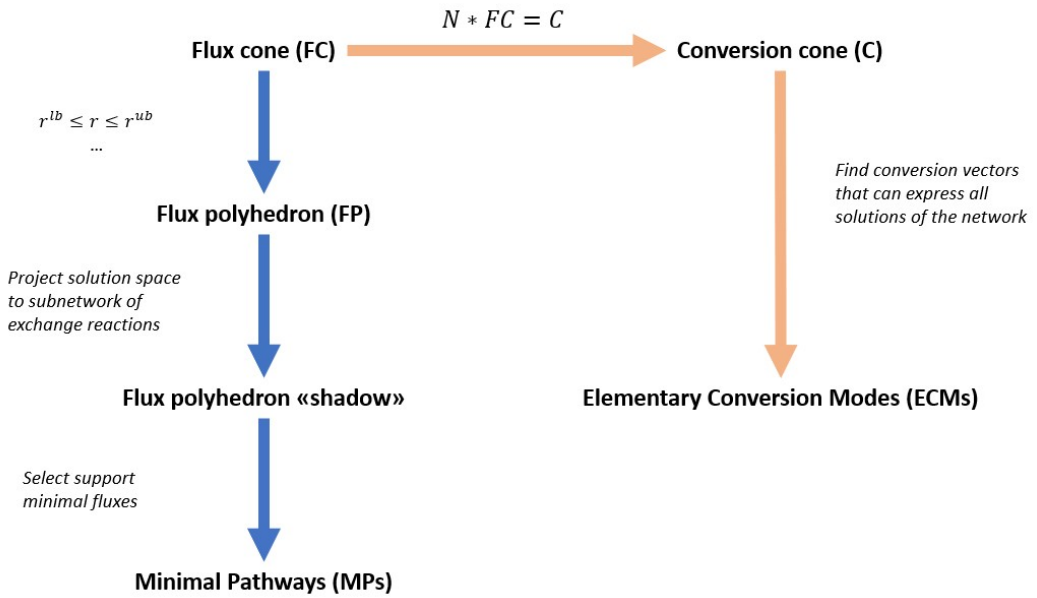

Figure 1.3: Schematic figure of the link between the solution spaces of MPs and of ECMs. The red arrows show the steps (simplified) from the flux cone to the conversion cone, and eventually to the ECMs. The blue arrows show the steps (simplified) from the flux cone to the flux polyhedron, and eventually to the MPs.

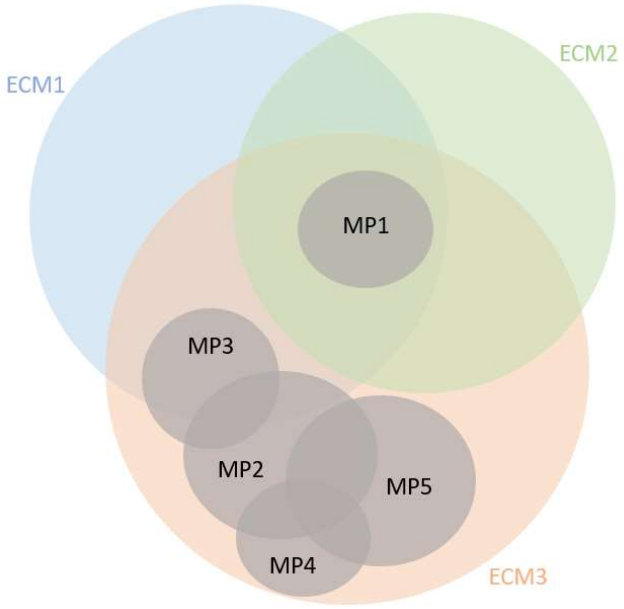

Figure 1.4 Schematic illustration of the relationships between MPs and ECMs. Each circle represents the set of metabolites present in an ECM or MP. Either there is a high number of MPs that are a subset of an ECM, which would look like the cluster of MPs in ECM3, or there could be a high number of ECMs are a superset of the same MP, which would look like the situation for MP1.

1.5 The aims and scope of this thesis

CBMM provides a framework for integrative analysis of molecular biology data and quantitative prediction of feasible phenotypic states (Heirendt et al., 2019). Methods of unbiased analysis can summarize all possible flux states of a cell, which is especially useful when the cell's objective is not clear. However, these methods are limited to smaller models or subsets of bigger models. Improving the scalability of characterization of metabolic exchanges to genome-scale or even multicellular models would be useful in understanding biological systems and to aid development of new biotechnology. Two methods in particular provide a solution to this challenge; enumeration of MPs (Øyås & Stelling, 2020) and enumeration of ECMs (Clement et al., 2021). However, these methods have yet to be widely applied to GEMs, and therefore their limits and capabilities are not well documented. This is the motivation for this thesis. Specifically, I have defined three main objectives:

1) Reproduce the results presented by Clement et al. (2021).
2) Compare the results from ECM enumeration with the results of MP enumeration.
3) Apply both methods to a large model that combines genome-scale models of multiple cells to test the limits of both methods.

2 Materials and methods

2.1 Software and models

The software packages *mptool* and *ecmtool* were applied, for enumeration of MPs and ECMs respectively, to the collection of models summarized in Table 2.1. All metabolite conversions were fully enumerated for e_coli_core and iIT341. Subsets of the metabolite conversions were enumerated for iJR904, using the "–hide" method in *ecmtool* and selecting an equivalent subset on *mptool*. For HS_BT_EC_FP_LL_LP_ST, MPs were randomly sampled because complete enumeration was infeasible. The following subchapters describe how each model was enumerated in detail.

Table 2.1: Metabolic models used for enumeration of ECMs and MPs in this thesis. Each row gives the name of the model, the model species, the number of reactions and the number of metabolites.

Model name	**Species**	**No. Reactions**	**No. Metabolites**	**Reference**
e_coli_core	*Escherichia coli*	95	72	(Orth et al., 2010)
iIT341	*Helicobacter pylori*	485	554	(Thiele et al., 2005)
iJR904	*Escherichia coli*	1075	761	(Orth et al., 2011)
HS_BT_EC_FP_LL_LP_ST	*Homo sapiens, Bacteriodes thetaiotaomicron, Escherichia coli, Faecalibacterium prausnitzii, Lactococcus lactis, Lactobacillus plantarum, Streptococcus thermophilus*	22343	13862	(Magnúsdóttir et al., 2017)

2.2 Enumeration of e_coli_core pathways

The model of *E. coli* core metabolism was downloaded from the BiGG database (King et al., 2016) and the enumeration was performed on a personal laptop (Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz and 8 GB RAM). Both *mptool* and *ecmtool* were installed according to the installation guides provided on their respective GitHub pages (Clement et al., 2021; Øyås & Stelling, 2020). In addition, an academic license was obtained for the Gurobi solver to enable enumeration for *mptool* (Gurobi Optimization LLC, 2022).

Both *ecmtool* and *mptool* were run on the default model without alterations. For enumeration of MPs, a subset of exchange reactions was selected, the lower bound of the biomass reaction was set to 0.1 h^{-1}, and the reversible exchange reactions were made irreversible by splitting them into two separate reactions before finding MPs. For enumeration, the options "iterative", "graph" and no preset bounds were used. When no bounds are set, the program performs FVA to obtain tight flux bounds.

For enumeration of ECMs, the example script provided on the GitHub page of the *ecmtool* project was used (Clement et al., 2021). Input and output metabolites were determined automatically before ECMs were enumerated. The metabolites exchanges were made irreversible by splitting reversible reactions into two separate irreversible reactions. Afterwards the whole network was compressed. This resulted in a stoichiometric matrix that could be used for ECM enumeration.

2.3 Enumeration of iIT341 pathways

Enumerations of all models larger than e_coli_core were executed on the Orion High Performance Computing Center at the Norwegian University of Life Sciences (NMBU) (Orion, 2022). All dependencies for *ecmtool* were installed directly on the HPC environment. CPLEX was used as the solver for *ecmtool* (Cplex, 2009). The git repository of Clement et al. (2021) was downloaded and used to run *ecmtool*. For *mptool*, Gurobi was installed with site-wide license (Gurobi Optimization LLC, 2022) on the HPC environment, and *mptool* itself was installed in a singularity container. *COBRApy* was used to inspect and edit the models when needed (Ebrahim et al., 2013).

