NTNU – Trondheim
Norwegian University of
Science and Technology

Identification of Antimicrobial Drug Targets from Robustness Properties of Metabolic Networks

Ove Øyås

Chemical Engineering and Biotechnology
Submission date: June 2015
Supervisor: Eivind Almaas, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Declaration

I declare that this work has been performed independently and in accordance with the rules and regulations for examinations at the Norwegian University of Science and Technology (NTNU).

Ove Øyås,
Trondheim, June 11, 2015

ABSTRACT

A reaction universe containing all 13,849 metabolic reactions known to exist was constructed and found to share many topological properties with real-world metabolic networks. Integration of the reaction universe into 43 different microbial genome-scale metabolic reconstructions led to improved viability and robustness. Five metabolic reactions remained essential in more than 70 % of these reconstructions after integration of the reaction universe and these absolutely superessential reactions were identified as potential targets for broad-spectrum antimicrobial drugs. One of the five reactions was involved in peptidoglycan biosynthesis and the remaining four were part of riboflavin metabolism. No reactions were absolutely superessential in all 43 cellular contexts, meaning that no set of reactions that are always essential in any metabolic network is likely to exist.

Ten of the reconstructions into which the reaction universe was integrated were used to generate large ensembles of random viable metabolic networks. The method used for metabolic network randomization was evaluated and it was found that it produced networks with large fractions of blocked reactions. Aside from this, the reaction contents of random viable metabolic networks correlated very strongly with network size. Most importantly, small networks were less randomized than large ones. Even so, the increased size of the reaction universe relative to past studies allowed greater network randomization than what has previously been achieved.

Many reactions that were essential or part of synthetic lethal pairs in random viable metabolic networks were capable of being so in all investigated cellular contexts. Based on this, it was postulated that essentiality and synthetic lethality is often caused by factors that are shared between different organisms and environments.

Superessentiality indices, which indicate how frequently reactions are expected to be essential in metabolic networks in general, were calculated and found to correlate positively between cellular contexts. However, these correlations were only strong between indices obtained from very similar models, indicating that superessentiality is sensitive to cellular context. Also, a great deal of deviation between indices calculated in this study and previously reported ones was observed, primarily due to the increased size of the reaction universe. An average superessentiality index revealed that some reactions were highly superessential in all investigated cellular contexts and the ten reactions with highest average superessentiality indices, all of them involved in purine or histidine metabolism, were identified as potential antimicrobial drug targets.

Synthetic lethality data obtained from random viable metabolic networks was used to construct graph representations of pairwise synthetic lethal interactions between reactions. All of these synthetic lethality networks contained a giant component in which most nodes were found and in all cases this giant component was highly clustered and single-scale and exhibited small-world properties. Indications of assortative network organization were also found.

Finally, an algorithm was developed for identifying alternative metabolic pathways of essential reactions in metabolic networks and applied to all essential reactions in two models of potentially pathogenic bacteria. It was found that more than 500 alternative metabolic pathways existed in the reaction universe for most essential reactions in these models. The remaining essential reactions generally had few alternative pathways, most of which consisted of few reactions. Comparison to superessentiality indices showed that the key determinant for reaction superessentiality was most likely a combination of the number of alternative pathways and the lengths of these pathways.

Abstract (Norwegian)

Et reaksjonsunivers bestående av alle 13 849 kjente biokjemiske reaksjoner ble konstruert. Mange felles topologiske egenskaper mellom dette universet og reelle metabolske nettverk ble funnet. Integrering av reaksjonsuniverset i 43 ulike rekonstruksjoner av mikrobielle metabolske nettverk forbedret disse nettverkenes levedyktighet og robusthet. Fem reaksjoner forble essensielle i mer enn 70 % av disse nettverkene etter integrering av reaksjonsuniverset og disse absolutt superessensielle reaksjonene ble identifisert som potensielle mål for bredspektrede antimikrobielle midler. Én av de fem reaksjonene var involvert i peptidoglykansyntese og de fire andre var del av riboflavinmetabolismen. Ingen reaksjoner var essensielle i alle disse cellulære kontekstene, noe som betyr at det sannsynligvis ikke finnes noe sett av reaksjoner som alltid er essensielle i alle metabolske nettverk.

Ti av nettverkene som reaksjonsuniverset ble integrert i ble brukt til å generere store samlinger av tilfeldige levedyktige metabolske nettverk. Metoden som ble brukt til nettverksrandomisering ble evaluert og det ble funnet at nettverkene den produserte inneholdt store andeler blokkerte reaksjoner. Reaksjonsinnholdet i nettverkene korrelerte forøvrig sterkt med nettverkenes størrelse. Blant annet ble nettverk med få reaksjoner mindre randomisert enn de med mange. Nettverkene ble likevel mer randomisert enn i tidligere studier som følge av at reaksjonsuniverset som ble brukt her var større.

Mange reaksjoner som var essensielle eller del av syntetisk letale par i de tilfeldige metabolske nettverkene var i stand til å være essensielle i alle de cellulære kontekstene som ble undersøkt. Basert på dette ble det postulert at essensialitet og syntetisk letalitet ofte er forårsaket av faktorer som deles mellom ulike organismer og miljøer.

Superessensialitetsindekser, som indikerer hvor ofte reaksjoner ventes å være essensielle i metabolske nettverk generelt, ble beregnet og positiv korrelasjon ble funnet mellom ulike cellulære kontekster. Sterk korrelasjon ble imidlertid kun funnet mellom indekser beregnet fra svært like modeller, noe som indikerer at superessensialitet avhenger av cellulær kontekst. Mye variasjon ble også funnet mellom indeksene som ble beregnet i denne studien og de som tidligere har blitt rapportert, men dette skyldtes primært den økte størrelsen på reaksjonsuniverset. En gjennomsnittlig superessensialitetsindeks viste at noen reaksjoner var svært superessensielle i alle undersøkte cellulære kontekster og de ti reaksjonene med høyest gjennomsnittlig indeks ble identifisert som potensielle mål for antimikrobielle midler. Alle disse reaksjonene var del av purin- eller histidinmetabolismen.

Syntetisk letalitet observert i tilfeldige nettverk ble brukt til å sette opp nettverksrepresentasjoner av syntetisk letale interaksjoner mellom reaksjonspar. I alle disse nettverkene var de fleste nodene samlet i en stor sammenkoblet komponent som inneholdt mange tett koblete klynger av noder, hadde én skala, og utviste «liten verden»-egenskaper. Indikasjoner på assortativ nettverksorganisering ble også funnet.

En algoritme ble utviklet for identifisering av alternative biokjemiske spor for essensielle reaksjoner i metabolske nettverk. Denne algoritmen ble anvendt på alle essensielle reaksjoner i to modeller av potensielt patogene bakterier. Mer enn 500 alternative spor ble funnet i reaksjonsuniverset for de fleste av disse reaksjonene. De øvrige essensielle reaksjonene hadde generelt få og korte alternative spor. Sammenligning med superessensialitetsindekser avdekket at den viktigste determinanten for superessensialitet sannsynligvis var en kombinasjon av antall alternative spor og lengden til disse sporene.

Preface

The work presented in this master's thesis was carried out at the Norwegian University of Science and Technology (NTNU) in the spring of 2015. The thesis marks the end of my time as a student in the Biotechnology specialization of the five-year M.Sc. program in Chemical Engineering and Biotechnology.

My supervisor has been Professor Eivind Almaas at the Department of Biotechnology, to whom I would like to express my sincere gratitude. His dedication and insights have motivated me greatly throughout the time spent working on this thesis and our discussions have invariably been fruitful.

I am also grateful to the Wagner lab at the University of Zürich, in particular Dr. Aditya Barve and Professor Andreas Wagner, for welcoming me so warmly when I visited them in the spring of 2014. Aditya Barve deserves special thanks for his interest in my work and for taking the time to follow up on me both during and after my stay in Zürich.

Peter Wad Sackett and Professor Mikael Rørdam Andersen at the Technical University of Denmark both deserve thanks as well as credit for this thesis. Without attending their courses, I would not have developed the skills and knowledge needed to pursue these topics. They were also kind enough to give me recommendations for jobs that I applied for.

I thank Bjørn Lindi and Vegard Eide at NTNU's Section for Scientific Data Processing for their kind assistance with getting my software up and running on the supercomputer Vilje.

Høiskolens Chemikerforening has been at my side throughout my studies at NTNU and deserves credit for way too much fun to list here.

Finally, my biggest thanks go to my parents, Trine Hjertås Østlyng and Ola Øyås, for always being there for me and letting me follow my interests, and Stine Marie Hoggen for her incredible kindness, patience, and support.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 The emergence of systems biology

In 1966, Francis Crick wrote that "the ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry" [1]. This statement succintly summarizes the reductionist paradigm that has dominated the biological sciences for the better part of the last century. Biological systems have been described and analyzed in terms of their basic components with the fundamental assumption that elucidating the function of these components separately would lead to an understanding of how the systems work as a whole [2–4]. Although this approach has proved very effective and enabled scientists to explain the chemical basis of many biological processes, there is growing awareness of its limitations [5, 6].

Simply put, reductionism has failed to explain how living organisms function because they are staggeringly complex, displaying properties that often cannot be explained or even predicted through the study of their individual parts alone [3, 5]. In this way they are similar to other complex systems that are studied in physics, but biological systems differ from these systems as well. Importantly, whereas the emergent properties of nonbiological complex systems are often explained as consequences of homogeneous parts interacting more or less randomly with no particular purpose, biological systems consist of heterogeneous parts that have evolved to become highly organized in space and time in response to functional requirements [7].

To account for the complexity and unique properties of life, a way of thinking focused on systems properties has emerged in biology over the past few decades [2]. This new paradigm has been termed systems biology [8]. It

is interdisciplinary by nature, combining biology with fields such as mathematics, physics, computer science, and others, and is fundamentally about the integration of parts into a larger whole. As stated by Denis Noble in *The Music of Life: Biology beyond genes* [9]:

> *"Systems biology (...) is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different. (...) It means changing our philosophy, in the full sense of the term."*

The emergence of systems biology has been driven largely by the development of high-throughput experimental technologies that allow simultaneous systems-level characterization of biological components. These technologies have caused the pace of data generation to increase exponentially, leading biology into the information-rich era of genomics, transcriptomics, proteomics, metabolomics, and so on. Today, *omics* data can provide descriptions of virtually all molecular components and interactions that occur within a cell, paving the way for mechanistic formulations of the genotype-phenotype relationships that are at the heart of the life sciences. [2, 10]

Connections between genotype and phenotype are currently close to being unraveled for metabolic functions, much thanks to genome-scale computational models of metabolic networks and their steadily increasing ability to accurately replicate experimental data [2, 4, 11, 12]. Since the first models of this type were published around the turn of the last century [13, 14], they have evolved into powerful research tools with a wide range of applications [15–17]. Further expansion of the scope and predictive capablities of these and other systems-level modeling frameworks promises to help make 21st century biology a truly quantitative, integrative, and predictive science [4, 18, 19].

1.2 Antimicrobial drugs and resistance

Microorganisms, or microbes, are free-living organisms that are usually single-celled. They are very diverse and include all bacteria and *Archaea* as well as many eukaryotes. Some microorganisms are pathogens that infect other organisms, and some of these pathogens cause disease in humans. The discovery of antibiotics revolutionized 20th century medicine, enabling easy treatment of many bacterial infections and nearly eradicating diseases such as tuberculosis, and antibiotics and other antimicrobials remain absolutely cru-

cial today. It is estimated that more than 10,000 metric tons of antimicrobial drugs are manufactured and used annually worldwide. [20]

Despite the successes of antimicrobial drugs, the microbes that are targeted have become progressively more resistant to the drugs that are commonly employed, much due to overuse [21, 22]. Also, research on the development of such drugs has largely stagnated [23] and the combination of these two factors has caused antimicrobial drug resistance to emerge as a serious threat to human health over the past few decades [24]. A major crisis looms in medicine and enormous investments as well as new approaches are needed to get antimicrobial research back on track [23, 25].

Several antimicrobial drugs are antimetabolites that target pathogens through their metabolism. Prominent examples include the sulfonamides, a class of growth factor analogs that act on enzymes involved in the biosynthesis of folate [20, 26], and trimethoprim-sulfamethoxazole, which also interferes with folate metabolism [27]. Resistance to antimetabolites is developed either through *de novo* mutation or via acquisition of genes from other organisms through horizontal gene transfer. [26]. In the latter case, the acquired genes may enable the microbe to produce enzymes that destroy an antimicrobial drug, to express transport proteins that allow it to excrete the drug before it asserts its action, or to produce an alternative metabolic pathway that bypasses the target of the drug [26].

1.3 Thesis objective

In this thesis, the following fundamental question is asked: How difficult is it for microorganisms to replace the reactions that are essential in their metabolic networks? Answers are sought in two complementary ways, both of which are computational and based on exploration of a reaction universe consisting of all metabolic reactions known to exist. The first approach builds upon previous work by Barve, Rodrigues, and Wagner [28] and involves integration of the reaction universe into microbial genome-scale metabolic reconstructions followed by generation of large ensembles of randomized, theoretically viable metabolic networks from these reconstructions. The second aims to identify the reaction sets in the reaction universe that are capable of replacing essential reactions in metabolic networks and thus constitute alternative metabolic pathways.

The focus of both approaches is to identify metabolic reactions that are predicted to be both essential and difficult or impossible to bypass in any metabolic network. This addresses the problem of antimicrobial drug resistance caused by acquisition of enzyme-encoding genes that produce alterna-

tive metabolic pathways. The fewer ways there are to replace a reaction that is targeted by an antimicrobial drug, the less likely it should be for resistance to the drug to develop through this mechanism. It is hoped that this study can aid the identification of broad-spectrum antimicrobial drug targets that are not only essential for the survival of pathogens but also minimize the risk of resistance.

CHAPTER 2

THEORY AND LITERATURE REVIEW

In this chapter, theory and previously published research that is considered relevant for the thesis is presented. The main topics that will be reviewed are the basic concepts of linear programming, the theory and applications of constraint-based reconstruction and analysis (COBRA), robustness properties of metabolic networks, the concept of metabolic genotype space, and some fundamental network theory.

2.1 Linear programming

Linear programming (LP) is a mathematical optimization method that is used to determine the best outcome in a situation where requirements can be formulated as linear relationships. It is fundamentally about distributing a limited number of resources among competing activities in an optimal way and has found widespread use in business, economics, and engineering. Here, the elementary theory of linear programming and some of its varieties is presented, all of it based on Hillier and Lieberman [29].

2.1.1 Defining a linear programming problem

There are four main components in an LP problem:

- The objective function that one wishes to minimize or maximize.
- Coefficients and constants that represent known data.

- Decision variables representing the levels of activities.
- Restrictions that limit the allowed values of the decision variables.

The objective function as well as all functions describing restrictions are required to be linear. The standard way to express the model in matrix form is as follows:

$$\begin{aligned} \text{maximize} \quad & Z = \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ \text{and} \quad & \mathbf{x} \geq 0 \end{aligned}$$

Here, Z is the objective function, $\mathbf{x}$ is the vector of decision variables, $\mathbf{c}$ and $\mathbf{b}$ are vectors of coefficients and constants, and $\mathbf{A}$ is a matrix of coefficients. Note that $\mathbf{c}^{\mathrm{T}}$ signifies the transpose of $\mathbf{c}$. In economical terms, activities are represented by $\mathbf{x}$, resources by $\mathbf{b}$, the costs of activities by $\mathbf{A}$, and the profit of activities by $\mathbf{c}$.

The form of the LP problem shown above is referred to as the *standard form*. It should be added that four legitimate variations of the standard form exist:

- Minimization of the objective function rather than maximization.
- Greater-than-or-equal-to inequality constraints ($\geq$).
- Equality constraints ($=$).
- No nonnegativity constraints for one or more decision variables.

Any situation that can be formulated mathematically using the standard form and any of these variations is an LP problem.

2.1.2 Solutions and solution space

In linear programming, any set of values for the decision variables of a problem is refered to as a solution, regardless of whether it is optimal or even allowed. There are, however, different types of solutions, the most fundamental of which are the feasible and infeasible ones. Feasible solutions are solutions for which all constraints are satisfied, whereas infeasible solutions are solutions for which at least one constraint is violated. The feasible region, also called the solution space, is the collection of all feasible solutions and has as many dimensions as there are decision variables in the problem.

An optimal solution is a feasible solution that has the largest or smallest allowed value, depending on whether the objective function is maximized or minimized. A problem can have zero, one, or an infinite number of optimal solutions. If no optimal solution exists for a problem, it is either because no feasible region exists or because infinitely high or low objective values are allowed due to the feasible region not being closed. If exactly one solution exists, it can be shown that it must always lie at a corner of the feasible region, and if infinitely many solutions exist, they must all lie along the border of the feasible region and at least two of them must lie at a corner. Note that this means that any LP problem that has a feasible region has an optimal solution in at least one of the corners of this region. For this reason, the corner solutions have special significance and are refered to as corner-point feasible (CPF) solutions. Figure 2.1 illustrates the feasible region of an LP problem with two decision variables and constraints.

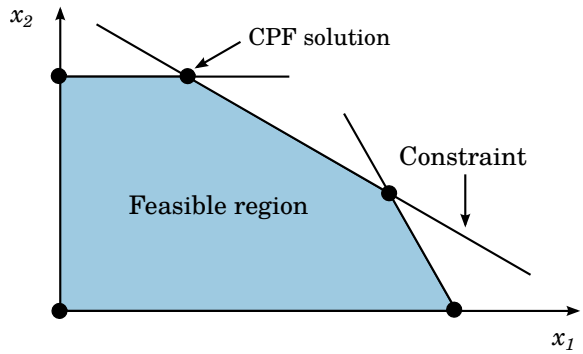


Figure 2.1: A two-dimensional closed feasible region defined by constraints. One or two of the corner-point feasible (CPF) solutions that are marked with dots must be optimal.

2.1.3 Solving linear programming problems

LP problems can be solved very quickly, even for huge numbers of decision variables and constraints. Many algorithms exist for this purpose, but the most fundamental one is the simplex method, which will be presented here. Developed by George Dantzig in 1947, it has proven to be remarkably efficient and is still widely used today. It is based on a very straight-forward idea, namely to move between neighboring CPF solutions until an optimal solution is reached, and can be understood in terms of six key concepts:

1. Only CPF solutions are considered. As previously discussed, if a problem has at least one optimal solution, an optimal CPF solution must exist.

2. The method is iterative and each iteration asks a simple question: Is the currently considered CPF solution optimal? If it is, the procedure stops, if not, it moves to the next CPF solution.

3. Whenever possible, the origin is chosen as the initial solution, as this eliminates the need to find and solve the initial solution.

4. It is computationally easier to evaluate a neighboring CPF solution than another solution. This is because only one restriction needs to be changed from the previous solution, as opposed to all restrictions for nonneighboring solutions.

5. Each CPF solution has two neighbors and will be connected to them by edges along which the value of the objective function strictly increases or decreases. This eliminates the need to solve for the neigboring solutions in order to identify which one has the better value. Instead, the rate of improvement of the objective function along the edges connecting a CPF solution to its neighbors is identified.

6. A CPF solution is optimal if none of its neighboring solutions are better.

The application of the simplex algorithm to an LP problem in standard form is illustrated in Figure 2.2.