The model iIT341 was enumerated using both methods. The model file in the GitHub repository of *ecmtool* was used. For this model, enumerations of two different sets of metabolites were performed: A full enumeration of all metabolites in the model on *minII* medium presented by Thiele et al. (2005), and another enumeration where outputs were hidden, thereby only analyzing the inputs. *ecmtool* uses an internal ID system, which is printed out each time the program runs, to label metabolites. These IDs were used to define which metabolites should be considered as outputs, inputs and hidden in the runscript. The inputs for *ecmtool* did not deviate from the corresponding runscripts found in the supplementary literature of the paper of Clement et al. (2021). A new round of enumerations was performed, on both the full model and of only inputs, after a bug was discovered in the program. The runscripts were nearly identical, the only difference was the inclusion of the command "—remove_infeasible=true". This command was included after the bug was discovered and fixed (although further testing is required), and it was set to "false" by default such that the code where the bug existed did not execute unless specified.

Enumeration of MPs was done on the same model. The full set of MPs was enumerated by imposing a lower bound on the biomass reaction. Because input and output metabolites were defined with a special ID system for both enumerations of iIT341 with *ecmtool*, it was necessary to ensure that the same subset of metabolites were defined correctly as inputs and outputs for MP enumeration. To achieve this, the metabolite IDs used for defining inputs and outputs for ECM enumeration were used to obtain the subset of the corresponding reactions for MP enumeration. The flux bounds of

glucose, water, ammonium, S-adenosyl-L-homocysteine, S-adenosyl-4-methylthio-2-oxobutanoate and protoheme were changed to allow uptake and/or secretion according to how they were defined for the ECM enumeration. The reactions were subsequently made irreversible, and MPs were enumerated with settings “iterative” and “graph”. Almost the same procedure was used for enumeration of iIT341 with hidden outputs, the only difference being the subset that was selected.

2.4 Enumeration of iJR904 pathways

Enumeration of ECMs was done with the iJR904 model provided on the GitHub repository of *ecmtool*. Enumeration was attempted multiple times but was not completed. On the last attempt nearly the same arguments were passed to the program as detailed in the supplementary literature of the paper of Clement et al. (2021). The only differences were: 1) The addition of argument “—remove_infeasible=true” after the previously mentioned bug was fixed, 2) Using the direct method instead of the indirect method, and 3) reserving 10 nodes with 32 cores each for the job instead of 20 nodes with 16 cores each. The enumeration ran for 8 days and 11 hours before it was shut down. Before the job was stopped, the output file was examined. The number of computations left was not exactly known but based on the amount of time it had already used on the previous computations there was not enough time available to keep the job running.

2.5 Enumeration of human-microbe exchanges

The human-microbe model is made of six separate metabolic models. These models were downloaded from the GitHub page of *mptool* and processed similarly as described by Øyås and Stelling (2020). The combined model consists of five microbe models and one model of a human cell (species of each model is listed in Table 2.1). Some additional modifications to the models were made. Non-essential sink and demand reactions were turned off for the microbiota, while all sink and demand reactions were turned off completely for the human model. In addition, a lower bound of 0.1 h^{-1} was set for the biomass reaction in the microbial models and a lower bound of 0.1 mmol gDW^{-1} h^{-1} was set for the ATP maintenance reaction (which consumes ATP) in the human model. The models were constructed to have a shared compartment where the cells exchange metabolites among each other. These metabolites are marked as internal exchange metabolite reactions, or “IEX”. FVA was conducted before MP enumeration to obtain tight flux bounds. Internal exchange reactions, along with external exchange reactions for metabolite uptake were used as the subset for pathway analysis. The MPs were found using the iterative method and random sampling. 1000 MPs were sampled initially: An array of 10 parallel jobs was sent to Orion, where 100 MPs were sampled per core. Next, 1000 more MPs were sampled using only one core for all 1000 samples. Lastly, another 8000 MPs were sampled by submitting an array of 8 jobs, each sampling 1000 MPs on one core each.

Enumeration of ECMs was attempted on this model. The plan was to enumerate the metabolite conversions of each of the models separately and compare which metabolites were produced by the different microbes. ECM enumeration of the human model was not considered. Enumeration of ECMs on the *Lactobacillus_plantarum_WCFS1* model (Magnúsdóttir et al., 2017) was started. Most settings were run on default, but “--hide_all_in_or_outputs” was set to “inputs”. The enumeration was not completed.

2.6 Data analysis

The output files produced by *ecmtool* and *mptool* are different. For each ECM, *ecmtool* calculates the stoichiometry of all metabolites that are included in the ECM. If the coefficient is negative, the metabolite is consumed by the cell, while a positive coefficient signifies secretion of the metabolite by the cell. The output of mptool however, only considers whether the metabolite is produced or consumed in the first place. In addition, all MPs will produce biomass due to the constraints imposed on the network. These constraints are not present for calculation of ECMs ergo, an ECM does not necessarily produce biomass. To be able to compare MPs and ECMs directly, only ECMs that produced biomass and were unique when only considering the reaction direction of each metabolite were used in further analysis. Thus, each MP and ECM was transformed to a combination of metabolites that are either consumed or produced by the cell. Consumption and production of the same metabolite was considered as two separate metabolites.

Hierarchically clustered heatmaps were produced of the output of both enumerations for the iIT341 and e_coli_core models. In addition, MPs and ECMs were clustered with respect to each other by transforming each MP and ECM into sets of metabolites as described above. Jaccard index was calculated for each combination ECM and MP, by dividing the intersection of the sets with the union of the sets, to measure the similarity of each pair of MP/ECM. The indices were subsequently hierarchically clustered. All clustered heatmaps were generated using *seaborn's* clustermap (Waskom, 2021). Default metric and method were used for clustering, Euclidean and average linkage respectively. The set of ECMs for all metabolite conversions of iIT341 was too large for clustering within reasonable time. A regular heatmap was therefore generated for visualization. For comparison between MPs and ECMs, the number of MPs that were a subset of an ECM was counted for each ECM. Likewise, the number of ECMs that were a superset of an MP was counted for each MP. The relation between these counts is illustrated in Figure 1.4. After it was observed that no ECMs included production of H^+, but some of the MPs did, the H^+ metabolite was removed from all sets and subsets were counted anew. The counts of subsets for each MP and each ECM were also performed on the results for e_coli_core and iIT341 inputs only. The distributions of counts were visualized in histograms.

2.7 Data and code availability

All models and code can be found on the GitLab repository of this master thesis (Wedmark, 2022): https://gitlab.com/YlvaKaW/masteroppgave-pathway-analysis

All code and data related to the analyses of the e_coli_core model can be found under the folder "simple_networks". Code and data related to the analyses of all other models can be found under the folder "genome_scale_networks".

3 Results

3.1 Enumeration and comparison MPs and ECMs

Table 3.1 summarizes the results of all enumerations that were carried out with both *ecmtool* and *mptool*. Both model size and subset size affect the total number of pathways or conversion modes that are found, and the time it takes to compute all of them. The number of ECMs seems to grow faster with subset size and model size than the number of MPs. The total time required for enumeration was also consistently lower for MPs than for ECMs.

Table 3.1: Summary table of ECM and MP enumerations in e_coli_core, iIT341 and iJR904. Model size is determined by the number of reactions in the model, subset size is the number of metabolites that were chosen for enumeration, no. ECMs and no. MPs are the total number of ECMs, and MPs found, respectively, and the two last columns describe the total time in seconds it took to complete the enumeration for each method.

	Model size	Subset size	No. ECMs	No. MPs	Time ECMs (s)	Time MPs (s)
e_coli_core	95	20	689	33	7,49	0,25
iIT341 inputs	554	17	3 652	2	7 603,36	0,02
iIT341 minII	554	77	874 236	1 382	14 961,26	84,74
iJR904	1 075	2	N/A	1	>730 800	<0,00

3.1.1 Enumeration of MPs and ECMs in e_coli_core

For the e_coli_core model a total number of 689 ECMs were found. The same number of ECMs were found by Clement et al. (2021). MPs and ECMs are visualized by a hierarchically clustered heatmap in Figures 3.1 and 3.2, respectively. Each ECM and MP is an alternative collection of inputs and outputs of the network. After filtering for biomass and stoichiometry (Section 2.6), the total number of ECMs was 346. Common for all MPs and all ECMs is that they produce H^+ and consume phosphate, D-glucose and ammonium. They can grow without oxygen, but often produce ethanol in those situations. The core metabolism of *E. coli* can, according to both MP and ECM analysis, both consume and produce CO_2.