2.1.4 Integer and nonlinear programming

Integer programming (IP) is a variation of linear programming in which additional constraints are placed on the decision variables. In an LP problem, all variables must be continuous, whereas in an IP problem they are restricted to be integers. Several classes of integer programming exist, notably binary integer programming (BIP), in which all variables are restricted to be one or zero, and mixed integer programming (MIP), which mixes continuous and integer variables. Solving any kind of IP problem is computationally hard and requires different solution procedures than those used to solve LP problems.

In nonlinear programming (NLP) the LP assumptions of linear objective and constraint functions are not made. This leads to nonlinear mathematical

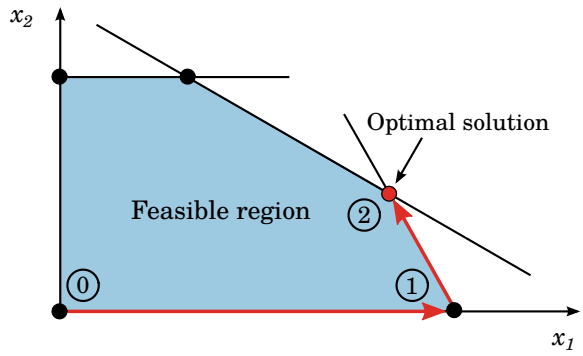


Figure 2.2: Illustration of the simplex algorithm applied to an LP problem in standard form with two decision variables. Nonoptimal CPF solutions are marked by black dots and the optimal CPF solution by a red dot. The iteration at which a CPF solution is evaluated is indicated by a circled number and the red arrows signify movement between neighboring CPF solutions. The procedure starts at the origin and identifies the edge along which the rate of change for the objective function is highest. It moves along this edge to the next CPF solution and checks the rate of change along the edge connecting this edge to its nonvisited neighbor. The rate of change is positive, and so it moves again. The next CPF solution is identified as optimal and the procedure stops.

optimization problems that can generally be expressed as

$$\begin{aligned} \text{maximize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq b_i, \quad i = 1, 2, \ldots, m \\ \text{and} \quad & \mathbf{x} \geq 0 \end{aligned}$$

where $f(\mathbf{x})$ and $g_i(\mathbf{x})$ are functions of the decision variables in $\mathbf{x}$. There are many types of NLP problems, depending on the characteristics of $f(\mathbf{x})$ and $g_i(\mathbf{x})$, and different algorithms are used to solve the different types. Some classes of NLP problems can be solved quite efficiently, while others are challenging to solve even for small problems.

2.2 Constraint-based reconstruction and analysis

Constraint-based reconstruction and analysis (COBRA) encompasses the reconstruction of metabolic networks from genome sequences and available data as well as the assessment of the metabolic capabilities of the resulting models [4, 12, 30]. It is the most well-established framework for metabolic

modeling and the only modeling technique that has been successfully applied to metabolism on the genome scale [19, 30]. This section outlines all the key aspects of COBRA, including the process of reconstructing metabolic networks *in silico*, the contents and basic mathematical properties of genome-scale metabolic reconstructions, and the identification of optimal cellular states.

2.2.1 Genome-scale metabolic reconstructions

Like the rest of biology, the metabolic processes of living cells have historically been studied and described in reductionist terms, with sets of reactions defining biochemical pathways that play discrete roles in cellular function. In reality, however, these pathways are all entangled in a complex network of metabolic reactions that is highly organized in space and time [2, 31]. For any given cell, this *metabolic network*, defined by the totality of the chemical reactions that occur within it, is what underlies its cellular processes and produces its key physiological properties [2, 16]. An illustration of a metabolic network is shown in Figure 2.3.

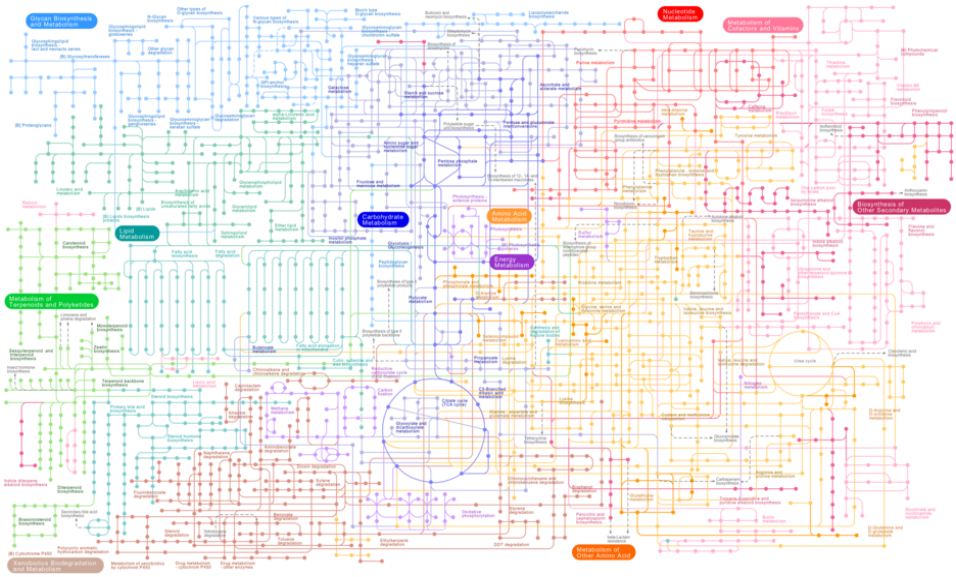


Figure 2.3: An overview of the metabolic network defined by pathways found in the KEGG PATHWAY database [32, 33]. Image obtained via the KEGG API (http://rest.kegg.jp/get/map01100/image).

A metabolic network can be reconstructed *in silico*; that is, available information about the metabolism of an organism can be integrated into a computational model [16, 34, 35]. These models, which are often referred to as genome-scale metabolic reconstructions, encompass all metabolites and reactions that are known to exist in an organism [36].

The metabolic network reconstruction procedure

The procedure for metabolic network recontruction has been reviewed by Feist, Herrgard, and Thiele [16] and a detailed protocol has been published by Thiele and Palsson [35]. In general, the reconstruction procedure consists of four main phases:

1. *Assembling a draft reconstruction*. In the first phase, a preliminary model is created based on the genome sequence of the organism that is to be modeled. The genome must be annotated, meaning that the sequence must be associated with structural and functional biological information, most importantly the presence of enzyme-encoding genes [37]. From genome annotation, unique gene identifiers are obtained, and these in turn lead to the identification of the metabolic enzymes that are thought to be present in the target organism. This procedure, which has been fully automated in recent years [35, 38], leads to a draft reconstruction.

2. *Curation of the draft reconstruction*. In phase two, the draft reconstruction from phase one is inspected and corrected. This manual curation is a laborious and time-consuming process that involves the removal of erroneously added reactions as well as the filling of network gaps through the addition of new reactions. The result should be a high-quality reconstruction of the metabolic network of the target organism [16, 35].

3. *Converting the reconstruction to a computational model*. Phase three involves converting the network reconstruction into a mathematical representation, effectively turning it into a true genome-scale model of metabolism [16]. A crucial part of this step is the definition of a biomass reaction based on experimental data. This allows computation of physiological properties that can be compared to experimental results in the fourth and final step. The ATP requirement for non-growth-associated maintenance should also be estimated through experiments and added to the model as a reaction. At the end of the phase, one should obtain a draft model that can be used to make phenotype predictions.

4. *Model evaluation*. This phase represents the final debugging of the genome-scale metabolic model in which its predictive power is evaluated by comparison to *in vivo* experiments [35]. Examples of physiological properties that can be computed using constraint-based methods and used to validate the model include viability on minimal growth media, growth rates, uptake and secretion rates, and gene essentiality [16]. The results of comparisons between model-predicted phenotypes and experimentally determined ones are used to guide further improvement of the model – the first step in an iterative process of computation, comparison and model improvement that leads to a finalized model. The decision of when to stop model improvement will depend on its desired scope and purpose [35].

Contents of genome-scale metabolic reconstructions

The most basic components of a genome-scale metabolic reconstruction are reactions and the metabolites that participate in them. Several different categories of reactions can be defined. First of all, there are what can be called metabolic or chemical reactions that transform reactants of one kind into products of another kind. Second of all, virtually all genome-scale metabolic models contain at least two distinct compartments, the cytoplasm and the extracellular space, necessitating the inclusion of intercompartmental transport reactions. The function of these reactions is not to convert metabolites into other metabolites, but to move one or more metabolites between two different compartments. Finally, exchange, or boundary, reactions define the uptake and secretion rates of all metabolites that can be exchanged between an organism and its environment. [2, 16, 35]

The biomass reaction accounts for the fractional distribution of all constituents that are known to be necessary for the organism to grow and reproduce. It is a necessity for computing the ability of the metabolic network to support growth [16, 35, 39]. As an example, take the following biomass reaction from a published model of *Aspergillus niger* [40]:

$$0.263\ \text{Protein} + 0.00244\ \text{DNA} + 0.01814\ \text{RNA} + 0.10899\ \text{Lipids} + 0.131\ \text{Pool} + 0.38\ \text{Cell wall} + 61\ \text{ATP} + 61\ \text{H}_2\text{O} \longrightarrow 61\ \text{ADP} + 61\ \text{P}_\text{i} + \text{Biomass}$$

Growth-associated maintenance (GAM), i.e. the energy, in the form of ATP, that is consumed as the cell replicates, is included in the biomass reaction. Cellular systems also expend energy for maintenance functions that are not associated with growth, and more recent models therefore include a non-growth-associated maintenance (NGAM) reaction as well. The general form of an NGAM reaction is simply the consumption of one molecule of ATP,

which is scaled to match experimental data via adjustment of the reaction rate [35]:

$$\mathrm{ATP} + \mathrm{H_2O} \longrightarrow \mathrm{ADP} + \mathrm{P_i} + \mathrm{H^+}$$

In addition to reactions and metabolites, genome-scale metabolic reconstructions usually incorporate the genes of the organism and information about how these genes relate to its proteome and reactome [41, 42]. Such gene-protein-reaction associations (GPRs) are implemented as boolean expressions that indicate which gene products are needed to form the enzymes that are necessary for a reaction to occur [16, 35]. In recent years, models have emerged that take this one step further by integrating descriptions of macromolecular synthesis [43–45]. Finally, it should be mentioned that high-throughput data such as gene expression levels can be integrated in different ways, expanding the scope of models and potentially improving their predictive power [46–48].

Examples of high-quality genome-scale metabolic reconstructions

The metabolic networks of a large number of organisms have been reconstructed and the number of available reconstructions is growing at a pace similar to that of genome sequencing [35]. These reconstructions encompass all three domains of cellular life, ranging in complexity from bacterial models containing only a few hundred reactions, metabolites, and genes, to eukaryotes with several thousand distributed over multiple intracellular compartments. Some representative examples of growth-predictive genome-scale metabolic reconstructions that have been validated against experimental data are presented in Table 2.1.

Table 2.1: Examples of high-quality genome-scale metabolic reconstructions, ordered by number of genes accounted for in the model. The names of the organism and model are listed along with the number of genes, metabolites, reactions, and compartments. Year of publication is also included along with a reference.

Organism name	Model name	Genes	Met.	React.	Comp.	Pub. year
Mycoplasma genitalium	*i*PS189	189	274	262	2	2009 [49]
Helicobacter pylori	*i*IT341	341	485	476	2	2005 [50]
Methanosarcina barkeri	*i*AF692	692	558	619	2	2006 [51]
Yersinia pestis	*i*AN818m	818	825	1,020	2	2009 [52]
Aspergillus niger	*i*MA871	871	1,045	1,190	3	2008 [40]
Saccharomyces cerevisiae	*i*MM904	904	713	1,412	8	2010 [53]
Escherichia coli	*i*AF1260	1,260	1,039	2,077	3	2007 [41]
Arabidopsis thaliana	AraGEM	1,419	1,748	1,567	6	2010 [54]
Homo sapiens	Recon 1	1,496	2,766	3,311	8	2007 [55]

Applications of genome-scale metabolic reconstructions

The application areas of genome-scale metabolic models can be broadly assigned to five major categories [16, 17]:

1. *Contextualization of high-throughput data*. Models allow integrated evaluation and linking of omics data sets by placing them in a functional and structured context. For example, gene microarray data can be overlaid on a model to determine condition-dependent phenotypes.

2. *Guidance of metabolic engineering*. Model-guided engineering strategies can be used to optimize production of cellular compounds or improve other desired phenotypes.

3. *Direction of hypothesis-driven discovery*. The discovery of new cellular properties such as novel genes and enzymes can be facilitated by models.

4. *Analysis of multi-species relationships*. Different models can be combined to compare species or predict interactions between cells.

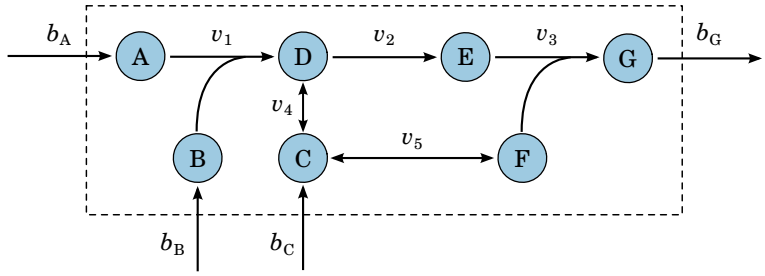
5. *Network property discovery*. Models can be used to study topological network properties such as metabolite connectivity.

2.2.2 The stoichiometric matrix

As previously stated, a genome-scale metabolic reconstruction is fundamentally a large system of metabolic reactions, the basic components of which are the reactions themselves and the metabolites they interconvert. The standard way to represent this system is the stoichiometric matrix, often simply called the S-matrix, in which every row corresponds to a metabolite and every column corresponds to a reaction. The elements of the S-matrix are the stoichiometric coefficients of metabolites in reactions. In each reaction, reactants and products have negative and positive coefficients, respectively, and the coefficients of metabolites that do not participate are zero. Figure 2.4 shows an example of a simple metabolic system and its stoichiometric matrix.

In general, for a system of m metabolites and n reactions, $\mathbf{S}$ is an $m \times n$ matrix:

$$\mathbf{S} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{pmatrix} \tag{2.1}$$



	v_1	v_2	v_3	v_4	v_5	b_A	b_B	b_C	b_G
A	-1	0	0	0	0	1	0	0	0
B	-1	0	0	0	0	0	1	0	0
C	0	0	0	-1	-1	0	0	1	0
D	1	-1	0	1	0	0	0	0	0
E	0	1	-1	0	0	0	0	0	0
F	0	0	-1	0	1	0	0	0	0
G	0	0	1	0	0	0	0	0	-1

Figure 2.4: A simple metabolic system and its stoichiometric matrix. There are seven metabolites and nine reactions in the system. Four of the reactions are boundary reactions. The stoichiometric matrix is divided into two parts, the one on the left describing internal reactions and the one on the right describing boundary reactions. Figure adapted from Lewis, Nagarajan, and Palsson [12].

where $c_{i,j}$ is the stoichiometric coefficient of metabolite i in reaction j. In any realistic large-scale model of metabolism, $n > m$, meaning that there are more reactions than metabolites [39]. Each of the m metabolites in the system has a concentration, denoted x_i for metabolite i. Together these concentrations define the concentration vector:

$$\mathbf{x} = (x_1, x_2, ..., x_m) \tag{2.2}$$

Also, each of the n reactions in the system is associated with a flux, denoted v_j for reaction j. The flux vector is

$$\mathbf{v} = (v_1, v_2, ..., v_n) \tag{2.3}$$

The vectors $\mathbf{x}$ and $\mathbf{v}$ and the matrix $\mathbf{S}$ are related through the dynamic mass balance of the system:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{S}\mathbf{v} \tag{2.4}$$

2.2.3 Identifying optimal cellular states

Cellular functions are limited by constraints that can be classified into four main categories [56]:

- *Physico-chemical constraints*. These are inviolable constraints on cell function that are imposed by physical and chemical laws. Examples include conservation of mass and energy, intracellular diffusion rates, osmotic pressure, the thermodynamics of chemical reactions, and more.
- *Biotopological constraints*. These constraints relate to the physical location of intracellular components. For example, the cell has a limited volume and its components need to be in the same physical location in order to be able to interact.
- *Environmental constraints*. Time-dependent external factors such as nutrient availability, pH, temperature, osmolarity, and the availability of electron acceptors place constraints on cellular functions.
- *Regulatory constraints*. These constraints are also time-dependent and may be implemented in many different ways, examples being cellular regulation of transcription, translation, and enzyme activites.

The constraints summarized above are represented mathematically as balances and bounds [39]. Balances are equalities that represent the conservation of quantitites such as mass, energy, or redox potential, and bounds are inequalities that define the allowable ranges of variables such as fluxes and concentrations. These balances and bounds restrict the attainable metabolic flux distributions of a metabolic system to a limited solution space. This space has the same number of dimensions as there are fluxes in the system and each point within it corresponds to a flux distribution. It is usually convex, meaning that any two points within the space can be connected by a line segment that is completely contained in the space [56]. The phenotype of a metabolic system is represented by its flux distribution, and the solution space thus acts as a summary of phenotypic potential [57].

Linear programming (see Section 2.1) and related mathematical optimization techniques allow the identification of optimal states within the allowable solution space. An optimal state is a phenotype that is better than any other based on some assumed cellular objective [39]. Biologically, this objective can be interpreted as the evolutionary goal of maximizing fitness, and mathematically it takes the form of maximizing or minimizing an objective function [56]. Figure 2.5 illustrates a soluton space defined by constraints and the identification of an optimal solution within this space.

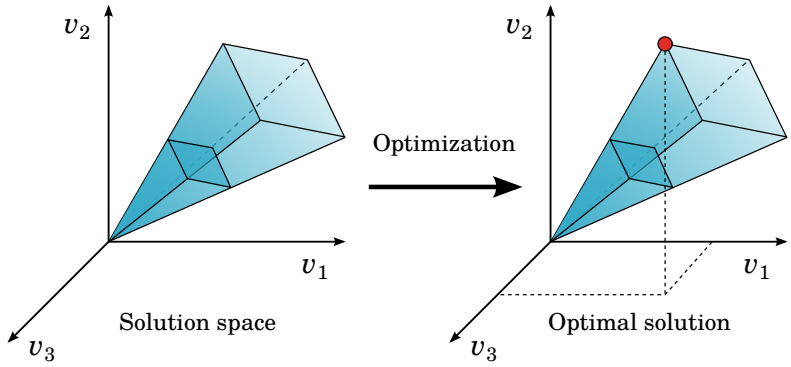


Figure 2.5: Optimization for an assumed cellular objective allows identification of an optimal flux distribution within a solution space defined by constraints. In this example, the metabolic system consists of three reactions and the solution space can therefore be visualized in three dimensions. Figure adapted from Orth, Thiele, and Palsson [39].

The most widely used approach for identifying optimal states in metabolic networks is flux balance analysis (FBA). This method requires that all reactions are elementally balanced and relies on the assumption of steady-state, meaning that the fluxes in the system do not change with time. At steady-state, the dynamic mass balance given in Equation (2.4) becomes

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{S}\mathbf{v} = 0 \tag{2.5}$$

This defines a system of linear equations with as many equations as there are metabolites. The variables are the reaction fluxes. As stated before, there are more reactions than there are metabolites in any realistic genome-scale metabolic reconstruction, so the system is underdetermined: there are more variables than equations and no unique solutions exist. [39]

The second fundamental assumption of FBA is that a metabolic system seeks to distribute its fluxes within the allowable solution space in a way that optimizes for a cellular objective that can be expressed as a linear combination of fluxes. Finding the optimal flux distribution of the system then becomes a linear programming problem that can be expressed in the following way:

$$\begin{aligned} \text{maximize} \quad & Z = \mathbf{c}^{\mathrm{T}}\mathbf{v} \\ \text{subject to} \quad & \mathbf{S}\mathbf{v} = 0 \\ \text{and} \quad & l_i \leq v_i \leq u_i, \quad i = 1, \ldots, n \end{aligned}$$

Here, Z is the objective function, $\mathbf{c}$ is a vector of coefficients indicating how much each reaction contributes to the objective function, $\mathbf{v}$ is the flux vector, $\mathbf{S}$ is the stoichiometric matrix, and l_i and u_i are the lower and upper bounds of the flux v_i, respectively. There are n fluxes in the system, equal to the number of reactions. The flux bounds and the steady-state mass balance define the constraints. The problem can be solved very efficiently, as described in Section 2.1, yielding a flux distribution that optimizes the system for its objective. The typical example of an objective function is to maximize growth rate, i.e. to achieve the highest possible flux through the biomass reaction, but any other sum of weighted fluxes can be maximized or minimized as well. The identification of biologically meaningful objective functions and optimality principles is a theme of ongoing research [58, 59].

2.3 Robustness of metabolic networks

Metabolic networks are inherently robust, meaning that they show a high degree of tolerance to genetic and environmental perturbations in the form of deletions of network components such as genes, reactions, or metabolites [60, 61]. This robustness is closely intertwined with the complexity of metabolic networks and is a common denominator of many networks found both in nature and elsewhere [62, 63]. The investigation of metabolic network robustness in response to perturbations is often associated with the concepts of essentiality and synthetic lethality, which refer to the ability of an organism to maintain its viability in response to single or multiple deletions of genes or reactions [49].

2.3.1 Essentiality and synthetic lethality

Essential genes or reactions are those whose individual deletion are lethal, meaning that the metabolic network is left unable to produce the precursors that are necessary for biomass production [39, 49]. Synthetic lethals are sets of multiple genes or reactions whose deletions are lethal, but where the deletion of any single member of the set is not lethal in itself [64]. Synthetic lethality can arise in multiple ways. For example, two enzymes can be isozymes, meaning that they are interchangeable with respect to catalyzing an essential reaction, or they can catalyze reactions that occur in different pathways that perform the same essential network function [49].

2.3.2 Superessentiality

Essentiality and synthetic lethality are environment- and organism-specific properties. A reaction that is essential in one organism in one environment will therefore not necessarily be essential under different conditions in another metabolic network. However, reactions can be more than just essential – they can be essential in all, some, or no metabolic networks with a given phenotype – and to account for this fact the concept of superessentiality has been introduced [28, 65, 66].

As shown in Figure 2.6, a hierarchy of superessential reactions can be defined. At the top of this hierarchy are the absolutely superessential reactions that are essential in all metabolic networks. Barve, Rodrigues, and Wagner [28] identified 124 such reactions in their analysis of superessentiality. In the middle of the hierarchy, superessential reactions are found, the reactions that are essential in some, but not all metabolic networks. The nonessential reactions that are never essential in any metabolic network are found at the bottom of the hierarchy.

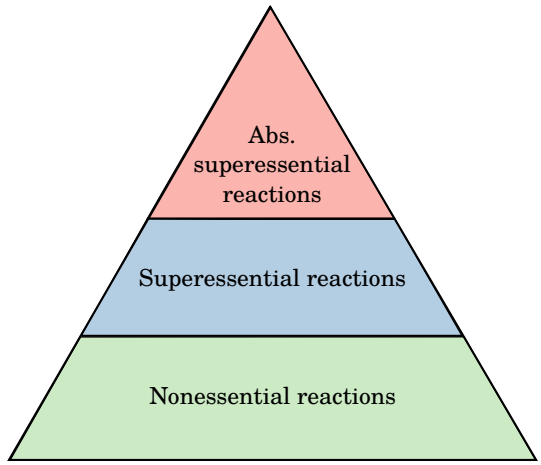


Figure 2.6: The hierarchy of superessentiality in metabolic networks. Absolutely superessential reactions that are essential in all metabolic networks are at the top of the hierarchy, followed by superessential reactions that are essential in some, but not all networks. At the bottom are reactions that are never essential in any metabolic network. Figure adapted from Barve, Rodrigues, and Wagner [28].

The superessentiality of a reaction can be quantified by a superessentiality index, I_{SE} [28]. This index must be a number between one and zero, where a value of one means that the reaction is absolutely superessential and a value of zero means that it is never essential in any network. Intermediate

values correspond to superessential reactions. Barve, Rodrigues, and Wagner [28] reported superessentiality indices for more than 1,400 reactions, as illustrated in Figure 2.7. They also showed that the superessentiality index of a reaction is not very sensitive to the environment or organism-specific biomass requirements.

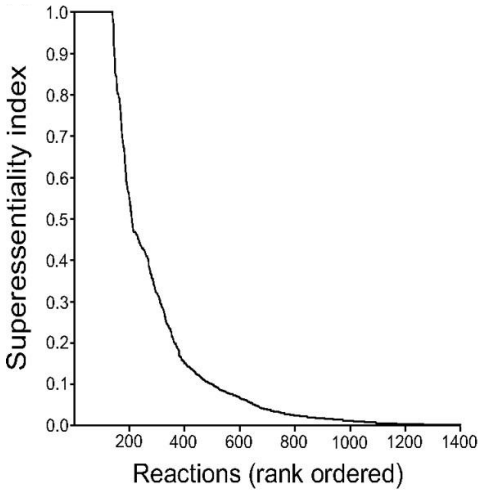


Figure 2.7: Rank plot of the superessentiality indices, I_{SE}, of more than 1,400 reactions. The plateau on the left corresponds to absolutely superessential reactions with $I_{\mathrm{SE}} = 1$. Figure obtained from Barve, Rodrigues, and Wagner [28].

2.4 Metabolic genotype space

As defined by Barve, Rodrigues, and Wagner [28], the reaction universe, or universal metabolic network, is the set of all reactions that are known to occur in one or more metabolic systems. Also, the metabolic genotype of an organism is the set of all enzyme-encoding genes present in its genome [14], or, equivalently, the set of all reactions that are catalyzed by enzymes encoded by these genes [67]. The reaction universe defines a vast set of such metabolic genotypes that has been called metabolic genotype space [68, 69].

Metabolic genotype space summarizes the current state of knowledge of metabolism and it has been shown that investigation of its metabolic properties can lead to the elucidation of general properties of metabolic systems [28, 65–68, 70]. For example, it has been found that genotypes that share the same phenotype form large genotype networks in which two genotypes are connected if they differ by the presence or absence of a single enzyme-

encoding gene. These networks extend throughout genotype space, implying that organisms can traverse a very wide range of different genotypes as they evolve without ever changing their phenotype [65, 69].

A procedure for sampling random genotypes with a given phenotype from metabolic genotype space has been developed. It is a Markov Chain Monte Carlo (MCMC) method [71] that involves long random walks through genotype networks, making use of their highly connected nature [65, 67]. The sampled networks, which have been called random viable metabolic networks, share the same phenotype and contain the same number of metabolic reactions, yet are otherwise randomly independent. Such networks have been used to investigate superessentiality [28] as well as other facets of the evolutionary plasticity and robustness of metabolic networks [65, 67, 68, 70].

2.5 Network theory

Unless otherwise stated, the theory presented here is based on Newman [72]. The basics of network representation as well as some important network measures are reviewed.

2.5.1 Network representation

In its most elementary form, a network, also commonly called a graph, is a collection of nodes that are connected by edges. These edges may be directed, meaning that they point from a source node to a target node but not the other way around, or they may be undirected. Nodes and edges may also be associated with additional information, a common example being edge weights that indicate the strengths of connections. An example of a simple network is shown in Figure 2.8.

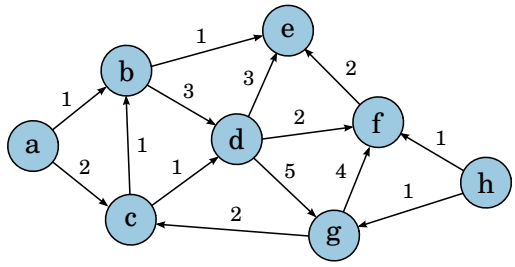


Figure 2.8: An example of a simple network. The nodes in the network are connected by directed edges with integer weights.

2.5.2 Network measures

The network measures presented here can be broadly categorized into those related to neighborhood and those related to paths. The focus is on undirected networks. In all cases, n denotes the number of nodes in a network and m the number of edges.

Measures related to neighborhood

In undirected networks, the neighborhood of a node is the set of nodes to which it is connected by an edge. The size of the neighborhood of a node i is its degree, k_i. One of the simplest characteristics of a network is the average node degree:

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n} \tag{2.6}$$

Related to the average node degree is the density, ρ, of a network:

$$\rho = \frac{\langle k \rangle}{n(n-1)} \tag{2.7}$$

When $\rho = 1$, all possible links between all nodes are present and the network is said to be complete.

The neighborhood connectivity of a node i, $k_{nn,i}$, can be defined as

$$k_{nn,i} = \frac{1}{k_i} \sum_{j=1}^{n} k_j a_{ij} \tag{2.8}$$

where k_i is the degree of node i, k_j is the degree of node j, and a_{ij} is the number of edges directly connecting nodes i and j. Neighborhood connectivity measures the affinity with which a node connects to other nodes of high or low degree. In other words, it expresses connectivity correlations between nodes.

The clustering coefficient C_i of a node i is defined as

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \tag{2.9}$$

where e_i is the number of edges between the k_i neighbors of i. It can be interpreted as the fraction of possible connections between neighbors that are actually present. The global clustering coefficient of a network is the average of the clustering coefficients for all nodes:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{2.10}$$

Measures related to paths

A path in a network can be defined as a sequence of nodes where every consecutive pair of nodes in the sequence is connected by an edge. Put more simply, a path is a route through a network that runs from node to node along the edges of the network. There may be many paths connecting two given nodes. The length of a path is the number of edges forming it and the shortest path between two nodes is, naturally, the one with the smallest length. The shortest path length between two nodes i and j is denoted $L(i,j)$.

Shortest path lengths are only defined between nodes in the same connected component. A connected component is a subset of the nodes of a network in which at least one path exists between all pairs of nodes. The number of connected components in a network is one of the simplest indicators of network connectivity, a low number suggesting high connectivity and vice versa.

The eccentricity, E_i, of a node i is the length of the longest shortest path between this node and any other node j:

$$E_i = \max_{j=1}^{n} L(i,j) \tag{2.11}$$

The diameter of a network, D, is the maximum eccentricity of any node in the network:

$$D = \max_{i=1}^{n} E_i \tag{2.12}$$

and the network radius, R, is the minimum eccentricity of any node in the network:

$$R = \min_{i=1}^{n} E_i \tag{2.13}$$

The characteristic path length, L, of a network is the mean of all the $n(n-1)$ shortest path lengths between the n nodes in the network:

$$L = \frac{1}{n(n-1)} \sum_{j=1}^{n} L(i,j) \tag{2.14}$$

A concept closely related to the distance measures described above is the centrality of nodes in a network. A central node is important in the network, for example by being part of a large fraction of shortest paths. The centrality of nodes in a network may be expressed through the network centralization, C_D [73, 74]:

$$C_D = \frac{1}{n-2} \left(\frac{\max_{i=1}^{n} k_i}{n-1} - \rho \right) \tag{2.15}$$

where ρ is network density.

Distributions

The distribution of node degrees, $P(k)$, is one of the most frequently highlighted properties of networks. It gives the fraction of nodes with node degree k for all k in a network. In order to obtain smoother distributions, the cumulative node degree distribution, $P(K \geq k)$, is often used, indicating the fraction of nodes with degrees greater or equal to k [75].

The average neighborhood connectivity distribution of a network shows how the average neighborhood connectivity of nodes changes with node degree. It is an indicator of whether a network is assortative or disassortative. In an assortative network, high-degree nodes will tend to connect to other high-degree nodes and low-degree nodes will tend to connect to other low-degree nodes. There is positive correlation between the degrees of connected nodes and the average neighborhood connectivity is an increasing function of the node degree. In a disassortative network, the opposite is the case. High-degree nodes tend to connect to low-degree ones and the average neighborhood connectivity distribution is a decreasing function of the node degree.

Finally, it is often useful to consider the average clustering coefficient distribution, which shows how the average clustering coefficient of nodes varies with node degree.

2.5.3 Properties of real networks

Many general characteristics of large and complex networks observed in the real world have been unraveled over the past couple of decades. Systems as different as the World Wide Web, scientific collaborations, food webs, and metabolism have been represented and analyzed as networks and found to have many common features [76]. For example, it has been found that many are highly clustered, yet have small characteristic path lengths, and networks that exhibit these characteristics have been called small-world networks [77]. Some, but not all, small-world networks are scale-free as well, meaning that they have node degree distributions that closely follow power laws [78, 79]. The vast majority of nodes in these networks have low degrees, but a few are very highly connected hubs. Power-law behavior may also be found in the average clustering coefficient distributions of scale-free networks, and this has been associated with hierarchical network organization [80]. In a hierarchical network, the patterns of interactions that occur on the smallest scale between single nodes are also observed on larger scales between clustered groups of nodes [3].

CHAPTER 3

SOFTWARE AND METHODS

This chapter describes how the results presented in this thesis were obtained. All software that was used is described and procedures are provided for all applied methods.

3.1 Software

Here, the computer programs that were used are presented. Every performed task was accomplished using the software tools decribed in this section.

3.1.1 Python and COBRApy

Virtually all data was generated using computer programs written in the Python programming language [81]. Most programs made use of code from COBRApy, a Python module that provides basic COBRA methods [82], in combination with self-produced Python code. COBRApy is open-source software and part of the openCOBRA project, a community effort for promoting constraint-based research [83]. The COBRA Toolbox for MATLAB, one of the most commonly used software packages for constraint-based metabolic modeling, is also part of this project [30]. COBRApy is documented well online [84] and an example of a Python script using COBRApy is included in Appendix A.

3.1.2 LibSBML

All models were stored in the Systems Biology Markup Language (SBML) file format. SBML is a standardized XML-based [85] format designed for com-

puter representation of models of biological processes [86, 87]. The format is free in itself and working with it is facilitated by free and open-source software such as the libSBML programming library that was used in this study [88]. LibSBML is an optional dependency of COBRApy that, when installed, allows SBML files to be read and written easily using built-in COBRApy functions.

3.1.3 Gurobi

Optimization problems were solved using the Gurobi Optimizer [89], a commercial mathematical programming solver that is free to use with an academic license. COBRApy includes an interface to Gurobi through the Python module GurobiPy.

3.1.4 MATLAB

MATLAB (version R2013a) with the Statistics and Machine Learning Toolbox was used for statistical analysis [90].

3.1.5 Graph-tool, Cytoscape and NetworkAnalyzer

Graphs were created using the Python module graph-tool [91]. Visualization and analysis of graphs was performed using Cytoscape [92] with the NetworkAnalyzer plugin [93].

3.2 Parallel computing

Programs were generally executed in parallel using implementations of the Message Passing Interface (MPI) standard. Most programs were run on Vilje, a supercomputer at NTNU. Vilje has 1,404 nodes with a total of 22,464 cores and its theoretical peak performance is 467 teraflops/s [94].

3.3 Flux balance analysis

Flux balance analysis (FBA) was performed using Python and COBRApy (see Section 3.1.1) with Gurobi as optimizer (see Section 3.1.3). It was used to predict growth rates, both wild-type and after deletion of single reactions or reaction pairs, and to identify blocked reactions.

3.3.1 Growth rates

Growth rate predictions were made using FBA with maximization of the biomass reaction flux as objective.

3.3.2 Essential reactions and synthetic lethal reaction pairs

Essential reactions and synthetic lethal reaction pairs were identified through single and double reaction deletion analysis. A reaction was defined as essential if its deletion caused predicted growth rate to drop below a specified threshold value (10^{-3}) and two nonessential reactions were defined as a synthetic lethal pair if the simultaneous deletion of both reactions caused predicted growth rate to drop below the same threshold.

Single reaction deletion analysis

Single reaction deletion analysis was performed according to the following procedure:

1. The model was imported into COBRApy.

2. The predicted wild-type growth rate of the model was calculated and reactions with zero flux in the optimal solution were identified. Reactions that can have zero flux in an optimal solution cannot be essential and these reactions were therefore assigned the wild-type growth rate as their single deletion mutant growth rate and excluded from further analysis.

3. The remaining reactions were deleted from the model one by one and classified as essential if the predicted growth rate of the single deletion mutant was lower than the specified threshold value. Also, after each growth rate prediction, reactions with zero flux in the optimal solution were identified and stored for potential later use in double deletion analysis.

Double reaction deletion analysis

When double reaction deletion analysis was done on a network, it was always preceded by single reaction deletion analysis as described above. The steps below build on that procedure. For all pairs of reactions in the model, the following was done:

1. If the deletion of one or both reactions in a pair was known to be lethal from single deletion analysis, the deletion of the pair was defined as lethal as well, but the pair was not classified as a synthetic lethal.

2. If both reactions in a pair had zero flux in the optimal solution obtained when calculating the wild-type growth rate, the double deletion mutant growth rate was set equal to the wild-type growth rate and the pair was not classified as a synthetic lethal.

3. If one of the reactions in the pair was not found to be essential, but had nonzero flux in the initially calculated optimal solution, and the other reaction had zero flux in the single deletion solution of the first reaction, the double deletion mutant growth rate was set equal to the single deletion mutant growth rate of the first reaction.

4. All pairs of reactions for which none of the criteria above were found to be true were deleted from the model one pair at a time. A reaction pair was classified as a synthetic lethal if the predicted growth rate of the double deletion mutant was below the specified threshold.

3.3.3 Blocked reactions

Reactions that could not carry a nonzero flux in a model were considered blocked. They were identified as follows:

1. The model was imported into COBRApy.

2. The model was optimized and all reactions with nonzero flux in the resulting solution were identified. These reactions could not be blocked and were therefore excluded from further analysis.

3. The fluxes of all remaining reactions were maximized and minimized. Reactions whose maximum and minimum fluxes were both zero were classified as blocked.

3.4 Model construction

This section outlines the process of constructing the models that were used in this study.

3.4.1 Constructing the reaction universe

The reaction universe was constructed using MNXref [95], a recent effort to reconcile biochemical data from different databases and models within a single namespace. Flat files containing all reaction, metabolite, and compartment data in the MNXref namespace were downloaded from MetaNetX.org, a MNXref-based website for "accessing, analyzing and manipulating metabolic networks" [96]. The files were subsequently parsed as follows:

1. Identifiers and names of all metabolites were extracted.

2. Identifiers, names, and equations of all elementally balanced reactions in the data set were extracted. The reactions were then validated by examining their equations. A reaction was considered valid if its equation was not empty or ill-defined, did not contain any metabolites that were not extracted in the previous step, did not contain the same metabolite on both sides of the reaction equation (these were considered to be transport reactions), and did not contain the metabolite identified as biomass.

3. All valid reactions and the metabolites that participated in them were added to a COBRApy model.

4. The upper and lower flux bounds of all reactions were set to arbitrary large values (10^6 and -10^6, respectively). All reactions were considered reversible, in part due to a lack of reversibility data. Also, this approach limited the metabolic capabilities of the reaction universe as little as possible.