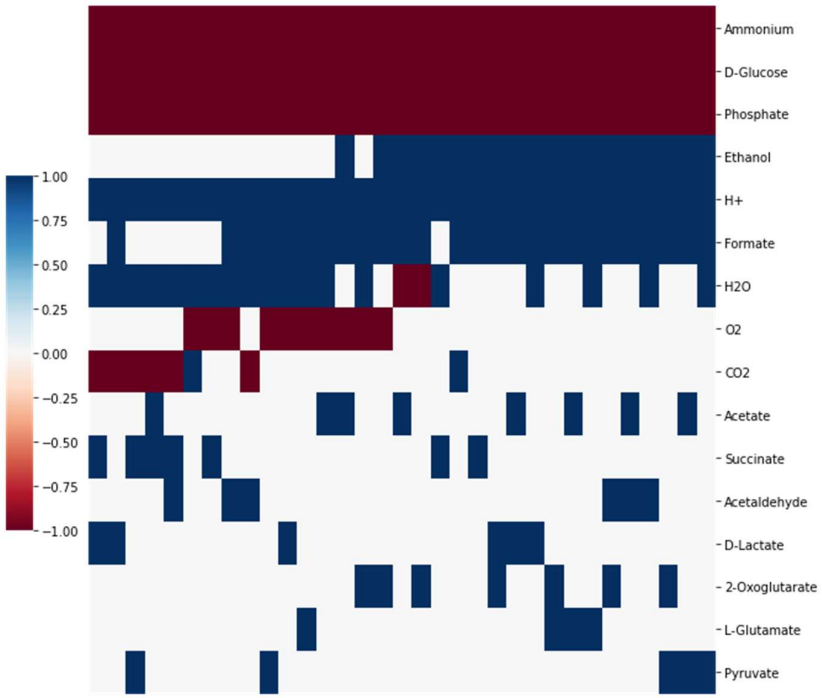

Figure 3.1: Hierarchically clustered heatmap of all MPs found for the e_coli_core model. Each column represents an MP and each row represents a metabolite. The color signifies whether the metabolite is consumed (red), produced (blue) or not included (white) for each MP.

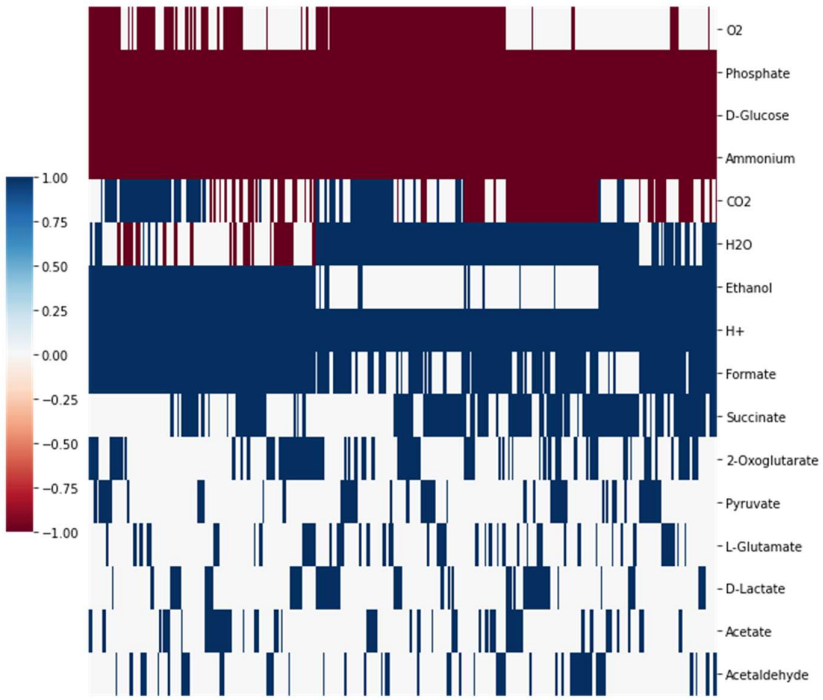

Figure 3.2: Hierarchically clustered heatmap of ECMs found for the e_coli_core model. The ECMs were processed as described in (Section 2.6). Each column represents an ECM, and each row represents a metabolite. The color signifies whether the metabolite is consumed (red), produced (blue) or not included (white) for each ECM.

MPs and ECMs were also clustered with respect to each other (Figure 3.3). Only MPs that were a subset of an ECM are colored for each ECM the heatmap. The majority of ECMs have at least one MP as a subset. Furthermore, each MP is a subset of multiple MPs. To the left of the heatmap there is a clear division between ECMs that are a superset of only one or two MPs, and ECMs that are supersets of more than two MPs.

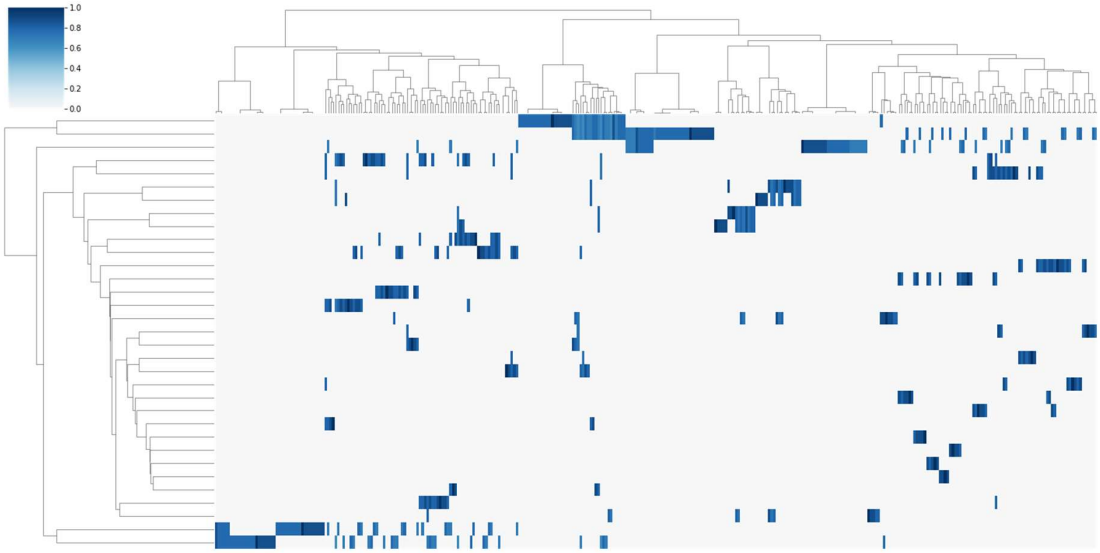

Figure 3.3: Hierarchically clustered heatmap of ECMs and MPs for exchange metabolites of e_coli_core. Each row represents an MP, while the columns represent an ECM. The color signifies each pair's Jaccard index. The darker the color, the more similar the MP/ECM pair is. Only MPs that were subsets of an ECM are colored.

Counts of MPs that were subset of ECM for each ECM, and for each MP were obtained for e_coli_core and are presented in Figure 3.4. Maximal number of MPs that are subset of one ECM is six. On the contrary, each MP is subset of ten or more ECMs. All ECMs, except for two, have either one or more MP(s) as a subset. The two ECMs that did not have an MP subset included the metabolites: H^+, formate, consumption of phosphate, H_2O, consumption of D-glucose, consumption of CO_2, consumption of ammonium and succinate. The only difference was the metabolite 2-oxoglutarate.

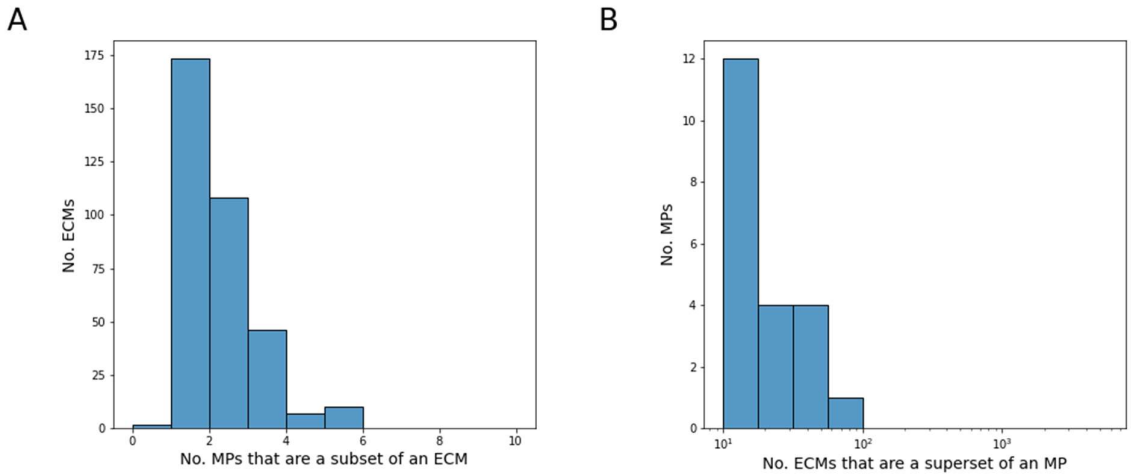

Figure 3.4: Counts of subsets and supersets of metabolites in ECMs and MPs respectively for e_coli_core. A: Number of MPs that were a subset of each ECM for the e_coli_core model. The x-axis represents the number of MPs that were a subset of an ECM. The y-axis is the number of ECMs which have a given number of MPs as its subset. B: Number of ECMs that were a superset of each MP for the e_coli_core model. The x-axis represents the number of ECMs that were a superset of an MP. The y-axis is the number of MPs which have a given number of ECMs as its superset.

3.1.2Enumeration of ECMs and MPs in iIT341 (only inputs)

The initial ECM enumerations of iIT341 were not consistent with results presented in earlier work (Clement et al., 2021). For the enumeration of all metabolite conversions 317 974 ECMs were initially found, while 302 ECMs were found for the enumeration of inputs only. The number of ECMs reported by Clement et al. (2021) was 874 236 ECMs for all metabolite conversions and 3 652 for only inputs. Not only were fewer ECMs found in total, but none produced biomass. This result was not expected, considering that the model grew on medium that supports growth (Thiele et al., 2005). The results were sent to the corresponding author of the *ecmtool* paper, which eventually led to the discovery of a bug in one of the network compression steps. After the bug was patched, new enumerations on iIT341, both on all metabolite conversions in the network and inputs only were performed. This time the same number of ECMs were found as were reported previously in both cases.

Clustered heatmaps of the results from enumeration of only inputs of iIT341 are shown below (Figure 3.5). After filtering the ECMs as described in Section 2.6, 32 ECMs were left. Metabolites such as sulfate, D-glucose, H+ and H2O are included in the set of ECMs, but not in the set of MPs. In fact, when these metabolites are removed, the set of ECMs reduces to the set of MPs. The only difference between the two MPs that were found was consumption of the two isomers of alanine. A clustered

heatmap of differences between ECMs and MPs for the inputs of the iIT341 model (Figure 3.6) were computed similarly to Figure 3.3. This heatmap clearly divides all ECMs in two clusters. Half of the ECMs are a superset of the MP that consumes L-alanine, while the other half are supersets of the MP that consumes D-alanine.

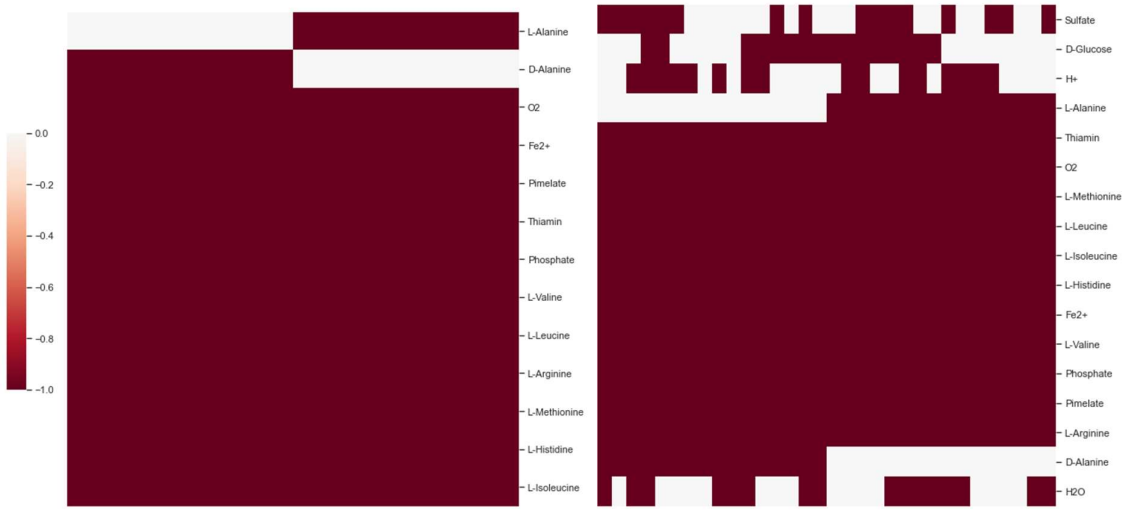

Figure 3.5 Hierarchically clustered heatmaps of all MPs found (left heatmap) and all ECMs found (right heatmap) for the iIT341 model where only inputs are considered. Each column represents an MP/ECM and each row represents a metabolite. The color signifies whether the metabolite is consumed (red) or not included (white) for each MP/ECM.

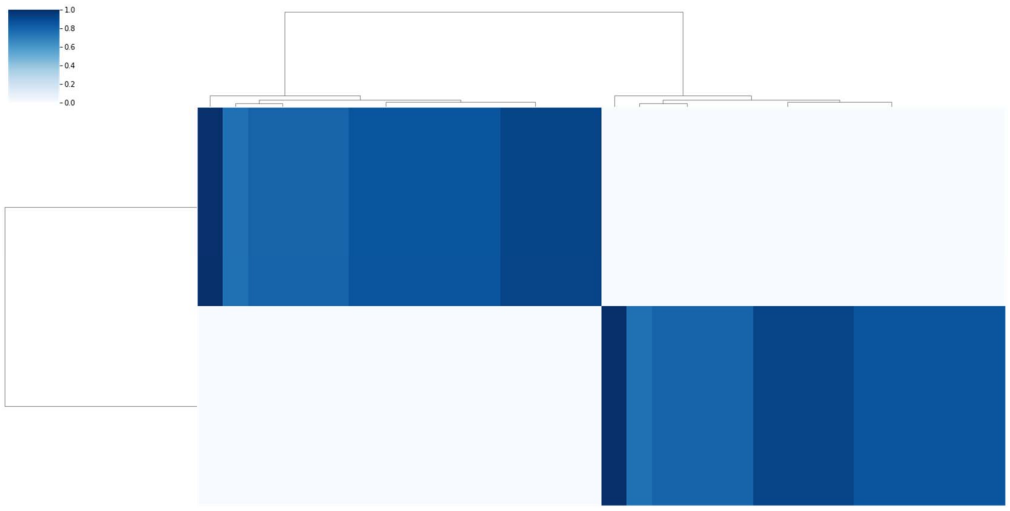

Figure 3.6: Hierarchically clustered ECMs against MPs for input metabolites of iIT341. Only MPs that were subsets of an ECM are colored. Each row represents the distance of each MP, while the columns represent the distance of each ECM. The color signifies each pair's Jaccard distance, which was computed after turning each ECM and each MP to a set of metabolites. The darker the color, the more similar the ECM/MP pair is.

3.1.3 Enumeration of MPs and ECMs in iIT341 (*minII* medium)

The results from enumeration using all exchanged metabolites of the iIT341 model are shown below. Figure 3.7 shows a heatmap of ECMs, while Figure 3.8 shows a clustered heatmap of MPs. After filtering ECMs for biomass and stoichiometry (Section 2.6), the total number of ECMs for all exchange reactions was 125 020. The metabolites that were always consumed in the enumerations of only inputs (Figure 3.5) were also always consumed in the enumerations of all exchange metabolites (Figure 3.6, Figure 3.7). H^+, however, is consumed in some MPs, and produced in others (Figure 3.8). This is not the case for ECMs, where it is either not included or it is consumed (Figure 3.7).