5. The model containing the reaction universe was saved in SBML format.

3.4.2 Integration of the reaction universe

The reaction universe was integrated into a selection of existing genome-scale metabolic reconstructions. This was done in two steps as decribed in the sections below.

Preparing models for merging with the reaction universe

First, all growth-predictive genome-scale metabolic models available through MetaNetX.org [96] were downloaded as flat files as well as in SBML format and prepared for merging with the reaction universe. The model preparation procedure consisted of the following steps:

1. The model and the reaction universe were imported into COBRApy. The flat text file was opened as well.

2. All metabolites in the model that did not participate in any reactions were removed.

3. If the model contained multiple biomass reactions, all but one of these were removed.

4. Names of all compartments in the model were checked and corrected if not consistent with the compartments defined in the reaction universe.

5. Boundary reactions, transport reactions, the biomass reaction, and reactions with fixed flux values were identified and defined as framework reactions.

6. Duplicate reactions in the model were merged into a single reaction.

7. The identifiers of reactions in the model were converted to the MNXref namespace using information from the flat text file. No changes were made to the identifiers of framework reactions nor to the identifiers of reactions for which no conversion into MNXref identifiers was available.

8. Flux bounds were adjusted. All nonzero, nonfixed flux bounds of internal reactions in the model were set to a fixed, large value. Excretion rates (the upper flux bounds of bundary reactions) were also set to the same value. The predicted growth rate of the model was then normalized by setting all nonzero uptake rates (the lower flux bounds of bundary reactions) equal to each other and adjusting them until a specified flux through the biomass reaction was achieved. This normalization was performed in order to make it easier to quickly compare the effect on growth rate of changes made in different models.

9. The identifier of the biomass reaction was changed in order to get the same identifier in all models.

10. The prepared model was saved in SBML format.

Merging prepared models with the reaction universe

After preparation, each model was merged with the reaction universe in a straight-forward fashion:

1. The model and the reaction universe were imported into COBRApy.

2. A copy of each reaction in the universe model was added to all intracellular compartments in the prepared model if not already present.

3. The merged model was saved in SBML format.

3.5 Randomization of metabolic networks

Random viable metabolic networks were generated using a Markov Chain Monte Carlo (MCMC) method developed by the Andreas Wagner Laboratory at the University of Zürich [28, 65–68, 70]. The theory behind random viable metabolic networks is presented in Section 2.4. In this study, the previously applied procedure was slightly modified in order to randomize not only one but all intracellular compartments of multicompartment models. The following routine was used:

1. Two models were imported into COBRApy: an existing viable model to be randomized and a version of the same model merged with the reaction universe (see Section 3.4.2). The framework reactions in the models – boundary reactions, transport reactions, the biomass reaction, and reactions with fixed flux values – were also imported along with the identifiers of blocked and essential reactions in the merged model

2. Reactions were classified by compartment. By definition, all nonframework reactions only contained metabolites from a single compartment and each reaction was assigned the compartment of its metabolites.

3. The predicted wild-type growth rate of the prepared model was determined.

4. Metabolic reactions, i.e. nonframework reactions, were swapped between the model being randomized and the merged model. Each swap consisted of the removal of a random reaction from the model being randomized followed by the addition of a random reaction from the merged model to the compartment from which a reaction was removed. Only reactions whose deletion did not cause the predicted growth rate to drop below a threshold of one percent of the predicted wild-type growth rate were allowed to be removed and only reactions not already present in the model being randomized were allowed to be added. Reactions that were known to be blocked in the merged model were never allowed to be added to a model, only removed, as it was known *a priori*

that these reactions could not be active in any random network that was generated. Similarily, removal of reactions that were known to be essential in the merged model was never attempted, as it was known that these reactions had to be present in all random networks in order for them to be viable.

5. Random viable metabolic networks were obtained by repeating the previous step. Before sampling the first random network, 50,000 swaps were performed. This sampled network was then randomized further, with new random networks being sampled every 10,000 swaps.

The number of swaps to perform before sampling the first randomized network and between sampling of two randomized networks was determined through testing as described in Appendix B.

3.6 Calculation of indices

The superessentiality index of a reaction, I_{SE}, was calculated by dividing the number of random viable metabolic networks in which the reaction was found to be essential by the total number of networks sampled. The index for synthetic lethal reaction pairs, I_{SSL}, was similarily calculated by dividing the number of random viable metabolic networks in which a reaction was part of a synthetic lethal pair by the total number of networks sampled.

The average superessentiality index of a reaction, $\langle I_{\mathrm{SE}} \rangle$, was calculated from the superessentiality indices that were obtained for that reaction in individual cellular contexts:

$$\langle I_{\mathrm{SE}} \rangle = \frac{1}{n} \sum_{i=1}^{n} I_{\mathrm{SE}}^{i} \tag{3.1}$$

Here, n is the number of different models and I_{SE}^{i} is the superessentiality index for model i.

3.7 Graph-based analyses

Much of the performed work involved analyses of graphs. This included topological analysis of the reaction universe and investigation of synthetic lethality networks. The software tools summarized in Section 3.1.5 were used in all cases.

3.7.1 Graph representation of the reaction universe

The reaction universe was represented as a graph in which each node corresponded to a metabolite and two nodes were connected by an edge if the metabolites corresponding to the nodes were ever a reactant-product pair in a reaction. Since all reactions in the reaction universe were allowed to proceed in both directions, the edges in the graph were considered to be undirected.

3.7.2 Randomizing synthetic lethality networks

Synthetic lethality networks were randomized through random rewiring of edges. This involved random swapping of the source and target nodes of edges while preserving the degree distribution of the graph and degree correlations between nodes.

3.7.3 Small-world analysis

Small-world analysis of synthetic lethality networks was done by using the criteria defined by Humphries, Gurney, and Prescott [97]. The following parameters were calculated:

$$\gamma = \frac{C}{C_r} \tag{3.2}$$

$$\lambda = \frac{L}{L_r} \tag{3.3}$$

$$S = \frac{\gamma}{\lambda} \tag{3.4}$$

Here, C and L are the clustering coefficient and characteristic path length of a synthetic lethality network, respectively, and C_r and L_r are the same parameters obtained from a randomized version of the same network. In the definition, C_r and L_r are obtained from a single randomized network, but here the average of these parameters obtained from 100 random networks were used instead, all of them generated from the synthetic lethality network for which parameters were calculated. The criteria for small-world properties are $\gamma > 1$ and $S > 1$.

3.8 Statistical analyses

MATLAB (see Section 3.1.4) was used to perform several different statistical analyses, as described in this section.

3.8.1 Correlations

A matrix was prepared in which the columns corresponded to different data series. All pairwise correlations between columns and the p-values of all these correlations were then calculated using functions found in MATLAB's Statistics and Machine Learning Toolbox. In all cases, Pearson's linear correlation coefficient, r, was used.

3.8.2 Two-sample *t*-test

A two sample t-test was used to compare two different mean degrees of randomization for random viable metabolic networks, μ_1 and μ_2. The null and alternative hypotheses were

$$H_0: \quad \mu_1 = \mu_2 \tag{3.5}$$

$$H_1: \quad \mu_1 \neq \mu_2 \tag{3.6}$$

and the test statistic was

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.7}$$

Here, $\overline{x}_1$ and $\overline{x}_2$ are the calculated means of the two samples, s_1 and s_2 are the calculated standard deviations, and n_1 and n_2 are the sample sizes. The reported p-value was the two-tailed one.

3.8.3 Curve and distribution fitting

Most curves and probability distributions were fitted using built-in functions in MATLAB. The exception was the fitting of power laws, which was done by calculating maximum likelihood estimates of the parameters $x_{\min}$ and α – the minimum x value for power-law behavior and scaling parameter, respectively – as described in detail by Clauset, Shalizi, and Newman [75]. The calculations were performed using code available online [98].

3.9 Identification of alternative pathways for essential reactions

Here, the algorithm that was developed for identifying alternative metabolic pathways is described. The procedure for identifying the alternative pathways of a single reaction is first presented in general terms, followed by a description of the specific implementation that was used in this study.

3.9.1 The algorithm

The alternative pathways of a single reaction are identified as follows:

1. Start out with a model that has been merged with the reaction universe. The identities of reactions native to the model and the identity of the reaction for which alternative metabolic pathways should be found must also be known.

2. Define a set for storing alternative pathways and add the pathway for which alternative reactions should be found as the first pathway.

3. Predict the growth rate of the model using flux balance analysis with modified and additional constraints. A binary variable, b_i, indicating whether flux i is zero or nonzero, is introduced for each of the n fluxes in the system, producing a mixed integer programming problem:

$$b_i \in \{0,1\}, \quad i = 1,\ldots,n \tag{3.8}$$

   This also requires the lower and upper bounds of fluxes, l_i and u_i, to be changed:

$$l_i b_i \leq v_i \leq u_i b_i, \quad i = 1,\ldots,n \tag{3.9}$$

   One additional restriction is added for each known alternative pathway as well:

$$\sum_{i \in p_j} b_i < \left|p_j\right|, \quad j = 1,\ldots,n_p \tag{3.10}$$

   Here, p_j is pathway j with length $\left|p_j\right|$ and n_p is the total number of pathways found. The constraint states that the number of active reactions in each known pathway must be smaller than the total number of reactions in the pathway. In other words, all reactions cannot be active in any known alternative pathway. If the model is not viable when subjected to these constraints, there are no more alternative pathways to find and the procedure can end. If it is viable, the procedure moves to the next step.

4. Reactions that are not native to the model are knocked out one by one. If a knock-out mutant is viable as predicted by ordinary flux balance analysis, the knocked out reaction is removed, if it is not, the knocked out reaction is turned back on. All reactions with zero flux in the latest optimal solution are removed, as they cannot be essential. This generally speeds up the procedure dramatically.

5. When no more reaction knock-outs yield viable mutants, the remaining reactions are the ones native to the model and an alternative pathway. The alternative pathway is stored in the set of alternative pathways and the steps above are repeated, starting at step three.

3.9.2 Implementation

The alternative pathways of all essential reactions in genome-scale metabolic reconstructions were identified according to the following procedure:

1. Two models were imported into COBRApy: an existing viable model and a version of the same model merged with the reaction universe (see Section 3.4.2). The identifiers of reactions for which alternative pathways should be found and the identifiers of blocked reactions in the merged model were also imported.

2. The wild-type growth rate of the original model was determined.

3. All blocked reactions were removed from the merged model.

4. For each reaction for which alternative pathways should be found, alternative pathways were identified by using the algorithm described in Section 3.9.1. Only reactions in the same compartment as the reaction being analyzed were allowed to be part of alternative pathways, so all nonnative reactions were removed from all other compartments in the merged model before starting the pathway-finding procedure.

Chapter 4

Results and discussion

In this chapter, obtained results are presented and discussed. It is divided into the following main parts, all of which focus on making predictions about potential antimicrobial drug targets:

- *Analysis of the reaction universe*. The structure of the reaction universe was investigated and its metabolic capabilities were quantified in different cellular contexts through integration into existing genome-scale metabolic reconstructions. Sets of reactions predicted to always be essential were identified for each context.

- *Analysis of random viable metabolic networks*. The reaction contents of existing genome-scale metabolic reconstructions were randomized using the reaction universe. Reaction essentiality and synthetic lethality was investigated in the resulting random viable metabolic networks.

- *Identification of alternative metabolic pathways*. An algorithm was developed to identify sets of reactions from the reaction universe capable of replacing essential reactions in metabolic networks. This was used to identify alternative metabolic pathways for all essential reactions in existing genome-scale metabolic reconstructions, and the properties of these pathways were investigated.

The workflow through which the results presented here were obtained, from input data to output data, is illustrated in Figure 4.1.

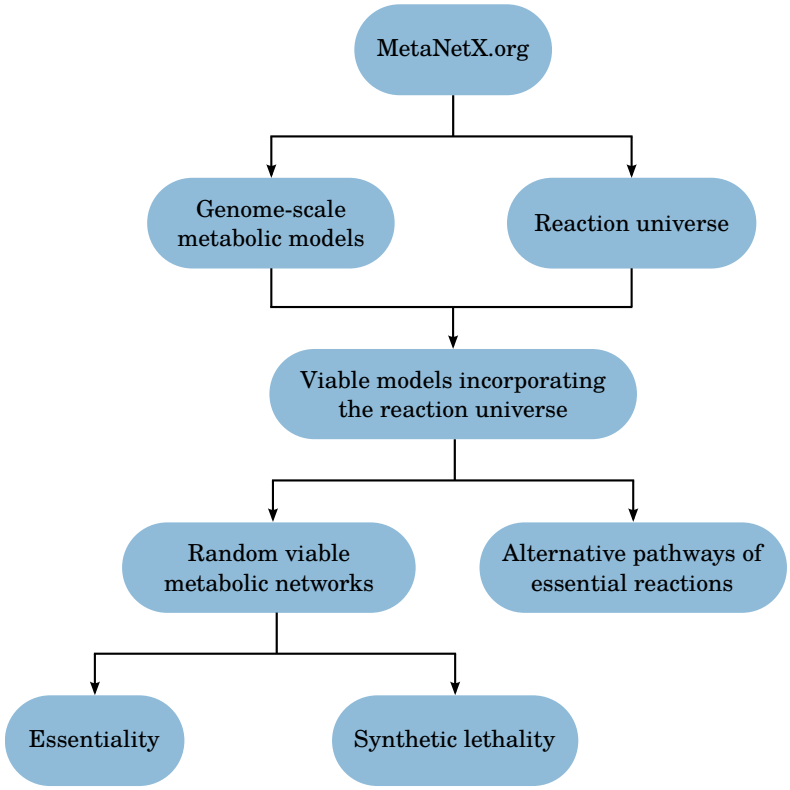


Figure 4.1: Flowchart describing the workflow of the project. Reaction, metabolite, and compartment data was obtained from MetaNetX.org [96] and used to construct a reaction universe, a data set containing all known metabolic reactions. This reaction universe was integrated into a collection of genome-scale metabolic models acquired from the same source and the resulting models were used as the basis of all the analyses that were performed. The two primary approaches involved the generation of random viable metabolic networks, in which essential reactions and synthetic lethal reaction pairs were identified, and the identification of alternative metabolic pathways capable of replacing essential reactions.

4.1 Analysis of the reaction universe

As defined in Section 2.4, the reaction universe is a collection of all known metabolic reactions. Such a data set was constructed as described in Section 3.4.1 and found to contain 13,849 reactions and 10,575 metabolites. It should be noted that only elementally balanced reactions that have been reconciled within the same namespace were included. The topology and metabolic capabilities of the metabolic network defined by this reaction universe were investigated and the reaction universe was compared to the one that has been used in previous studies.

4.1.1 Topology of the reaction universe

A graph representation of the reaction universe was prepared as described in Section 3.7.1. In this representation, each node corresponded to a metabolite and two nodes were connected by an edge if they were ever a reactant-product pair in a reaction. The graph contained 10,575 nodes, equal to the number of metabolites, divided across 36 different connected components, one of which contained more than 99 % of the nodes. Some simple parameters describing the topology of this giant component are listed in Table 4.1 and Figure 4.2 shows its node degree distribution. A visualization of the graph is included in Appendix C.

Table 4.1: Some simple parameters of the giant component found in the graph representation of the reaction universe. The shown parameters are average node degree ($\langle k \rangle$), centralization (C_D), clustering coefficient (C), and characteristic path length (L).

$\langle k \rangle$	C_D	C	L
7.56	0.457	0.339	2.81

Each metabolite in the graph was connected to 7.56 others on average. The relatively large clustering coefficient, in combination with a low characteristic path length, hints that the graph had small-world properties, a well-known feature of real metabolic networks [99]. In addition to this, the centralization parameter could indicate that some nodes in the graph were highly central hubs, another frequently highlighted feature of metabolism [100]. This is supported by the node degree distribution, which is very similar to typical distributions obtained from real metabolic networks [2, 75] and was found to follow a power law closely for a range of node degrees ($5 \leq k \leq 100$).

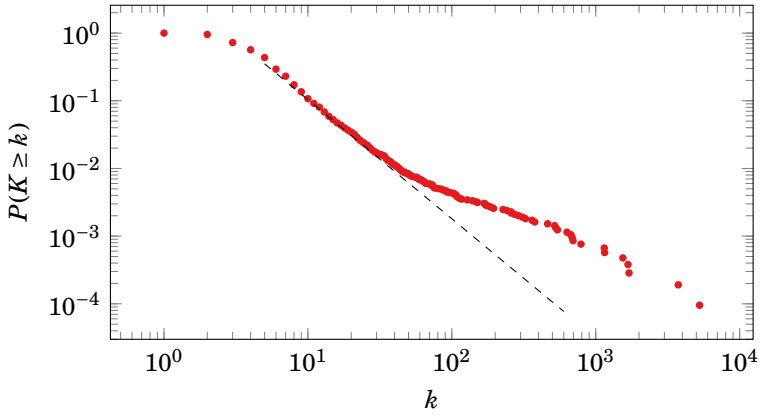


Figure 4.2: Log-log plot showing the node degree distribution of the giant component found in the graph defined by the reaction universe. $P(K \geq k)$ is the cumulative probability distribution of the node degree k. A power law with parameters $x_{\min} = 5$ and $\alpha = 2.76$ was fitted to the data. Its cumulative, $P(K \geq k) = 5.98k^{-1.76}$, was found to fit the cumulative distribution of low node degrees well with coefficient of determination $R^2 = 0.996$ in the range $5 \leq k \leq 100$.

The power law was fitted as described in Section 3.8.3. The deviation from the power law observed for higher node degrees, which can clearly be seen in Figure 4.2, could be explained in terms of currency and commodity metabolites. Currency metabolites are small and ubiquituous metabolites, such as H^+ and H_2O, and cofactors, such as ATP and NADH, that are expected to participate in a very large number of reactions. Commodity metabolites are the ones that are not currency. The ten metabolites that were found to be most highly connected, which are listed in Table 4.2, were confirmed to be among those often cited as currency metabolites [99, 101, 102]. Overall, it is interesting to observe that the reaction universe, despite being much larger than any realistically sized, evolved metabolic network, exhibits many of the same topological properties.

4.1.2 Metabolic capabilities of the reaction universe

In order to explore the metabolic capabilities of the reaction universe in a variety of cellular contexts, it was integrated into 43 different growth-predictive genome-scale reconstructions of microbial metabolisms. Most of these were models of bacteria, but one archaea and one unicellular eukaryote were also included. Cellular context here refers to all aspects of a metabolic model except the metabolic reactions themselves. This includes the metabolic re-

Table 4.2: The ten most highly connected metabolites in the reaction universe and their node degree, k.

Metabolite	k
H^+	5,267
H_2O	3,746
O_2	1,703
NADPH	1,676
$NADP^+$	1,544
NADH	1,153
NAD^+	1,146
ATP	790
Coenzyme A	696
CO_2	686

quirements for growth as expressed by the biomass reaction, available compartments and intercompartmental transport reactions, and the boundary reactions that define the extracellular environment.

The integration procedure is reported in Section 3.4.2. In brief, all reactions from the reaction universe were added to every intracellular compartment in every reconstruction as long as they were not already present. Properties of the reconstructions with which the reaction universe was merged as well as the models resulting from this merging are listed in Appendix D.

Effects on viability and robustness of metabolic reconstructions

Growth rates were determined and all essential metabolic reactions were identified for all models both before and after merging. The results are shown in Figures 4.3 and 4.4, respectively. In all cases, the growth rate of a model after merging was greater or equal to the growth rate of the same model before merging. This is as expected, since the addition of reactions to a metabolic network cannot diminish its ability to produce biomass precursors, only increase it or have no effect. In mathematical terms, the addition of new dimensions to the solution space of a model cannot lead to a less optimal solution. The effect that reactions from the reaction universe had on the growth rate of the models to which they were added varied greatly, from no effect to increases on the order of 10^2. A number of causes may have contributed to this variation, including growth medium and biomass composition, available intercompartmental transport reactions, and the size of the original network.

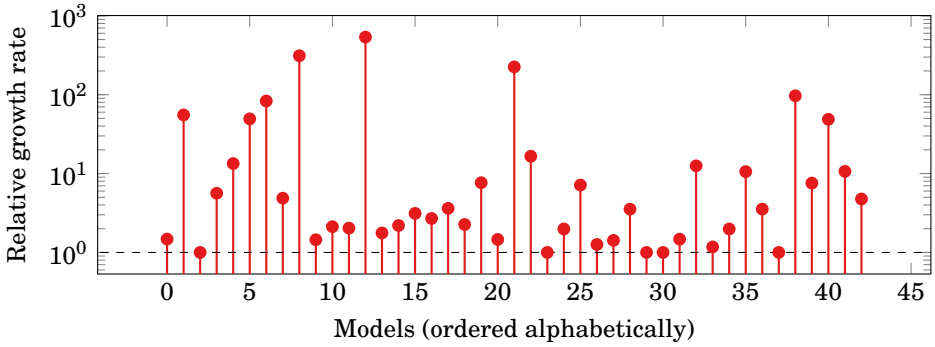


Figure 4.3: Growth rates of models after merging with the reaction universe relative to growth rates before merging. The growth rate of a model after merging was always greater or equal to its growth rate before merging. The vertical axis is logarithmic. The dashed line indicates a relative growth rate of one, i.e. that the growth rates before and after merging were equal.

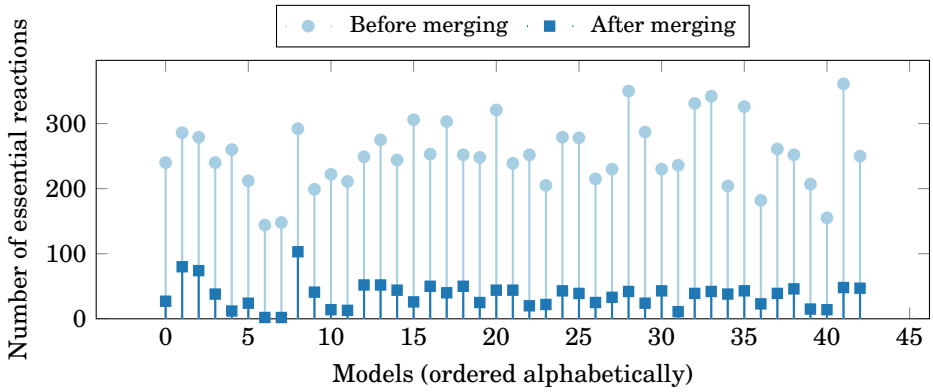


Figure 4.4: Number of essential reactions in models before and after merging with the reaction universe. Merging always reduced the number of essential reactions, indicating increased network robustness. The minimum number of reactions identifed in any model after merging was two.

The number of essential reactions decreased in all models upon addition of reactions from the reaction universe, indicating increased network robustness. Importantly, the essential reactions that were identified after merging were always found to be a subset of the reactions that were identified as essential before merging. This reflects the fact that adding reactions to a metabolic network cannot cause reactions to become essential, just as it cannot reduce growth rate. However, originally essential reactions may become nonessential if the reactions that are added connect to the original network in ways that allow the functions of the originally essential reactions to be performed without them.

The mean number of essential reactions in a model was 252 (s = 52) before merging and 36 (s = 20) after. The minimum number of essential reactions retained after merging was two. This was observed for the two eukaryotic models that were investigated, both yeast reconstructions with seven intracellular compartments each. The number of compartments is the most likely explanation for why these models contained fewer essential reactions than any other. Considering that all reactions from the reaction universe were added to every compartment in a model, the chance of replacing essential reactions should be expected to increase with the number of compartments in a model.

Absolutely superessential reactions

The reactions that remained essential in a model after merging with the reaction universe could not be replaced by any set of metabolic reactions. This means that they are predicted to be irreplacable in any metabolic network in the same cellular context, what Barve, Rodrigues, and Wagner [28] refers to as absolutely superessential (see Section 2.3.2). Here, 402 different reactions were found to be absolutely superessential in at least one cellular context. As shown in Figure 4.5, the majority of these reactions, 83 %, were identified as absolutely superessential only once, meaning that the majority of absolutely superessential reactions were context-specific. The reactions that were absolutely superessential in more than one context were identified 17 (s = 20) times on average. The largest number of contexts in which a reaction was absolutely superessential was 38, corresponding to 88 %. According to these results, no set of metabolic reactions that must always be essential in any metabolic network regardless of context exists. In other words, there are no context-general absolutely superessential reactions.

Among the reactions that were identified as absolutely superessential were five that were found to be so in more than 70 % of the analyzed contexts. These reactions are predicted to be irreplacable in all possible metabolic

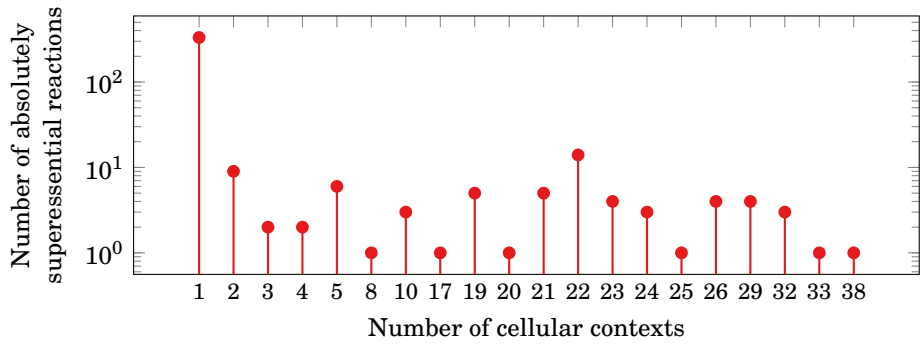


Figure 4.5: Number of cellular contexts in which reactions were identified as absolutely superessential. The vertical axis indicates the number of reactions and the horizontal axis the number of contexts in which these reactions were absolutely superessential. The vertical axis is logarithmic. All reactions that were found to be absolutely superessential in at least one context are included. Although some reactions were identified as absolutely superessential in many different contexts, most were context-specific and only identified once.

networks under a large range of different internal and external conditions and could therefore be promising candidates for broad-spectum antimicrobial drug targets. Table 4.3 lists Enzyme Commission (EC) numbers, and metabolic subsystem data for the enzymes catalyzing these reactions.

One of the enzymes listed in Table 4.3 is involved in peptidoglycan biosynthesis and the remaining four are associated with riboflavin metabolism. Known antimicrobial drug targets are found within both of these subsystems. Peptidoglycan forms the cell wall of most bacteria and its synthesis is the target of several widely used classes of antibiotics, notably β-lactams such as penicillins [20]. Reactions in riboflavin metabolism have been explored as targets for antimicrobial riboflavin and flavin mononucleotide (FMN) analogs. One example is roseoflavin, a natural antibiotic that inhibits the expression of genes encoding enzymes involved in riboflavin biosynthesis and transport through binding of associated riboswitches [103–105].

4.1.3 Comparison to a previously studied reaction universe

The reaction universe previously studied by Barve, Rodrigues, and Wagner [28] was compared to the one used in this study[1]. These two data sets will from now on be referred to as the old and the new reaction universe, respectively. The old reaction universe consisted of data obtained from databases in

[1]The previously studied data set was provided by courtesy of Dr. Aditya Barve.

Table 4.3: EC numbers, and metabolic subsystems of the enzymes catalyzing the five absolutely superessential reactions that were found in more than 70 % of the analyzed cellular contexts. Data obtained from MetaNetX.org [96], KEGG [32], and MetaCyc [106]. The percentage of cellular contexts in which the catalyzed reaction was absolutely superessential is shown for each enzyme.

EC number(s)	Subsystem	Contexts
2.7.8.13	Peptidoglycan biosynthesis	88 %
2.7.1.26	Riboflavin metabolism	77 %
2.5.1.9	Riboflavin metabolism	74 %
2.5.1.78, 2.5.1.9	Riboflavin metabolism	74 %
3.1.3.-	Riboflavin metabolism	74 %

the Kyoto Encyclopedia of Genes and Genomes (KEGG) [32, 33] merged with a genome-scale metabolic reconstruction of *Escherichia coli* [41], whereas the new one was constructed from reactions in the MNXref namespace available through MetaNetX.org [95, 96] (see Section 3.4.1). The old and new reaction universes contained 5,905 and 13,849 metabolic reactions, respectively. It was found that 4,801 reactions were shared between them, leaving 1,104 reactions unique to the old reaction universe and 9,048 reactions unique to the new one. This is illustrated in Figure 4.6.

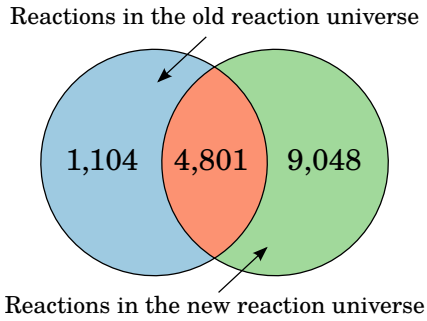


Figure 4.6: Venn diagram showing the overlap between reactions in the old reaction universe and the new. The old and new universes contained 5,905 and 13,849 reactions, respectively, and 4,801 reactions were shared between them.

The reasons for not including the reactions unique to the old reaction universe in the new one were elucidated. It was found that two of these reactions were not included because they were removed from the *E. coli* reconstruction, 38 were no longer present in KEGG, and 152 were duplicates as

defined by the namespace of the new reaction universe. Out of the remaining 912 reactions unique to the old reaction universe, 867 were not included in the new because they were not defined as elementally balanced and the rest, 45, because they were considered to be transport reactions.

The new reaction universe was much larger than the old one, containing more than twice as many reactions. There are at least two main reasons for this. First of all, the old reaction universe was constructed five to six years prior to the new one (Dr. Aditya Barve, personal communication, April 2014) and therefore does not incorporate information that has become available in KEGG in recent years. Second of all, the MNXref namespace that was used to construct the new reaction universe integrates metabolite and reaction data not only from KEGG, but from many other sources as well [95]. Such integration has been difficult to achieve until quite recently due to lack of standardization in nomenclature and conventions, but MNXref and several other recent efforts have successfully reconciled data from different databases and models within single namespaces [107–109].

4.1.4 Limitations of the reaction universe

Here, some potential caveats concerning the reaction universe are listed:

- Our current knowledge of metabolism is not complete and it is unlikely that the data set constructed here truly contains all the elementally balanced reactions that occur in metabolic systems. The number of reactions found in databases has grown continuously until now and it seems unrealistic to assume that this trend will not continue in the years to come.

- Databases can contain errors, for example putative reactions without experimental evidence.

- Databases are bound to be biased to some degree, as some organisms and metabolic subsystems are much more well-studied than others.

- Many reactions found in metabolic models are currently not incorporated into the MNXref namespace. This could for example have caused duplicate reactions to be included in the reaction universe.

- There is generally little reversibility data available for metabolic reactions. Here, all reactions were assumed to be reversible, but under real biological conditions this may not be the case.

4.2 Analysis of random viable metabolic networks

Random viable metabolic networks were generated from existing genome-scale metabolic reconstructions by repeatedly swapping reactions with the reaction universe. This randomization procedure is presented in detail in Section 3.5. In previous studies using the same approach, only one compartment in the metabolic network of a single organism has been randomized, but here the procedure was slightly modified to allow randomization of all well-defined intracellular compartments in many different models. All manually curated models found in the collection of growth-predictive models listed in Appendix D were randomized – ten models in total spanning all three domains of life. The names of these models and their randomized intracellular compartments are presented in Table 4.4. For each model, 5,000 randomized networks were sampled, all of them viable and containing the same number of metabolic reactions as the original model in all compartments.

Table 4.4: Model names and names of intracellular compartments that were randomized for the ten models from which random viable metabolic networks were generated.

Model name	Compartments randomized
*i*AF1260	Cytoplasm, periplasm
*i*AF692	Cytoplasm
*i*Cyt773	Cytoplasm
*i*IT341	Cytoplasm
*i*JN746	Cytoplasm, periplasm
*i*JR904	Cytoplasm
*i*MM904	Cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion, nucleus, peroxisome, vacuole
*i*ND750	Cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion, nucleus, peroxisome, vacuole
*i*NJ661	Cytoplasm
*i*YO844	Cytoplasm

The network randomization method was evaluated by investigating its ability to randomize metabolic networks and the metabolic reaction contents of the sampled randomized networks. To this end, all blocked reactions, essential reactions, and synthetic lethal reaction pairs were identified in all sampled networks. Essentiality and synthetic lethality data was further used to examine the potential for reaction essentiality and synthetic lethality in different cellular contexts and to investigate superessentiality and its

extension to synthetic lethal rection pairs. Finally, the large-scale organization of synthetic lethal interactions between reaction pairs was explored in a graph-based approach.

4.2.1 Evaluation of the randomization procedure

The degree of randomization was determined for each sampled random viable metabolic network as the fraction of reactions not shared with the model from which it was generated. The average degree of randomization and the average fractions of metabolic reactions that were essential, blocked, or participated in one or more synthetic lethal reaction pairs were calculated for each of the ten models from which random viable metabolic networks were generated. Figure 4.7 shows all of these averages for all models.

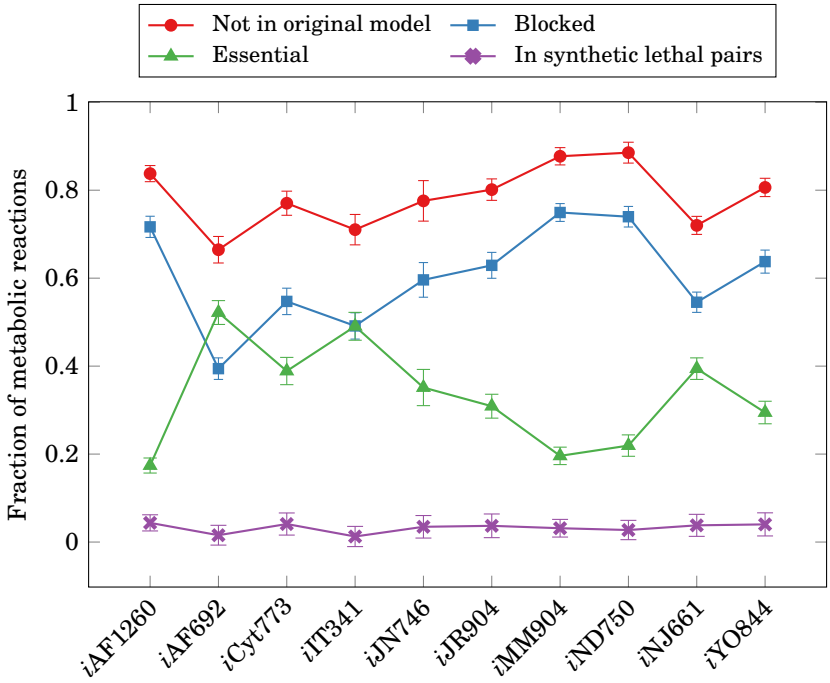


Figure 4.7: Average fraction of metabolic reactions in different categories found in random viable metabolic networks. The average fraction of reactions not found in the original model (degree of randomization) is shown along with the average fractions of reactions found to be blocked, essential, or participating in a synthetic lethal pair. Names of the models from which randomized networks were generated are indicated along the horizontal axis. The error bars indicate two standard deviations. For each model, 5,000 randomized networks were generated and analyzed.

The average degree of randomization over all models was 0.78 with minimum and maximum values of 0.67 and 0.89, respectively. The mean fraction of blocked reactions was 0.69, varying between 0.39 and 0.75, the mean fraction of essential reactions was 0.33, varying between 0.17 and 0.52, and the mean number of reactions participating in synthetic lethal reaction pairs was 0.03, varying between 0.01 and 0.04. With one exception, the generated random viable metabolic networks contained more than 50 % blocked reactions on average and a smaller average fraction of essential reactions. Moreover, very few reactions were neither blocked nor essential, the mean fraction of such reactions being 0.06. These reactions are the only ones that can possibly form synthetic lethal reaction pairs with each other in a metabolic network, explaining the low mean fraction of reactions found to participate in synthetic lethal reaction pairs.

Correlations with metabolic network size

A great deal of variation between random viable metabolic networks generated from different models is apparent in Figure 4.7. It was hypothesized that metabolic network size was the key determinant of this variation, and indeed, strong and significant correlations were found between the number of metabolic reactions in a model and the calculated averages. Pearson's r and p-values for these correlations are listed in Table 4.5. The identified correlations indicate that randomization of large models yielded random viable metabolic networks with a higher degree of randomization than networks generated from smaller ones. Also, networks generated from large models contained a higher fraction of blocked reactions, a lower fraction of essential reactions, and a higher fraction of reactions participating in synthetic lethal pairs. All of this may be understood in terms of the structure of metabolic networks in general and the way randomization was performed.

First of all, small metabolic networks should generally be expected to be less flexible and robust than large ones, as fewer reactions will usually mean fewer possible ways to produce the biomass precursors needed for growth. This in turn implies that a larger share of reactions is likely to be essential in small networks and explains why small random viable metabolic networks were found to contain more essential reactions than large ones. Second of all, randomization was performed by repeatedly and randomly removing nonessential reactions from a metabolic network and adding new ones. There was no inherent pressure on the reactions that were added to connect to the metabolic network being randomized and thus many reactions were probably blocked upon addition. This explains why large fractions of reactions were blocked in the sampled networks.

Table 4.5: Correlations between the size of metabolic networks and properties of random viable metabolic networks generated from them. Pearson's r and p-values are listed. Correlations were calculated between the number of metabolic reactions found in ten different models and the mean values of properties calculated for 5,000 random viable metabolic networks generated from each of these models.

Property	r	p
Degree of randomization	0.67	0.034
Fraction blocked reactions	0.73	0.001
Fraction essential reactions	-0.83	0.003
Fraction reactions in synthetic lethal pairs	0.68	0.031

Occasionally, added reactions must have connected to a network being randomized, possibly along with one or more previously blocked reaction, and in some cases, this must have caused previously essential reactions to become nonessential. The more nonessential reactions that were present in a network, including blocked ones, the higher the probability would have been for such replacement to happen. This means that it should have been harder to replace essential reactions in small networks than in large ones, and it follows that fewer essential reactions should have been replaced when randomizing small networks than large ones. This would explain the lower degrees of randomization observed for small random viable metabolic networks. Finally, the larger fraction of reactions found to participate in synthetic lethal pairs in large networks may be considered a consequence of the fraction of reactions that were neither blocked nor essential. As previously stated, all observed synthetic lethal reaction pairs must have been formed by these reactions and the large randomized networks contained more of them than the small ones.