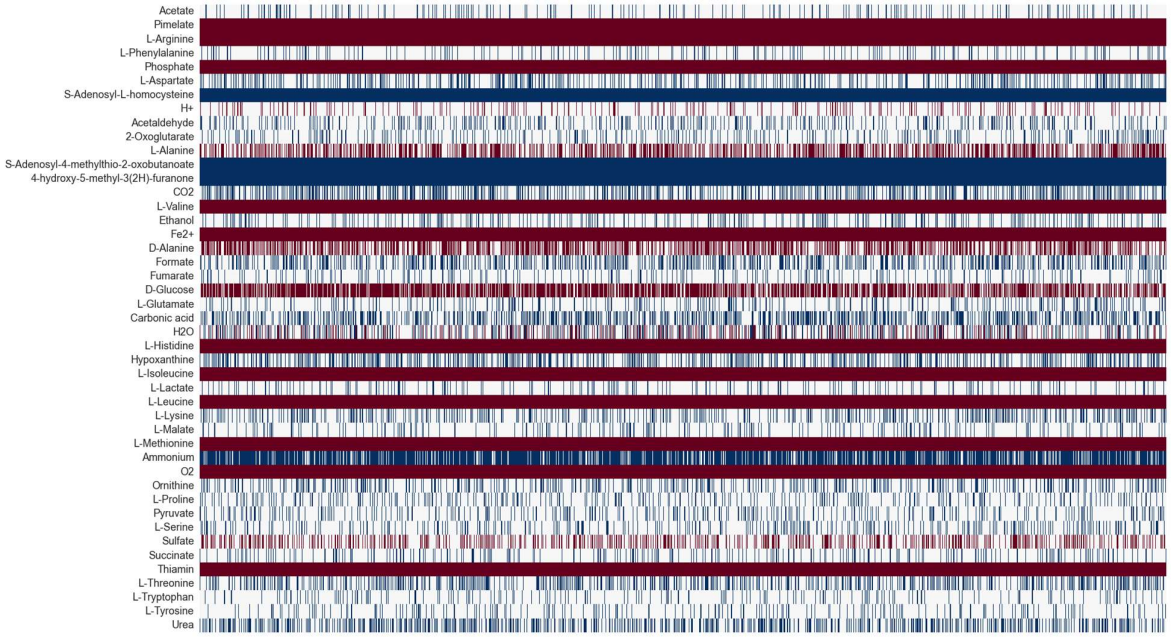

Figure 3.7 Heatmap of ECMs found using all exchanged metabolites for the iIT341 model. The ECMs were filtered for biomass and stoichiometry (Section 2.6). The ECMs were not clustered due to the size of the dataset. Each column represents an ECM, and each row represents a metabolite. The color signifies whether the metabolite is consumed (red), produced (blue) or not included (white) for each ECM.

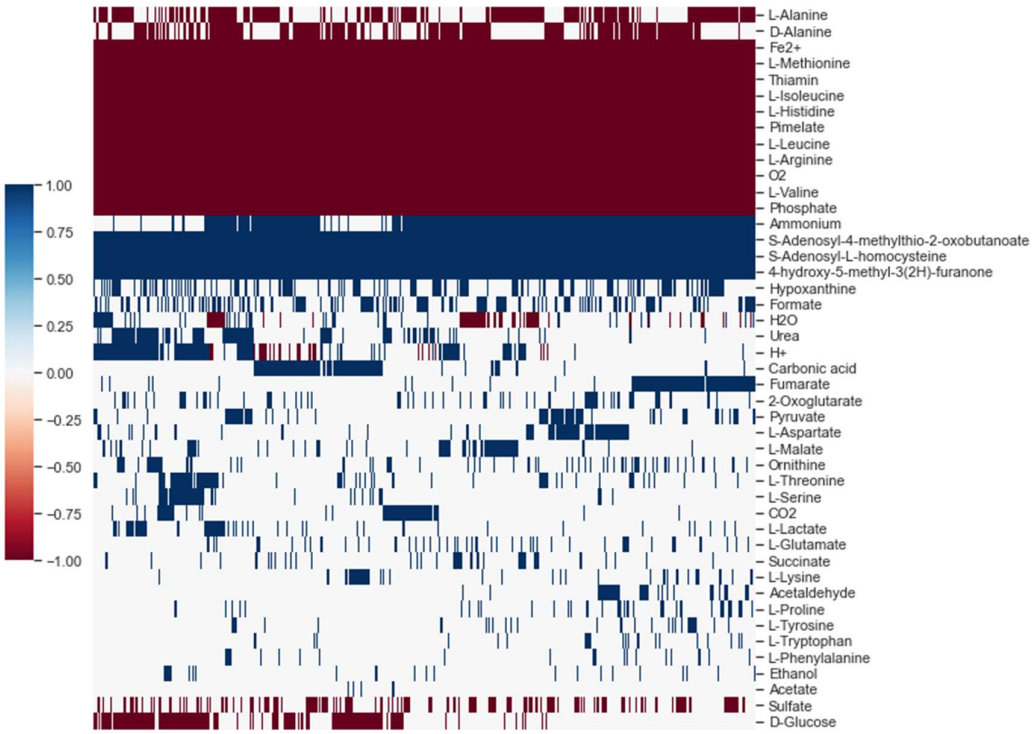

Figure 3.8: Hierarchically clustered heatmap of all MPs found using all exchange reactions of the iIT341 model on minII medium. Each column represents an MP and each row represents a metabolite. The color signifies whether the metabolite is consumed (red), produced (blue) or not included (white) for each MP.

Because the ECMs of all metabolite conversions of iIT341 were not clustered, the comparison was summarized by counting subsets of ECMs and supersets of MPs (Section 2.6). Initially, there were 12 954 of 125 020 ECMs that had no MP subsets, while 520 of 1 382 MPs had no ECM supersets. After H^+ was removed from all sets, 468 ECMs did not have any MPs as a subset, and 152 MPs did not have any ECMs as a superset. The distribution of the number of MPs that were a subset of each MP and the distribution of the number of ECMs that were a superset of each MP are shown in Figure 3.9. An ECM is a superset of 2 or 3 MPs on average (Figure 3.9A). An MP on the other hand, is often a subset of 100-1000 ECMs (Figure 3.9B). An ECM generally has fewer MPs as a subset than an MP has ECMs as a superset.

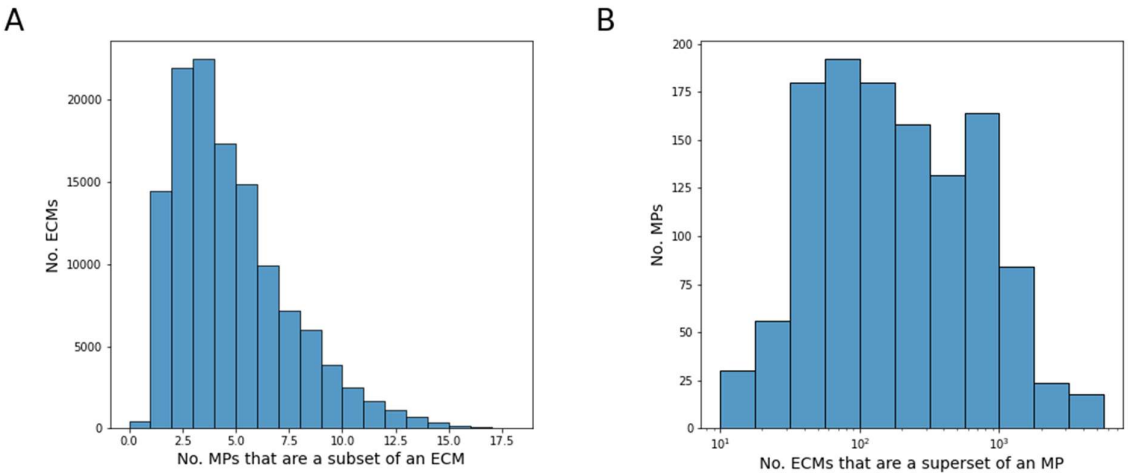

Figure 3.9: Counts of subsets and supersets of metabolites in ECMs and MPs respectively for iIT341. A: Number of MPs that were a subset of each ECM for the iIT341 model. The x-axis represents the number of MPs that were a subset of an ECM. The y-axis is the number of ECMs which have a given number of MPs as its subset. B: Number of ECMs that were a superset of each MP for the iIT341 model. The x-axis represents the number of ECMs that were a superset of an MP. The y-axis is the number of MPs which have a given number of ECMs as its superset. The MPs that did not have an ECM superset are not shown.