Comparison to previously obtained results

The *i*AF1260 *E. coli* model has been randomized before and the average degree of randomization achieved here for this model, 0.84, is higher than what has previously been reported (Dr. Aditya Barve, personal communication, April 2014). This indicates that the updated reaction universe and randomization of multiple compartments made it possible to generate random viable metabolic networks that shared fewer reactions with the models from which they were generated. To verify this, 100 random viable metabolic networks were generated from the *i*AF1260 model using the old reaction universe de-

scribed in Section 4.1.3 and the randomization method as described by Barve, Rodrigues, and Wagner [28]. This gave a mean degree of randomization of 0.67. A two-sample t-test led to the conclusion that randomization using the new reaction universe and modified method resulted in degrees of randomization that were significantly higher than what has been attainable in the past ($t = 175.6$, $p \approx 0$). The t-test is described in Section 3.8. The primary cause of the increased ability to randomize metabolic networks is thought to be the size of the reaction universe.

4.2.2 Potential for essentiality and synthetic lethality

Which reactions can possibly be essential or participate in a synthetic lethal reaction pair in different cellular contexts, and which synthetic lethal pairs can occur? The essential reactions and synthetic lethal reaction pairs that were identified in random viable metabolic networks were used to answer these questions. Table 4.6 lists the number of different essential reactions, reactions participating in synthetic lethal reaction pairs, and synthetic lethal pairs that were identified at least once in each of the ten cellular contexts that were studied. These sets of reactions and reaction pairs define the predicted potential for reaction essentality and synthetic lethality in different cellular contexts.

Table 4.6: Number of different essential reactions, reactions participating in synthetic lethal pairs, and synthetic lethal pairs found in random viable metabolic networks generated from different models. For each model, reaction essentiality and synthetic lethality was analyzed in 5,000 randomized networks.

Model name	Essential	Synthetic lethal	
		Reactions	Pairs
*i*AF1260	6,149	6,237	181,806
*i*AF692	4,780	3,759	43,271
*i*Cyt773	4,910	4,590	119,620
*i*IT341	4,335	3,074	26,060
*i*JN746	4,964	4,860	136,045
*i*JR904	4,931	4,074	125,028
*i*MM904	6,910	6,154	148,136
*i*ND750	6,830	5,623	134,308
*i*NJ661	5,126	4,948	161,497
*i*YO844	5,060	4,989	145,251

Correlations with metabolic network size

Again, correlation with network size was suspected and investigated. This led to the identification of strong and significant positive correlations between the number of metabolic reactions in a network and the number of essential reactions, reactions participating in synthetic lethal pairs, and synthetic lethal pairs identified. Pearson's r and p-values for these correlations are listed in Table 4.7. The observed correlations may be interpreted as an indication that not all potential for essentiality and synthetic lethality was uncovered. Although it is possible that the potential for reaction essentiality and synthetic lethality is lower in small metabolic networks than in large ones, there is no *a priori* reason to expect this. Instead, it is assumed that the number of identified reactions and reaction pairs was limited by the size of the network samples. Other than this, and besides the fact that a smaller network size means fewer potential reactions to sample from each network, the causes of the correlations are thought to be the same as those discussed in Section 4.2.1. Firstly, it is believed that the randomization procedure replaced essential reactions more rarely in small networks than in large ones, causing less variation among the essential reactions of random viable metabolic networks generated from small models. Secondly, there were fewer nonessential, nonblocked reactions to potentially form synthetic lethal pairs in small networks than in large ones.

Table 4.7: Correlations between the size of metabolic networks and the number of different essential reactions, reactions participating in synthetic lethal pairs, and synthetic lethal pairs that were identified in random viable metabolic networks generated from them. Pearson's r and p-values are listed. Correlations were calculated between the number of metabolic reactions found in ten different models and the mean values of properties calculated for 5,000 random viable metabolic networks generated from each of these models.

Property	r	p
Essential reactions	0.69	0.027
Reactions in synthetic lethal pairs	0.87	0.002
Synthetic lethal pairs	0.84	0.001

Compartmental variation

The compartmental variation in the potential for essentiality and synthetic lethality was also explored. An overview of the number of essential reactions,

reactions participating in synthetic lethal pairs, and synthetic lethal pairs found in different compartments in multicompartment models is provided in Appendix E. In all cases, the greatest potential for essentiality and synthetic lethality was uncovered in the cytoplasm, which contained the largest number of metabolic reactions in all models from which random viable metabolic networks were generated. Once more, the results indicate that network size – this time the number of metabolic reactions per compartment – was the primary factor determining the number of reactions and reaction pairs identified, and again it is argued that the cause of this is that the ability of the randomization method to randomize a metabolic network is diminished as network size decreases. Also, nearly all biomass precursors in all models were cytoplasmic metabolites whose production will virtually always require reactions in the cytoplasm.

Context-general essentiality and synthetic lethality

The ten sets of essential reactions, reactions participating in synthetic lethal pairs, and synthetic lethal pairs listed in Table 4.6 were compared in order to identify overlaps between contexts. Only cytoplasmic reactions were included in the comparison, as this was the only compartment shared between all cellular contexts. Figure 4.8 shows the distribution of the number of cellular contexts in which essential reactions and reactions participating in synthetic lethal pairs were found.

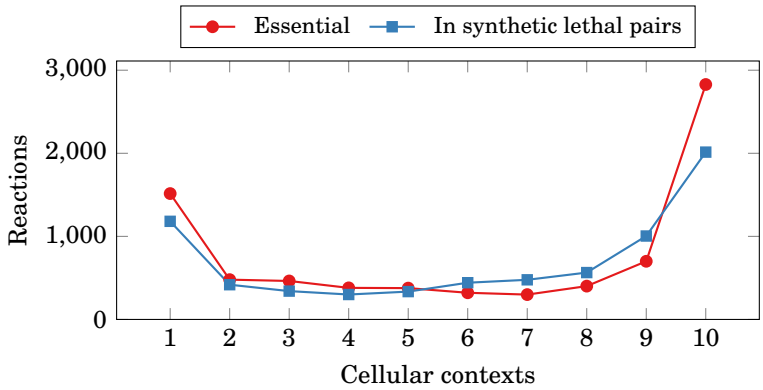


Figure 4.8: Number of different cellular contexts in which cytoplasmic reactions were essential or participated in a synthetic lethal pair. The vertical axis indicates the number of different reactions and the horizontal axis indicates the number of cellular contexts. A large share of cytoplasmic reactions were essential or participated in a synthetic lethal pair in all ten contexts.

It can be seen from Figure 4.8 that the largest fraction of essential cytoplasmic reactions, 36 %, were found in networks generated from all ten original models. This suggests that many essential reactions are context-general, meaning that they are essential due to factors that are common across organisms and environmental conditions. The distribution obtained for reactions participating in synthetic lethal pairs was very similar to the one found for essential reactions. As for essential reactions, the largest fraction of reactions that participated in synthetic lethal pairs, 28 % in this case, did so in all investigated cellular contexts. Results presented by Barve, Rodrigues, and Wagner [28] lend some support to these findings. Based on analyses of superessentiality in a wide variety of environments differing in their sole carbon source, they found that essential reactions are usually environment-specific or environment-general, the latter being the most common.

Figure 4.8 also shows that a larger-than-average fraction of identified essential reactions and reactions participating in synthetic lethal pairs were only found in networks generated from a single model. As shown in Figure 4.9, this was largely caused by essential reactions unique to networks generated from *i*Cyt773, a reconstruction of the photosynthetic cyanobacterium *Cyanothece* whose biomass equation includes many metabolites that are not found in any of the other models from which randomized networks were generated. Reactions that allow production of these unique biomass precursors are likely to be essential or part of synthetic lethal reaction pairs only in random viable metabolic networks generated from this model.

The distribution of the number of cellular contexts in which synthetic lethal reaction pairs were found is shown in Figure 4.10. In contrast to what was found for essential reactions and reactions participating in synthetic lethal pairs, the vast majority of synthetic lethal reaction pairs, 72 %, were identified only in a single cellular context. Even so, 1,475 different synthetic lethal pairs, 2 % of all the pairs that were identifed, were found to be context-general.

4.2.3 Superessentiality

The superessentiality index defined by Barve, Rodrigues, and Wagner [28], I_{SE}, was calculated for all reactions as described in Section 3.6. For each of the ten models from which random viable metabolic networks were generated, one set of superessentiality indices was obtained for each compartment. Rank plots of superessentiality indices for all compartments in which more than ten reactions were identified as essential at least once (see Appendix E) are shown for all cellular contexts in Figure 4.11.

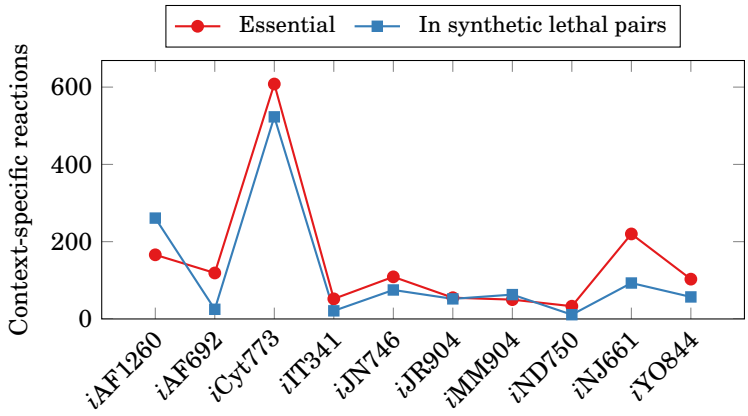


Figure 4.9: Number of context-specific essential reactions and reactions participating in synthetic lethal reaction pairs by cellular context. Only cytoplasmic reactions are included. Most essential reactions and reactions participating in synthetic lethal pairs that were only identified in random viable metabolic networks generated from a single model were found in networks randomized from the *i*Cyt773 *Cyanothece* reconstruction. The vertical axis indicates the number of different context-specific reactions and the horizontal axis indicates the names of the ten models from which random viable metabolic networks were generated.

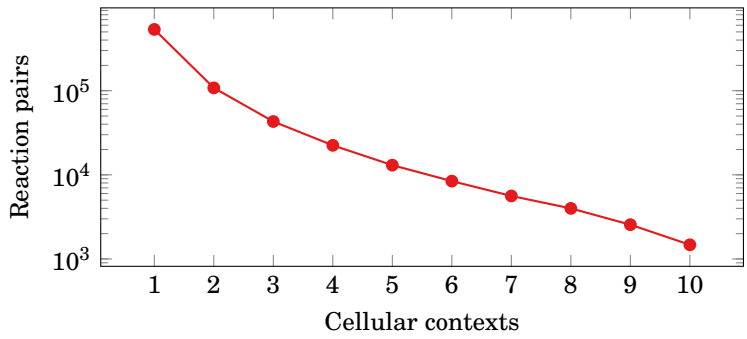


Figure 4.10: Number of different cellular contexts in which cytoplasmic synthetic lethal reaction pairs were identified. The vertical axis indicates the number of different synthetic lethal reaction pairs and the horizontal axis indicates the number of cellular contexts in which these reaction pairs were identified in random viable metabolic networks. The vertical axis is logarithmic. The vast majority of reaction pairs were only identified in one cellular context.

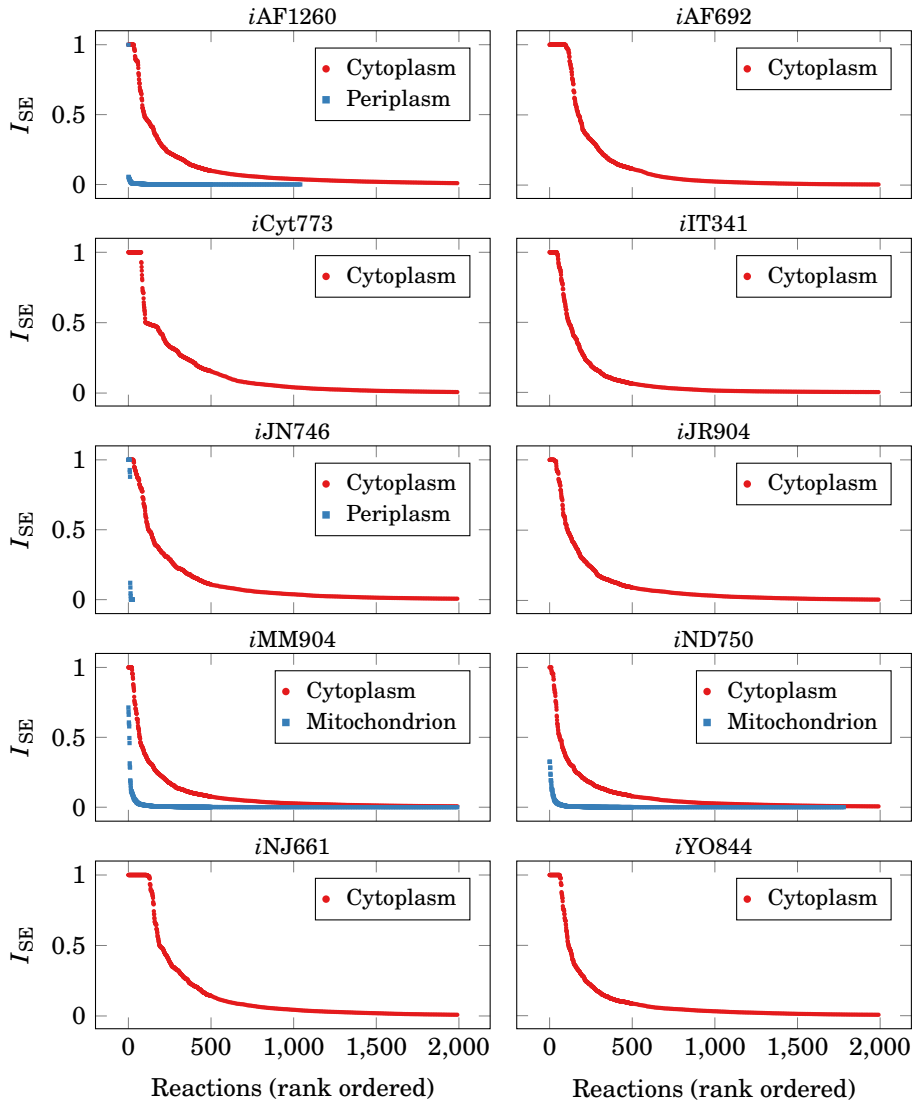


Figure 4.11: Rank plots of reaction superessentiality indices, I_{SE}, for all models from which random viable metabolic networks were generated. For each model, indices were determined from reaction essentiality in 5,000 random viable metabolic networks and one superessentiality index was calculated for each reaction in each intracellular compartment. Reactions are ordered by superessentiality index from high to low for each compartment. Only the 2,000 reactions with largest indices are included in cases where nonzero indices were found for more than 2,000 reactions, otherwise all nonzero indices are shown.

Only one intracellular compartment, the cytoplasm, was shared by all randomized models. Therefore, the only reactions for which superessentiality indices were calculated for all models were cytoplasmic ones. As can be seen from Figure 4.11, similarily shaped distributions were obtained for all cytoplasmic superessentiality indices, all of them closely resembling the one presented by Barve, Rodrigues, and Wagner [28] (see Figure 2.7 in Section 2.3). Each distribution consisted of a plateau of variable size for $I_{\mathrm{SE}} = 1$, corresponding to cytoplasmic reactions that were essential in all randomized networks, followed by a gradually flattening slope of intermediate superessentiality indices. Most reactions were identified as essential only in one or a few randomized networks and all distributions therefore had long tails, only the beginnings of which are shown in the plots.

The only noncytoplasmic compartments that yielded superessentiality indices for more than ten reactions were the periplasms of the *i*AF1260 and *i*JN746 models – reconstructions of *E. coli* and *Pseudomonas putida*, respectively – and the mitochondria of the two *Saccharomyces cerevisiae* reconstructions, *i*MM904 and *i*ND750. The two periplasm distributions differed greatly, the *E. coli* one consisting of many rections with close-to-zero superessentiality indices and the *P. putida* one only including a few reactions with comparatively large indices. The two mitochondrial distributions, on the other hand, were very similar. The cause of the differences between indices observed for the two periplasms is not entirely clear, but the periplasm of the *E. coli* model contained many more reactions than the *P. putida* one (see Appendix D). As discussed previously, for example in Section 4.2.1, larger networks were more easily randomized than small ones, meaning that it should have been easier for the randomization procedure to replace reactions in the periplasm of *E. coli* than in *P. putida*. The similarity between the mitochondrial superessentiality indices of the two yeast models is easily explained, as these two models are closely related. In fact, *i*MM904 is a modified version of *i*ND750 [53].

Comparison to absolutely superessential reactions

The sets of reactions with $I_{\mathrm{SE}} = 1$ were compared to the sets of absolutely superessential reactions identified in Section 4.1.2. The latter were always found to be subsets of the former, which was in fact guaranteed from the implementation of the randomization procedure (see Section 3.5). In all cellular contexts, some reactions that were not absolutely superessential were identified as essential in all random viable metabolic networks, meaning that, although it was known beforehand that these reactions were replacable, they were never replaced. The sizes of the sets of absolutely superessential re-

actions and reactions with $I_{SE} = 1$, all compartments included, are shown in Table 4.8. The mean difference between these sets over all cellular contexts was found to be 12 reactions ($s = 8$). Many of the false predictions of absolutely superessential reactions from random sampling were likely consequences of the sample size being too low, but it is not impossible that some reactions were in fact irreplacable in random viable metabolic networks despite being replacable in the corresponding model merged with the reaction universe. If this was ever the case, the cause would be the very large difference in network size between the reaction universe and the randomized metabolic networks.

Table 4.8: Number of reactions with $I_{\mathrm{SE}} = 1$ and number of absolutely superessential reactions identified for all models from which random viable metabolic networks were generated. The difference between the two numbers is also listed for each model. The absolutely superessential reactions were always subsets of the reactions with $I_{\mathrm{SE}} = 1$, and in all cases, more reactions belonged to the latter set than to the former.

Model name	$I_{\mathrm{SE}} = 1$	Abs. superessential	Difference
*i*AF1260	34	27	6
*i*AF692	93	80	13
*i*Cyt773	78	74	4
*i*IT341	45	38	7
*i*JN746	38	12	18
*i*JR904	27	24	3
*i*MM904	23	2	20
*i*ND750	13	2	9
*i*NJ661	110	103	7
*i*YO844	61	41	20

Average superessentiality indices

The superessentiality indices shown in Figure 4.11 gave insights into how difficult metabolic reactions were to replace in various cellular contexts considered separately. However, as previously discussed, potential targets for broad-spectrum antimicrobial drugs should be as difficult to replace as possible in as many cellular contexts as possible. Therefore, to identify promising drug targets, an average superessentiality index, $\langle I_{\mathrm{SE}} \rangle$ was calculated for each reaction as described in Section 3.6. This index quantifies how often a

reaction should be expected to be essential across metabolisms of different sizes under a range of internal and external conditions. The 100 largest average superessentiality indices that were found, all of them cytoplasmic, are shown in the rank plot in Figure 4.12. These indices were found to be fairly linearily distributed, ranging from a maximum value of 0.94 to a minimum of 0.31. Varying degrees of uncertainty were associated with the values. Some standard deviations of the mean were comparatively low and some were large enough to span the most of the spectrum from zero to one. The ten reactions with largest $\langle I_{\mathrm{SE}} \rangle$ were considered promising candidates for antimicrobial drug targets and were investigated in more detail. Information about the enzymes catalyzing these reactions can be found in Table 4.9.