3.2 Analysis of human-microbe interactions

Enumeration of ECMs and MPs were attempted for the human-microbe interaction model, HS_BT_EC_FP_LL_LP_ST, but ECM enumeration was never completed. MP enumeration was also infeasible in the unconstrained model, but a random sample of 10 000 MPs was obtained. Histograms of the time it took to find each MP, and the total time it took for each job is detailed in Appendix A1 and A2, respectively.

MPs were also enumerated in the same model, with additional constraints as described by Øyås and Stelling (2020). In the constrained model, a complete enumeration of 252 MPs was obtained. The metabolites that were exchanged in this model are shown in Figure 3.10 A. This figure shows clear interactions between human and the other microbes in the model. Bacteria produce a handful of essential amino acids and other metabolites, which are consumed by the human model. In contrast, Figure 3.10 B shows the interactions of the same metabolites (except malate and propionate which were not exchanged) captured by 10 000 random MP samples when the model is not constrained. The same metabolites are exchanged to a significantly lesser degree. Exchanges between the microbes of isoleucine and niacin are present in less than half of all the sampled MPs. The rest of the metabolites are exchanged in <2% of all MPs.

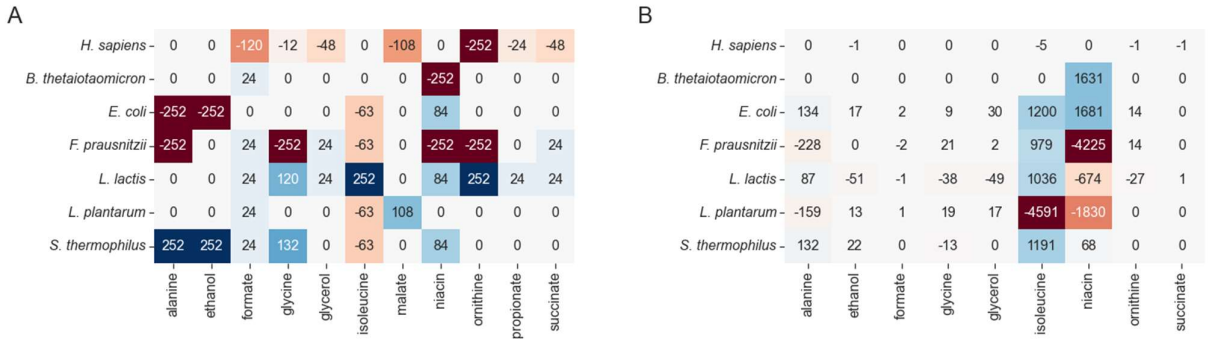

Figure 3.10: Heatmaps of metabolite exchanges between models of a human cell and human gut microbes. For both heatmaps, each row shows one species of the model, and each column shows a metabolite. The number in each square is the number of MPs a metabolite appeared in for a model of given species. Positive numbers (colored in shades of blue) signifies that the metabolite was produced by the species, negative numbers (colored in shades of red) signifies that the metabolite was consumed by the species. A: Heatmap of metabolites that were exchanged between the organisms when the model was constrained (Øyås & Stelling, 2020). B: Heatmap of the same metabolites from randomly sampled MPs from the same model without constraints. Metabolites propionate and malate are not shown, as there were no exchanges of these metabolites.

4 Discussion

The results presented here show the importance of reproducing others' results. An incongruence with the results from ECM enumeration in iIT341 by Clement et al. (2021) led to the discovery of a bug. The cause of the results from the initial enumerations for iIT341 was a numerical error found in one of the network compressions steps in *ecmtool*. This compression step is detailed in the supplementary literature of Clement et al. (2021) in Section 5.1. The network is compressed by first calculating the nullspace (solutions that satisfy $N_{int}\ r = 0$) of matrix N_{int} that contains all internal reactions. They can then identify reactions (columns) where the rate of each metabolite (row) is zero, implying that this reaction is not feasible when steady state is imposed. However, the reaction rates are based on the ratio of stoichiometric coefficients, and will therefore converge to zero, but never be zero. Before the bug was discovered, a reaction rate of $<10^{-8}$ was deemed to be zero. Realistically, reaction rates below 10^{-8} are very unlikely, but if the ratio of the stoichiometric coefficients is very large, the reaction rates become very small (de Groot, D. H., personal communication, 12th April 2022). This explains why significantly fewer ECMs were found, as the network compression step caused some reactions to drop out even though they were part of the null space, i.e., feasible under steady state. As the developers found this error, they were able to patch it and have warned users of the program about it.

The enumerations performed after *ecmtool* successfully reproduced the results of the ECM enumerations performed by Clement et al. (2021). The number of ECMs found matched the number of ECMs reported previously on both e_coli_core and iIT341 models. The data could not be compared directly, but as the same commands and the same model were used to perform the analyses it is reasonable to expect that the same ECMs were obtained. On the other hand, attempts were made to enumerate ECMs of glucose, oxygen, and biomass in the iJR904 model. This enumeration was expected to take about 14 hours, as this was the time it took for Clement et al. (2021). However, the enumeration was not done after 8 days had passed. It is not clear exactly what could have caused the large difference in computation time, but it was clear by the output file of the job that the number of computations needed was too large for the resources allocated to the job to finish the job within reasonable time. Even though the same number of cores were allocated to the same analysis compared to the analysis of Clement et al. (2021), the exact specifications of each core can differ. This could explain the difference in computational time.

Even though attempts were made to analyze ECMs and MPs for multiple models, analysis of both MPs and ECMs were only completed in two models, e_coli_core and iIT341. Therefore, all conclusions regarding the relations between ECMs and MPs will be based on the results of the analyses of these models. First, the full set of MPs was consistently smaller than the set of ECMs. This was expected as the solution space of the MPs is smaller compared to the solution space of the flux cone, due to the additional constraints and the requirement of support minimality of MPs. Therefore, there are fewer feasible solutions, which results in fewer MPs than ECMs.

The analysis of e_coli_core suggest that the set of MPs summarize the metabolite conversions found in the set of ECMs. The set of ECMs and MPs obtained for e_coli_core seem to reflect the same metabolic conversions (Figures 3.1 and 3.2). For example, there was consensus between MPs and ECMs of the essential metabolites, meaning the metabolites that were always consumed. This is as expected, given that essential metabolites always need to be consumed by definition. The two methods validate each other in this regard, as both were tasked to enumerate metabolic conversions of the same network and have found similar conversions. When the MPs and ECMs were clustered with respect to each other (Figure 3.3), almost every ECM was found to have at least one MP as a subset. This may indicate that MPs are able to capture the variation of conversions provided by ECMs, but still have fewer vectors in total. Moreover, most of the ECMs had multiple MPs as subsets. The union of MPs that are a subset of ECM may capture the full ECM, or the set of MPs may reflect alternative ways of expressing the same conversion when tighter flux bounds are imposed on the model, and only considering support minimal pathways. When comparing the solution space from the conversion cone generated by ECMs to the flux polyhedron generated by the MPs, the ECM might be reduced to either one or the other MP, analogous to how multiple flux modes have the same overall metabolite conversion (ECM). There were two ECMs that had no MPs as a subset. These ECMs were very similar, the only difference being the inclusion of 2-oxogluterate. This might imply that the MPs were not able to capture all the combinations of metabolic conversions that were present in the ECMs. Some degree of bias is introduced when determining the flux bounds, because they are calculated based on some cellular state of the network (Gudmundsson & Thiele, 2010). Some conversions might have been lost by imposing additional flux constraints on the network, or by selecting only support-minimal pathways. Otherwise, it could be that the MPs do not fully summarize all possible metabolite exchanges.

The same connections, and an additional connection, between ECMs and MPs are highlighted in the findings of enumerations using the input metabolites of iIT341 (Figure 3.6). They make it apparent that the ECMs cluster in groups where either L-alanine or D-alanine is consumed. In contrast to the e_coli_core analyses, the set of ECMs includes more metabolites than the set of MPs. The MPs are a set of support minimal vectors, in contrast to ECMs which are not support minimal. This implies that the metabolites that are present in the ECMs, but not present in MPs can be dropped without cutting the flow of the network.