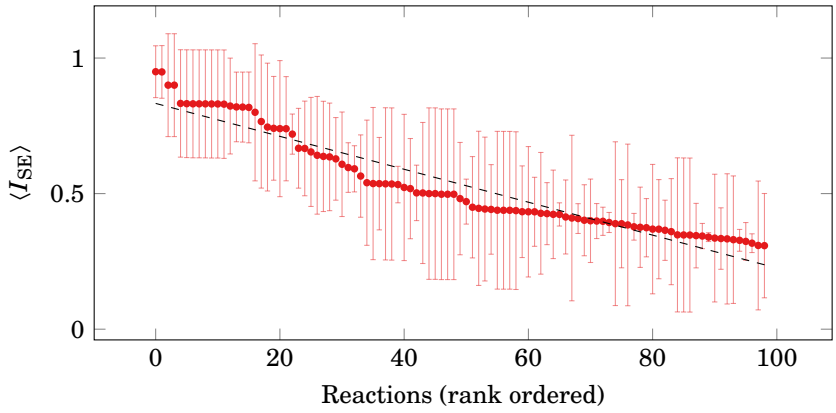


Figure 4.12: Rank plot of average superessentiality indices, $\langle I_{\mathrm{SE}} \rangle$. The 100 reactions with largest indices are included, all of them cytoplasmic. The error bars indicate standard deviations of the mean. The line $y = -6.07 \cdot 10^{-3} x + 0.833$ was fitted with coefficient of determination $R^2 = 0.919$.

All the ten reactions with largest average superessentiality indices were found in the same two metabolic subsystems, the top four being associated with purine metabolism and the six remaining being involved in the biosynthesis of histidine. Neither of these subsystems have seen much use as targets for antimicrobial drugs in the past, but both DNA, which would be an indirect target of drugs targeting purine metabolism, and enzymes involved in histidine metabolism have been highlighted as areas that are underexplored for antimicrobial purposes [110, 111].

One should note that none of the enzymes or subsystems listed in Table 4.9 are represented among the top absolutely superessential reactions presented in Table 4.3 in Section 4.1.2. Thus, none of the ten reactions with largest $\langle I_{\mathrm{SE}} \rangle$ were found to be irreplacable in more than 70 % of the models

Table 4.9: EC numbers, and metabolic subsystems of the enzymes catalyzing the ten reactions with largest average superessentiality indices. Data obtained from MetaNetX.org [96] and KEGG [32]. The average superessentiality index, $\langle I_{\mathrm{SE}} \rangle$, is included for each reaction with uncertainty equal to the standard deviation of the mean.

EC number(s)	Subsystem	$\langle I_{\mathrm{SE}} \rangle$
4.3.2.2	Purine metabolism	0.950 ± 0.096
2.1.2.3, 3.5.4.10	Purine metabolism	0.949 ± 0.097
6.3.5.3	Purine metabolism	0.900 ± 0.190
6.3.3.1	Purine metabolism	0.900 ± 0.190
5.3.1.16	Histidine metabolism	0.833 ± 0.198
2.6.1.9	Histidine metabolism	0.831 ± 0.199
4.2.1.19	Histidine metabolism	0.831 ± 0.199
2.4.2.-, 4.1.3.-	Histidine metabolism	0.831 ± 0.199
3.1.3.15	Histidine metabolism	0.831 ± 0.199
3.5.4.19	Histidine metabolism	0.831 ± 0.199

into which the reaction universe was integrated. The most likely reason for this is that absolute superessentiality was investigated in many more cellular contexts. Indeed, it is possible that the reactions listed in Table 4.9 would not be the same had all of these contexts been investigated through generation and analysis of random viable metabolic networks. Still, the results presented here claim validity across a wide range of microbial metabolisms with diverse properties and in a range of environments.

Correlation between superessentiality indices

Although the plots in Figure 4.11 revealed similarily shaped distributions for all cytoplasmic superessentiality indices, they conveyed no information about how the superessentiality index obtained for a reaction in one cellular context was related to those obtained for the same reaction in other contexts. Some correlation was revealed through the calculation of average superessentiality indices, where it was found that some reactions had large indices in most or all contexts, but a more thorough approach was taken as well. This led to the results shown in Figure 4.13, which shows pairwise linear correlations between all sets of superessentiality indices obtained for cytoplasmic reactions.

Highly significant positive correlations were found between all pairs of

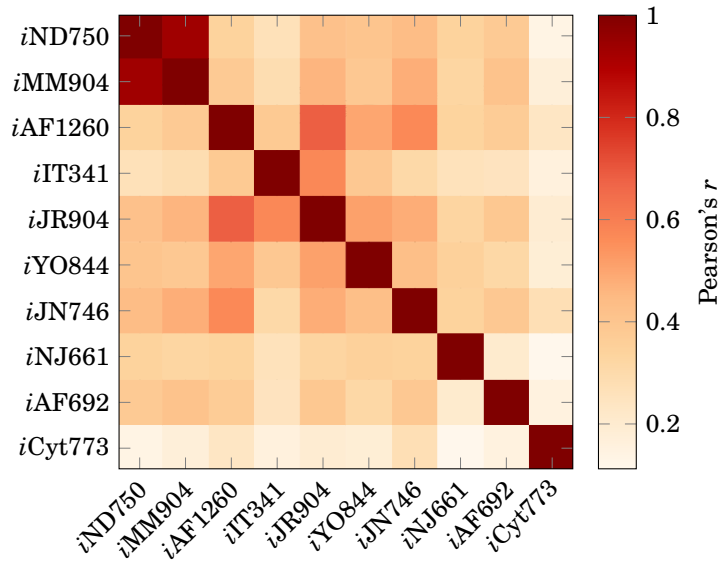


Figure 4.13: Pairwise linear correlations between all sets of cytoplasmic superessentiality indices. Pearson's r was used and the indices of 7,763 different reactions were compared. All non-diagonal correlations were highly significant ($p < 10^{-22}$). Models are clustered based on their correlation.

models, but correlation was only strong in cases where models were very similar. This indicates that the superessentiality index is quite sensitive to the cellular context in which it is calculated and thus results obtained from a single cellular context should be extrapolated to others with caution. The strongest correlation by far was observed between the two yeast models, *i*MM904 and *i*ND750. As stated before, *i*MM904 is a modified version of *i*ND750, and these two models are very similar. The second strongest correlation shown in Figure 4.13 is between *i*AF1260 and *i*JR904, two different *E. coli* reconstructions constructed by the same research group [41, 112]. The model whose superessentiality indices correlated the least with any other was, not surprisingly, *i*Cyt773, the reconstruction of the photosynthetic cyanobacterium *Cyanothece* (see previous discussion in Section 4.2.2).

Comparison to previously calculated superessentiality indices

The superessentiality index was defined by Barve, Rodrigues, and Wagner [28] who calculated it from reaction essentiality in random viable metabolic networks generated from the *i*AF1260 *E. coli* model. The superessentiality indices they found for *E. coli* reactions have been published and were com-

pared to the ones obtained from the same model in this study. These two sets will from now on be refered to as the old and the new indices, respectively. Although strong and significant positive correlation was found between the two sets of indices (Pearson's $r = 0.7539$, $p < 10^{-130}$, $n = 707$), large differences were found for many reactions. This can be seen in Figure 4.14, which shows the old indices plotted against the new.

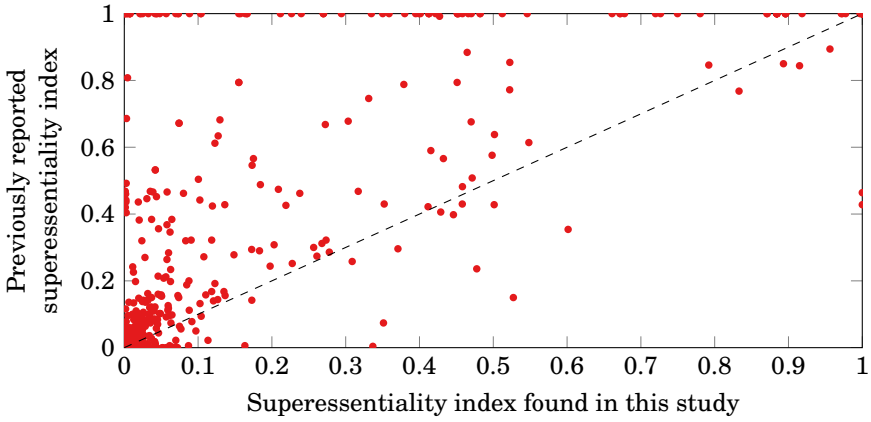


Figure 4.14: Plot showing the correlation between previously reported superessentiality indices of *E. coli* reactions and superessentiality indices calculated in this study. The previously reported values were obtained from Barve, Rodrigues, and Wagner [28]. Both sets of superessentiality indices were calculated from essentiality in random viable metabolic networks generated from the *i*AF1260 *E. coli* model, but different reaction universes were used for randomization, the sample sizes were not the same, and the randomization methods used differed slightly.

One thing that is evident from Figure 4.14 is that the reactions for which $I_{\mathrm{SE}} = 1$ was reported by Barve, Rodrigues, and Wagner were found to have a wide range of superessentiality indices in this study. In fact, the new indices of these reactions were quite evenly distributed between zero to one. Similar variation can also clearly be seen for reactions that were previously reported to be moderately superessential. For reactions with low superessentiality indices, a higher degree of consensus was generally found, and this is probably the main reason behind the strength of the observed correlation. In the pairs of superessentiality indices that were not found to be equal, the new superessentiality indices were more frequently lower than the old ones than vice versa. Specifically, 273 of the new indices were lower than the ones reported by Barve, Rodrigues, and Wagner, 411 were higher, and 23 were equal. On average, the new indices were 11 % lower than the old ones.

Three factors were nonidentical in the calculation of the old and new superessentiality indices, and there are therefore three possible explanations for the differences between indices that were discussed above. These are the sample size, the slightly modified randomization method, and the new reaction universe. The sample size was 500 for the old indices and 5,000 for the new. Barve, Rodrigues, and Wagner argued that their sample size was sufficient for good estimates of superessentiality, and the new sample size was ten times as large. This should exclude random errors due to insufficient sampling as a plausible explanation. The small modifications made to the randomization method, most importantly the randomization of other compartments than the cytoplasm, are not thought to have made a significant difference either. The reactions for which indices are shown in Figure 4.14 were all found in the cytoplasm and the only other intracellular compartment in the *i*AF1260 model was the periplasm. The reaction contents of the periplasm should not be capable of dictating the essentiality of cytoplasmic reactions to such a great degree, and it follows that the updated reaction universe is thought to be the main cause of the observed deviations.

As explained in Section 4.1.3, the reaction universe that was used in this study was much larger than the old one and included the majority of its reactions. With a larger reaction universe, superessentiality indices should mostly be expected to stay the same or drop due to the possibility of additional alternative pathways. This was observed for a large share of reactions, notably formerly absolutely superessential ones, and the superessentiality indices of investigated reactions did drop on average. A plausible explanation for the cases where the opposite was observed is the presence of reactions in the old reaction universe that were not found in the new. These reactions could potentially have formed alternative pathways that were not possible with the new reaction universe, thus lowering some of the old superessentiality indices relative to the new ones.

Extension of superessentiality

The concept of superessentiality was extended to incorporate information about synthetic lethality. An index analogous to the superessentiality index – a "super-synthetic-lethality" index, I_{SSL} – was calculated for all reactions that participated in synthetic lethal reaction pairs in random viable metabolic networks. This was done as described in Section 3.6 and the resulting indices are shown in Figure 4.15. One set of indices was calculated for each compartment in which more than ten different reactions were ever part of a synthetic lethal pair for each model from which random viable metabolic networks were generated.

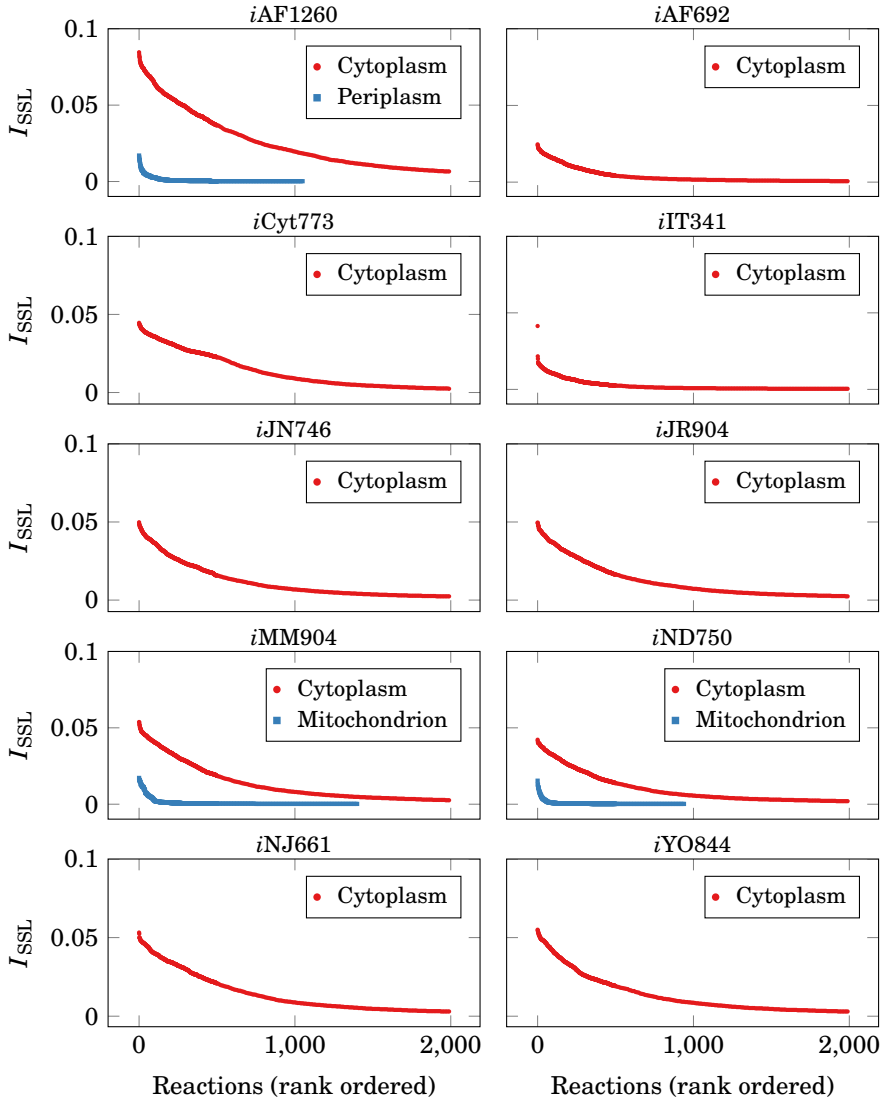


Figure 4.15: Rank plots of "super-synthetic-lethality indices", I_{SSL}, for all models from which random viable metabolic networks were generated. For each model, indices were determined from reaction synthetic lethality in 5,000 random viable metabolic networks and one index was calculated for each reaction in each intracellular compartment. Reactions are ordered by index from high to low for each compartment. Only the 2,000 reactions with largest indices are included in cases where nonzero indices were found for more than 2,000 reactions, otherwise all nonzero indices are shown.

The distributions obtained for cytoplasmic indices were all similar, gradually sloping down from a maximum value towards a long tail of low indices. The maximum value varied between 0.085 for the *i*AF1260 *E. coli* model, the largest one to be randomized, and 0.025 for the *i*AF692 reconstrution of *Methanosarcina barkeri*, the second smallest one. As discussed in Sections 4.2.1 and 4.2.2, respectively, larger fractions of synthetic lethal reaction pairs were generally identified in large networks than in small ones, and the total number of identified reactions participating in synthetic lethal pairs was positively correlated with network size. In other words, reactions should be expected to be observed more frequently in synthetic lethal pairs in random viable metabolic networks than in small ones. The same goes for compartment sizes, and it can be seen from Figure 4.15 that fewer and lower indices were obtained for other compartments than the cytoplasm. Little was concluded based on the I_{SSL} distributions other than the fact that some reactions participated in synthetic lethal reaction pairs more frequently than others.

The indices shown in Figure 4.15 are fundamentally tied to the superessentiality indices discussed in Section 4.2.3, since a reaction could only participate in a synthetic lethal pair in a random viable metabolic network if it was not essential on its own. This implies that information about how often a reaction was part of a synthetic lethal pair complements information about how often it was essential. Therefore, the indices calculated for synthetic lethal reaction pairs were combined with their corresponding superessentiality indices to produce combined indices. These indices express how frequently a reaction should be expected to be essential or part of a synthetic lethal reaction pair in different cellular contexts. However, as one might have expected from the low I_{SSL} values shown in Figure 4.15, the effect of combining these indices with superessentiality indices was very small. This can be seen from the rank plots comparing combined and superessentiality indices for cytoplasmic reactions in Appendix F. The distributions of the two indices were found to be virtually identical in all cellular contexts for all the 1,000 reactions with highest superessentiality indices, and although minor differences were discernible for reactions with lower superessentiality indices, these were very small. All in all, the combined indices added little or no information to the superessentialiy indices and did not produce insights likely to be useful in the identification of novel antimicrobial drug targets.

4.2.4 Synthetic lethality networks

The synthetic lethal reaction pairs identified in random viable metabolic networks were investigated as graphs – synthetic lethality networks. One graph

was constructed for each model from which random viable metabolic networks were created. The nodes in the graphs were metabolic reactions, two nodes were connected by an edge if their corresponding reactions ever formed a synthetic lethal pair together, and each edge was assigned a weight equal to the number of times its synthetic lethal reaction pair was observed. The number of nodes in a graph was equal to the number of different reactions participating in synthetic lethal pairs in the cellular context in question and the number of edges was equal to the number of different synthetic lethal reaction pairs identified (see Section 4.2.2).

The synthetic lethality networks were prepared for analysis by removing edges corresponding to synthetic lethal reaction pairs that were observed only once or twice. In all cases, this left a graph containing a giant component in which most nodes were found and on which analyses were performed. The number of nodes and edges in the giant components that were analyzed are listed in Table 4.10 and the giant components are visualized in Appendix G.

Table 4.10: Number of nodes and edges in the giant components of synthetic lethality networks. The numbers were obtained after removal of edges with weights lower than three.

Model name	Nodes	Edges
*i*AF1260	2,602	21,900
*i*AF692	754	2,811
*i*Cyt773	1,641	10,156
*i*IT341	494	1,500
*i*JN746	1,699	11,181
*i*JR904	1,790	10,533
*i*MM904	1,922	12,603
*i*ND750	1,585	9,699
*i*NJ661	1,998	15,345
*i*YO844	1,907	12,246

A selection of parameters describing the topology of the giant components found in synthetic lethality networks are listed in Table 4.11. Some variation can be observed due to varying graph sizes (revealed by comparison to Table 4.10), but, on the whole, there was agreement between networks. The average node degree, $\langle k \rangle$, which indicates the mean number of other reactions with which a reaction formed synthetic lethal pairs, varied between a minimum of 6.1 and a maximum of 16.8, all networks exhibited low cen-

tralization and relatively large clustering coefficients, and the characteristic path lengths ranged from 4.00 to 5.18.

Table 4.11: Simple topological parameters of synthetic lethality networks. Average node degree ($\langle k \rangle$), centralization (C_D), clustering coefficient (C), and characteristic path length (L) are shown for each model from which random viable metabolic networks were generated.

Model name	$\langle k \rangle$	C_D	C	L
*i*AF1260	16.8	0.0478	0.468	4.00
*i*AF692	7.5	0.0500	0.356	5.18
*i*Cyt773	12.4	0.0553	0.413	4.20
*i*IT341	6.1	0.0589	0.336	4.90
*i*JN746	13.2	0.0530	0.390	4.01
*i*JR904	11.8	0.0471	0.370	4.25
*i*MM904	13.1	0.0510	0.394	4.18
*i*ND750	12.2	0.0473	0.380	4.17
*i*NJ661	15.4	0.0765	0.409	4.24
*i*YO844	12.8	0.0463	0.385	4.04

The large clustering coefficients and seemingly low characteristic path lengths led to the question of whether the graphs could be considered small-world networks. To test this, the small-world test formulated by Humphries and Gurney [113] was used as described in Section 3.7.3. This method is based on comparison to a random graph, and for this and other purposes, collections of random graphs with the same number of nodes and the same node degree distributions as the synthetic lethality networks were generated (see Section 3.7.2). The random graphs were analyzed and the parameters γ, λ, and S were calculated. The obtained values are listed in Appendix H. The small-world criteria $\gamma > 1$ and $S > 1$ were found to be true for all synthetic lethality networks, indicating small-world properties.

The node degree distributions of the giant components are shown in Figure 4.16. All distributions were found to be similar, with most nodes having low degrees and a few having comparatively high ones. No good fit was found between the data and power laws. This can be seen directly from Figure 4.16, where the plots obviously do not fit well with straight lines, and was also tested more rigorously (see Appendix H). The lack of power-law fit means that the synthetic lethality networks were not scale-free. Rather, the exponentially decaying tails of their degree distributions imply that they were

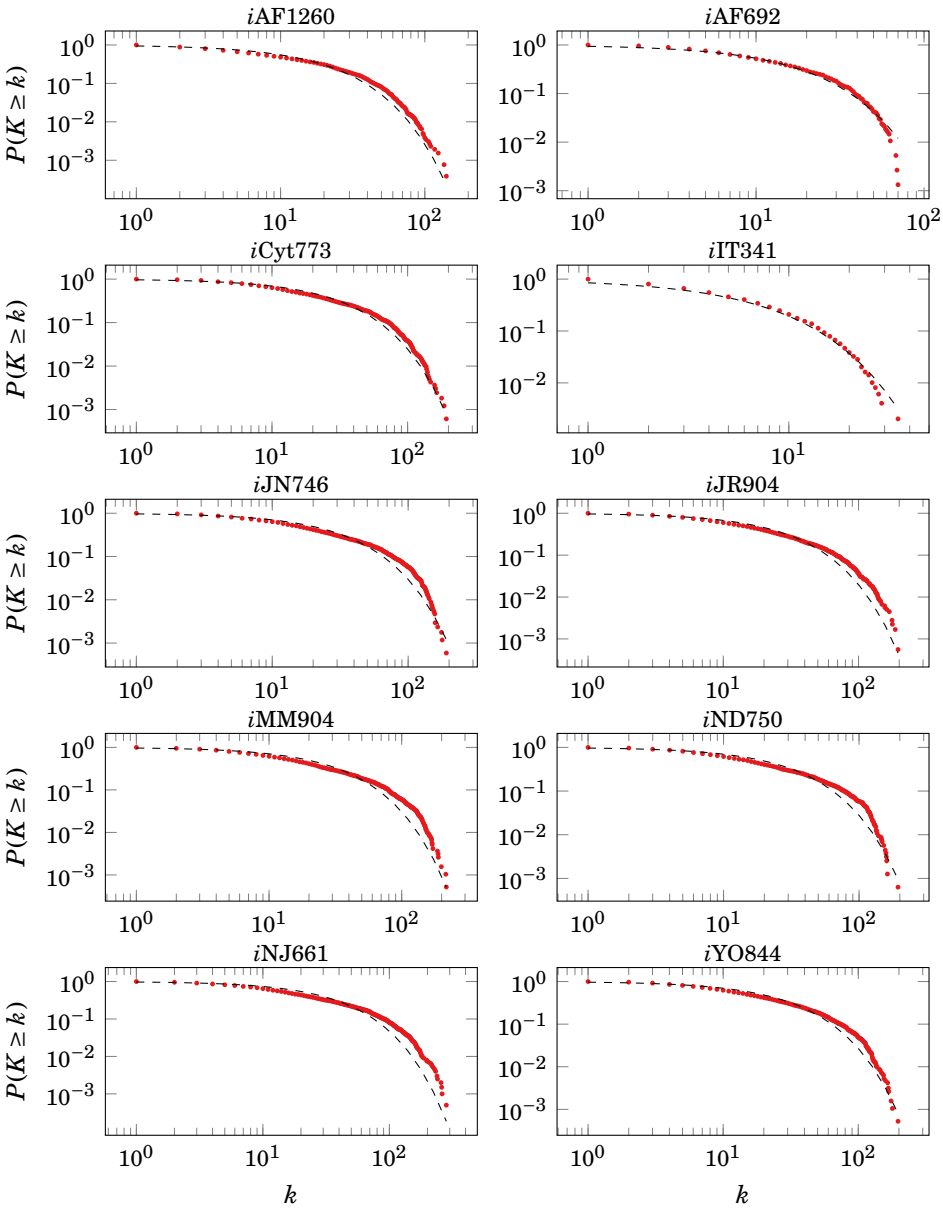


Figure 4.16: Degree distributions of synthetic lethality networks. $P(K \geq k)$ is the cumulative probability distribution of the node degree k. The fitted curves are cumulative exponential distributions whose general forms are $P(K \geq k) = e^{-k/\mu}$.

single-scale [114]. As indicated in Figure 4.16, the data were found to be closely matched by exponential distributions, the parameters and coefficients of determination of which are listed in Appendix H.

The organization of clustering and neighborhood connectivity was investigated further. Figure 4.17 shows the average clustering distributions of all synthetic lethality networks and Figure 4.18 shows the average neighborhood connectivity distributions. The parameters of the fitted lines can be found in Appendix H. Average clustering was found to decay with increasing node degree, so nodes with low degrees tended to have more densely connected neighborhoods than those with high degrees. This indicates that the networks contained clusters of highly interconnected nodes that were, to some degree, further organized as more sparsely connected units on a larger scale. For average neighborhood connectivity, increasing trends were observed in all cases. This suggest that the networks were weakly assortative in terms of node degrees, meaning that nodes with low degrees tended to be connected to other nodes with low degrees and nodes with high degrees tended to be connected to other nodes with high degrees. The significance of the clustering and neighborhood connectivity distributions was evaluated by comparison to distributions obtained from randomized networks (see Section 3.7.2). As can be seen in Figures 4.17 and 4.18, the trends observed for synthetic lethality networks deviated from those found in randomized networks.

4.2.5 Limitations of random viable metabolic networks

The following list summarizes some known and potential limitations concerning the generation and use of random viable metabolic networks:

- The networks were found to always contain large fractions of blocked reactions. These reactions will normally not be useful for analysis and also lead to a large fraction of essential reactions in the functional cores of networks.
- Small networks become less randomized than large ones.
- Network generation is computationally expensive.
- The degree to which the randomized networks are representative of real metabolic networks has not been thoroughly studied.
- The achievable degree of randomization and the quality of the randomized networks depend on the size and quality of the reaction universe. Thus, the limitations of the reaction universe listed in Section 4.1.4 apply here as well.

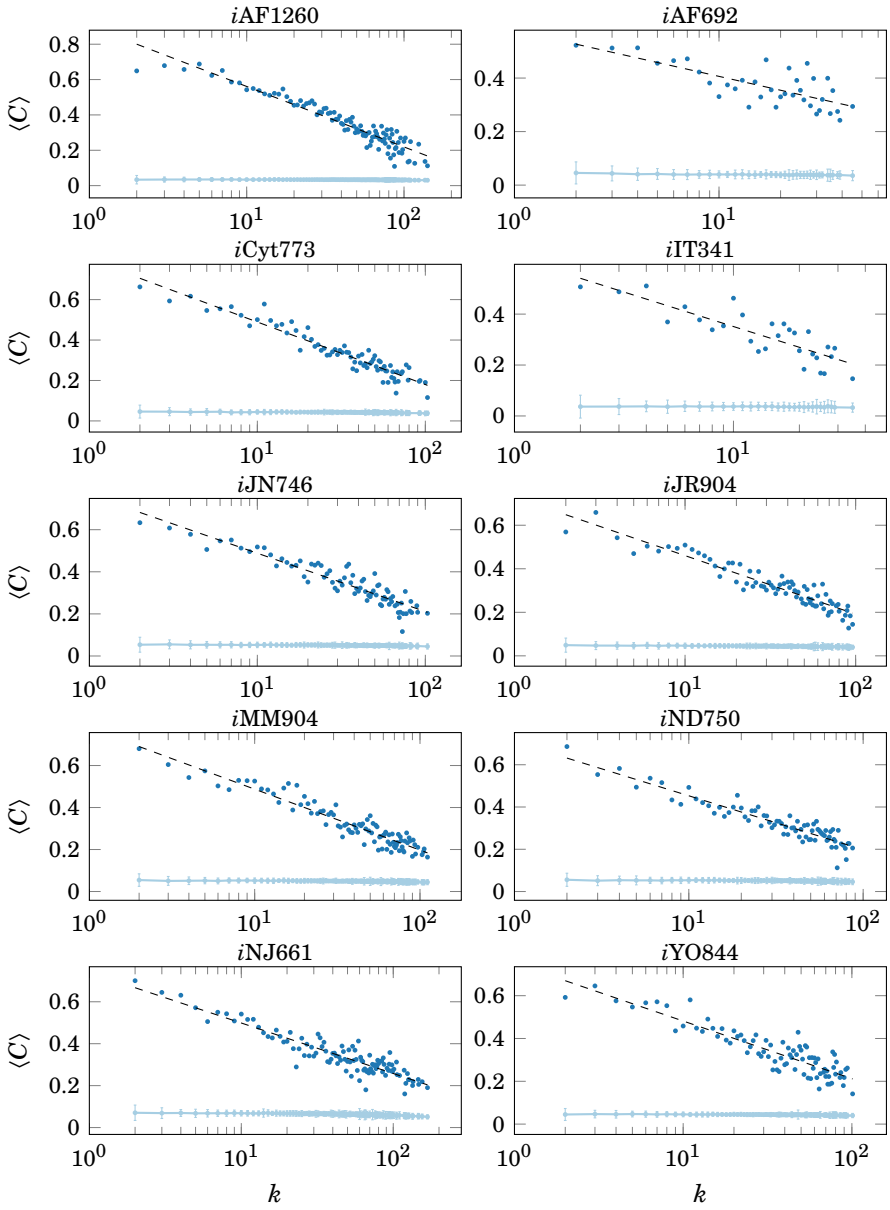


Figure 4.17: Average clustering coefficient distributions of synthetic lethality networks. The average clustering coefficient, $\langle C \rangle$, is plotted against node degree, k. The horizontal axis is logarithmic. The fitted lines have the general form $\langle C \rangle = a \log k + b$. Average data obtained from 100 randomized networks is also shown, with error bars corresponding to two standard deviations.

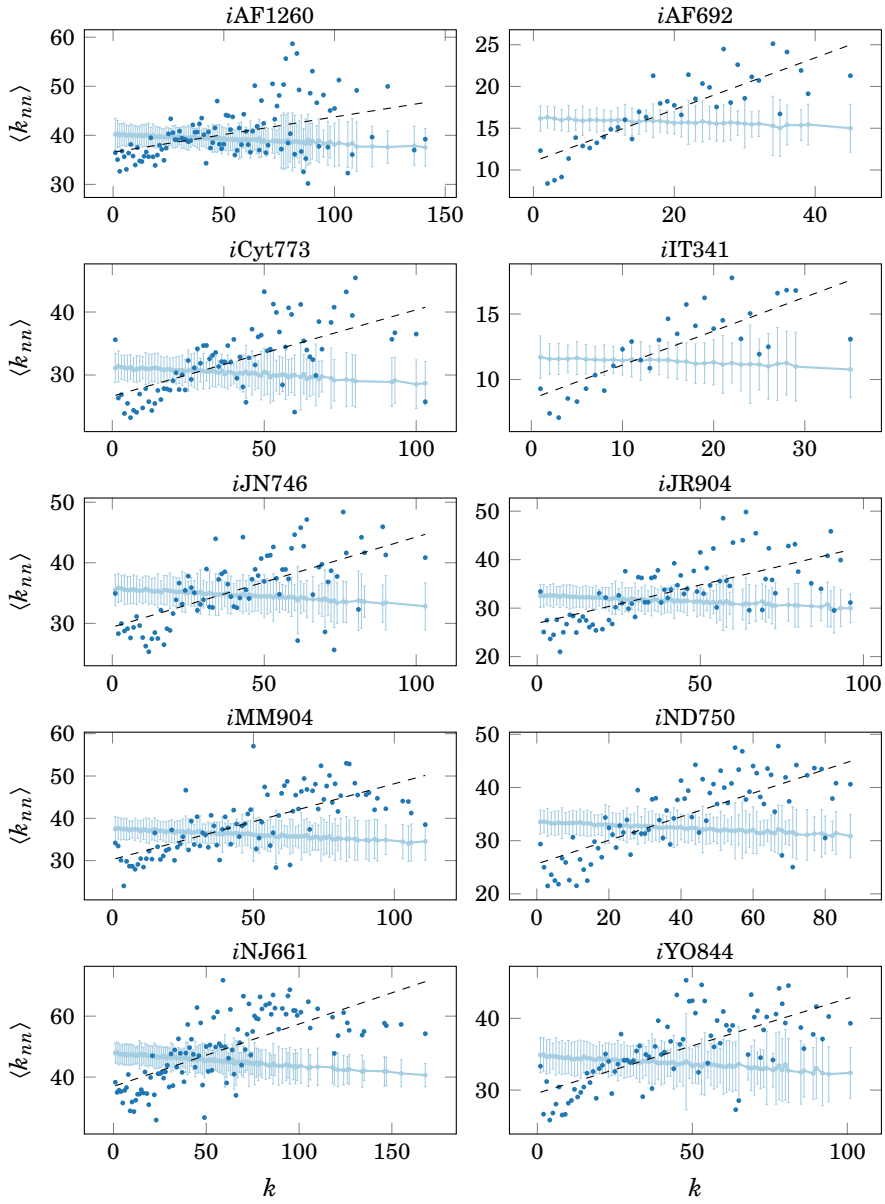


Figure 4.18: Average neighborhood connectivity distributions of synthetic lethality networks. The average neighborhood connectivity, $\langle k_{nn} \rangle$, is plotted against node degree, k. The fitted lines have the general form $\langle k_{nn} \rangle = ak + b$. Average data obtained from 100 randomized networks is also shown, with error bars corresponding to two standard deviations.

4.3 Alternative metabolic pathways of essential reactions

In Section 4.1.2 it was found that the number of essential reactions in genome-scale metabolic reconstructions was reduced upon integration with the reaction universe. This implied that the reaction universe contained alternative metabolic pathways – sets of reactions that were capable of connecting to the original networks and forming alternatives to originally essential reactions. A novel algorithm that makes use of mixed integer programming (see Section 2.1.4) was developed for the purpose of identifying such pathways. In short, it works by removing an originally essential reaction from a metabolic network, adding all reactions from the reaction universe, and then removing as many of these reactions as possible while maintaining network viability. This leaves an alternative metabolic pathway, defined as a minimal set of metabolic reactions from the reaction universe capable of replacing an essential reaction in a metabolic network. The algorithm and its implementation is presented in detail in Section 3.9.

The algorithm was applied to identify alternative metabolic pathways for all essential reactions in reconstructions of two potentially pathogenic bacteria, the *i*AF1260 and *i*NJ661 models of *E. coli* and *M. tuberculosis*, respectively. Information about these models can be found in Appendix D. It was found that a very large number of alternative pathways existed in the reaction universe for most essential reactions in both models. Essential reactions were therefore divided into two categories: those with fewer and those with more than 500 alternative pathways. The number of reactions in each category is shown in Table 4.12.

Table 4.12: Number of essential reactions for which fewer and more than 500 alternative metabolic pathways were found. The number of alternative pathways is signified by n_p.

Model name	$n_p < 500$	$n_p > 500$
*i*AF1260	57	153
*i*NJ661	32	157

For reactions that had fewer than 500 alternative metabolic pathways, all pathways were identified. For those that had more, varying numbers of pathways were identified but all were not found in any case. For this reason, and because reactions with few alternative pathways were the ones that were po-

tentially interesting as antimicrobial drug targets, only reactions with fewer than 500 alternative pathways were analyzed further. Two properties were explored for each of these reactions: the number of alternative pathways, n_p, and the average length of alternative pathways, l_p. The length of a pathway is defined as the number of reactions participating in it. Distributions of these properties are shown in the histograms in Figures 4.19 and 4.20. It can be seen that only a few alternative pathways existed in the reaction universe for most of the analyzed essential reactions. The average pathway length was also small for most reactions, the largest identified average path lengths being 6.9 and 6.8 for *i*AF1260 and *i*NJ661, respectively.

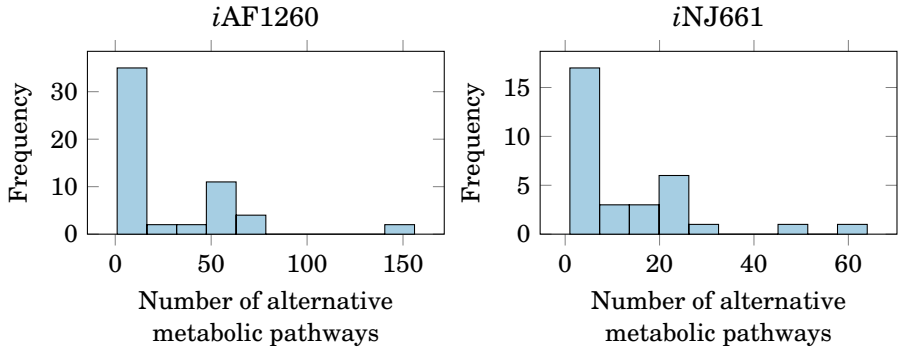


Figure 4.19: Histograms of the number of alternative metabolic pathways for essential reactions in the *i*AF1260 and *i*NJ661 models. Only reactions with fewer than 500 alternative pathways are included.

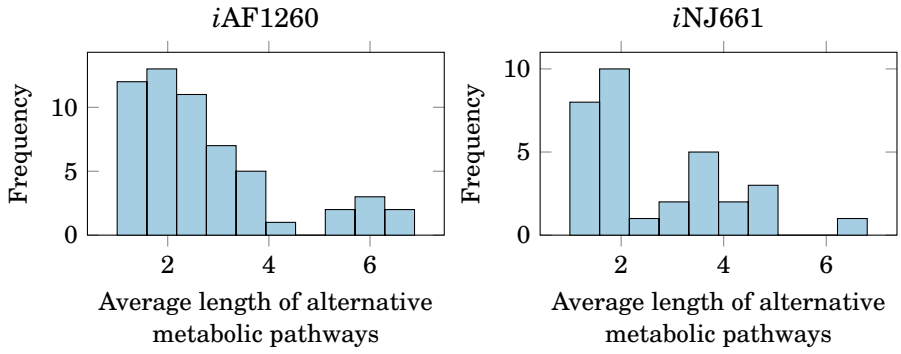


Figure 4.20: Histograms of the average length of alternative metabolic pathways for essential reactions in the *i*AF1260 and *i*NJ661 models. Only reactions with fewer than 500 alternative pathways are included