Expanding the analysis of iIT341 further to a larger set of metabolites yielded interesting results. 12 954 of the ECMs did not have any MP subset, and 520 of the MPs did not fit as a subset in any of the ECMs. One explanation for this discrepancy is that no ECM included production of H^+, while several MPs did. The set of metabolites of all MPs should be either equal to or less than the set of metabolites of all ECMs, which is supported in the findings for e_coli_core and consumed metabolites of iIT341. If iIT341 can produce H^+, it implies that the ECMs does not report all possible conversions. It is not clear from the results whether this is the case or what could cause these results in the first place. Therefore, it is challenging to find the root cause of this inconsistency. However, this reasoning was further explored by removing H^+ from all sets and comparing again. The number of unmatched MPs and ECMs both decreased, but there were still some MPs that were not a subset of an ECM, and ECMs that were not a superset of any MP. As for the case of the e_coli_core analysis, this could indicate that the MPs were not able to fully summarize all possible metabolite exchanges. On the other hand, the consistent mismatch of ECMs and MPs could also be explained by the lack of H^+ production in the ECMs. Assuming again that the iIT341 model should produce H^+, a different set of ECMs would have been obtained. Each ECM in this set could hypothetically have at least one MP subset. The MPs with missing ECM supersets, and the ECMs with missing MP subsets could then be a result of the ECMs not being properly enumerated.

Enumeration of ECMs was not completed for the human-microbe model but it was thought to be possible to enumerate each microbe individually. To limit the amount of time needed to enumerate the models even further, only outputs were analyzed, as it was considered most biologically interesting to analyze what the human gut can consume from the microbes. The reasons for the failed attempt at enumeration of ECMs of the iJR904 model applies to the gut microbe models as well. Furthermore, both the set of metabolites chosen for enumeration and the number of models that would have to be enumerated are considerably larger than the case for the iJR904 model. It made the ECM enumeration of this model even more challenging. Nevertheless, given enough time and resources it might be possible to enumerate the ECMs of all models.

Although 10 000 MPs were sampled from the HS_BT_EC_FP_LL_LP_ST model, very few exchanges between the microbes and the human model were found in contrast to the exchanges found for the constrained model (Figure 3.10). It is important to have a large enough sample size to capture the full range of possible pathways. For random flux sampling, one way to know when the sample size is sufficiently large could be to check if the flux distributions have converged (Herrmann et al., 2019). Similarly, it could be investigated if the MPs had converged. For this instance, it would take multiple rounds of randomly sampling 10 000 MPs to check. Nevertheless, the sample size cannot be known beforehand. All things considered, the discrepancy between the results of the constrained model and the unconstrained model may be explained by the constraints themselves or an insufficient sample size. The number of alternative flux routes could be relatively large compared to the number of flux routes requiring the human to exchange metabolites with the microbes. Thus, a too small sample size would not accurately reflect the full variation in metabolite exchanges, both between the human and the microbes and with their environment. On the other hand, assuming that the flux distribution had converged, the difference between the exchanges observed for the constrained model and the unconstrained model would likely occur due to the constraints themselves. The human and the microbes would be forced to interact with each other, even when they would not do it otherwise.

5 Conclusions and outlook

The methods of unbiased analysis are not yet readily available for larger models. The biggest hurdle in this field is the combinatorial explosion of possible metabolic conversions as the model size increases (Klamt & Stelling, 2002). Despite this problem, great advances have been made. Of the two types of analyses explored in this thesis, MPs scale to larger models than the ECMs do. The restriction of solution space allowed for further scalability. However, the more constraints that are imposed, the more important it is that each constraint is both reasonable and experimentally backed. Both enumeration methods had their unique challenges, however. The enumeration of ECMs worked well on models with <1 000 reactions, but the computational power required to calculate conversions for larger models was too great for this project. The reproducibility of the method was tested, and a bug was discovered and subsequently fixed (further testing is required), highlighting the importance of reproducibility in research. Although the MP analyses scaled better than the ECM analyses, it is uncertain whether the random MP sampling of the human-microbe interactions revealed all variation in flux distributions. To combat this problem, further work could focus on finding rules that dictate how many samples are needed for random sampling. Further work could also be directed to making both methods more scalable. In this regard, combining the two methods might be a good place to start, making unbiased analysis more applicable to larger models.

References

Aurich, M. K., Fleming, R. M. T., & Thiele, I. (2016). MetaboTools: A Comprehensive Toolbox for Analysis of Genome-Scale Metabolic Models [Protocols]. *Frontiers in Physiology*, *7*. https://doi.org/10.3389/fphys.2016.00327

Clement, T. J., Baalhuis, E. B., Teusink, B., Bruggeman, F. J., Planque, R., & De Groot, D. H. (2021). Unlocking Elementary Conversion Modes: ecmtool Unveils All Capabilities of Metabolic Networks. *Patterns*, *2*(1). https://doi.org/ARTN 100177 10.1016/j.patter.2020.100177

Cplex, I. I. (2009). V12. 1: User's Manual for CPLEX. *International Business Machines Corporation*, *46*, 157.

de Mas, I. M., Torrents, L., Bedia, C., Nielsen, L. K., Cascante, M., & Tauler, R. (2019). Stoichiometric gene-to-reaction associations enhance model-driven analysis performance: Metabolic response to chronic exposure to Aldrin in prostate cancer. *Bmc Genomics*, *20*(1). https://doi.org/ARTN 652 10.1186/s12864-019-5979-4

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*(1), 74. https://doi.org/10.1186/1752-0509-7-74

Gottstein, W., Olivier, B. G., Bruggeman, F. J., & Teusink, B. (2016). Constraint-based stoichiometric modelling from single organisms to microbial communities. *Journal of the Royal Society Interface*, *13*(124). https://doi.org/ARTN 20160627 10.1098/rsif.2016.0627

Gu, C. D., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*. https://doi.org/ARTN 121 10.1186/s13059-019-1730-3

Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, *11*(1), 489. https://doi.org/10.1186/1471-2105-11-489

Gurobi Optimization LLC. (2022). Gurobi Optimizer Reference Manual. https://www.gurobi.com

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., Magnusdóttir, S., Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Clancy, C. M., Modamio, J., Sauls, J. T., Noronha, A., Bordbar, A., Cousins, B., El Assal, D. C., Valcarcel, L. V., Apaolaza, I., Ghaderi, S., Ahookhosh, M., Ben Guebila, M., Kostromins, A., Sompairac, N., Le, H. M., Ma, D., Sun, Y., Wang, L., Yurkovich, J. T., Oliveira, M. A. P., Vuong, P. T., El Assal, L. P., Kuperstein, I., Zinovyev, A., Hinton, H. S., Bryant, W. A., Aragón Artacho, F. J., Planes, F. J., Stalidzans, E., Maass, A., Vempala, S., Hucka, M., Saunders, M. A., Maranas, C. D., Lewis, N. E., Sauter, T., Palsson, B. Ø., Thiele, I., & Fleming, R. M. T. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nature protocols*, *14*(3), 639-702. https://doi.org/10.1038/s41596-018-0098-2

Herrmann, H. A., Dyson, B. C., Vass, L., Johnson, G. N., & Schwartz, J.-M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Systems Biology and Applications*, *5*(1), 32. https://doi.org/10.1038/s41540-019-0109-0

Kim, O. D., Rocha, M., & Maia, P. (2018). A Review of Dynamic Modeling Approaches and Their Application in Computational Strain Optimization for Metabolic Engineering [Review]. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01690

King, Z. A., Lu, J., Drager, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., & Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515-D522. https://doi.org/10.1093/nar/gkv1049

Klamt, S., Regensburger, G., Gerstl, M. P., Jungreuthmayer, C., Schuster, S., Mahadevan, R., Zanghellini, J., & Muller, S. (2017). From elementary flux modes to elementary flux vectors:

Metabolic pathway analysis with arbitrary linear flux constraints. *Plos Computational Biology*, *13*(4). https://doi.org/ARTN e1005409 10.1371/journal.pcbi.1005409

Klamt, S., & Stelling, J. (2002). Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports*, *29*(1-2), 233-236. https://doi.org/Doi 10.1023/A:1020390132244

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, *10*(4), 291-305. https://doi.org/10.1038/nrmicro2737

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T., & Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, *35*(1), 81-89. https://doi.org/10.1038/nbt.3703

Nelson, D. L., & Cox, M. M. (2017). *Lehninger Principles of Biochemistry* (7th ed.). Macmillan Learning.

Orion, H. P. C. C. (2022). *Norwegian University of Life Sciences (NMBU)*. https://orion.nmbu.no

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., & Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Molecular Systems Biology*, *7*. https://doi.org/ARTN 535 10.1038/msb.2011.65

Orth, J. D., Fleming, R. M. T., Palsson, B. Ø., & Karp, P. D. (2010). Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. *EcoSal Plus*, *4*(1). https://doi.org/10.1128/ecosalplus.10.2.1

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, *4*(9). https://doi.org/ARTN R54 DOI 10.1186/gb-2003-4-9-r54

Reimers, A. M., & Reimers, A. C. (2016). The steady-state assumption in oscillating and growing systems. *Journal of Theoretical Biology*, *406*, 176-186. https://doi.org/10.1016/j.jtbi.2016.06.031

Schuster, S., Fell, D. A., & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, *18*(3), 326-332. https://doi.org/Doi 10.1038/73786

Theorell, A., & Stelling, J. (2022). Metabolic Networks, Microbial Consortia, and Analogies to Smart Grids. *Proceedings of the IEEE*, 1-16. https://doi.org/10.1109/JPROC.2022.3158396

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, *5*(1), 93-121. https://doi.org/10.1038/nprot.2009.203

Thiele, I., Vo, T. D., Price, N. D., & Palsson, B. O. (2005). Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *Journal of Bacteriology*, *187*(16), 5818-5830. https://doi.org/10.1128/Jb.187.16.5818-5830.2005

Wagner, A. (2012). Metabolic networks and their evolution. *Adv Exp Med Biol*, *751*, 29-52. https://doi.org/10.1007/978-1-4614-3567-9_2

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wedmark, Y. (2022). *Master Thesis - Pathway Analysis*. https://gitlab.com/YlvaKaW/masteroppgave-pathway-analysis

Øyås, O., & Stelling, J. (2020). Scalable metabolic pathway analysis. *bioRxiv*, 2020.2007.2031.230177. https://doi.org/10.1101/2020.07.31.230177

Appendix

A1 Computation time (s) of randomly sampling each MP

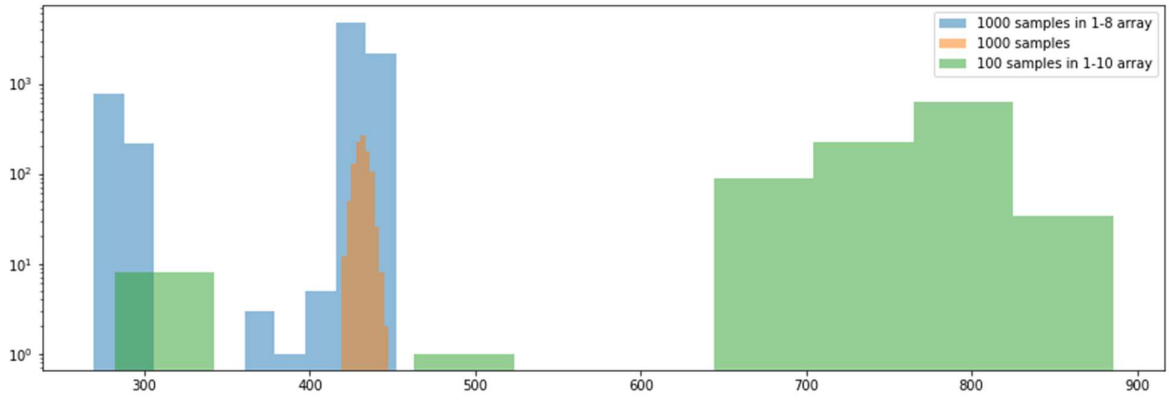

Figure A.1: Distribution of the time (s) it took to randomly sample one MP. The colors represent each job sent to Orion. The blue histogram is distribution of computation times per MP of the 8 parallel jobs where each sampled 1 000 MPs. The orange histogram is distribution the computation times per MP of one job that sampled 1 000 MPs. The green histogram is the distribution of computation times per MP of 10 parallel jobs where each job sampled 1 000 samples.

A2 Total computation time (s) of randomly sampling 100 MPs and 1 000 MPs

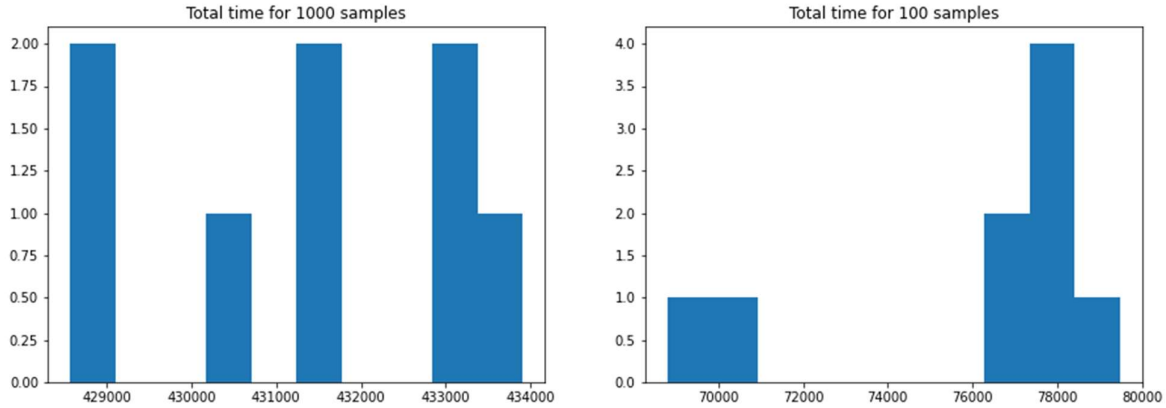

Figure A.2: Histograms of total computation time (s) for random sampling of 1 000 MPs (left) and of 100 MPs (right).

A3 All exchanges of 10 000 randomly sampled MPs of human-microbe model that are consumed by the human model

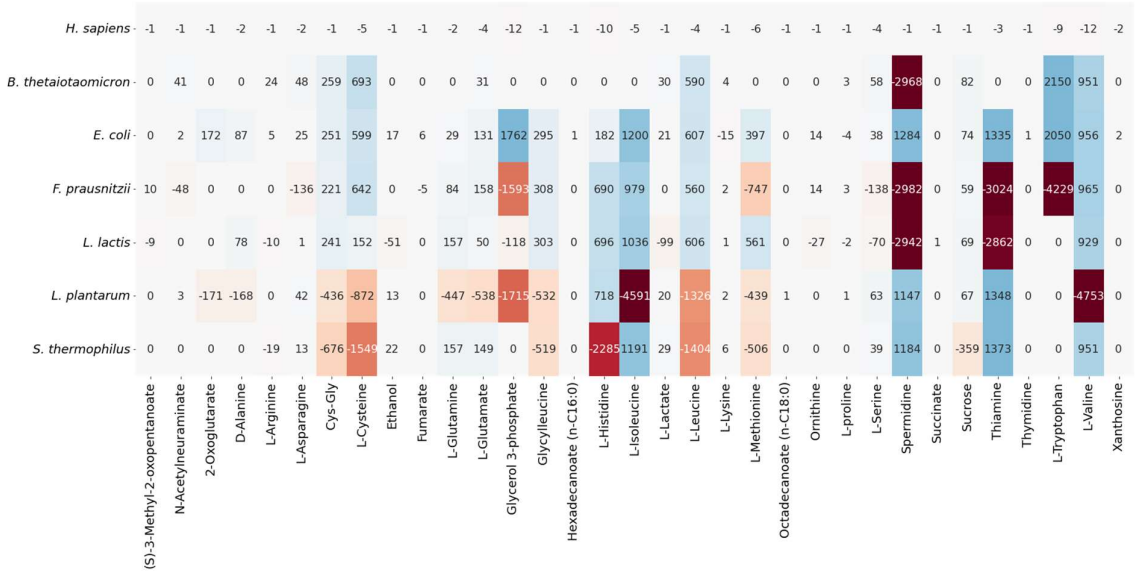

Figure A.3: Heatmap of metabolite exchanges between human model and human gut microbe models. Each row shows one species of the model, and each column shows a metabolite. The number in each square is the number of MPs a metabolite appeared in for a model of given species. Positive numbers (colored in shades of blue) signifies that the metabolite was produced by the species, negative numbers (colored in shades of red) signifies that the metabolite was consumed by the species. Only metabolites that were exchanged between the models and consumed by the human model across 10 000 samples are shown.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway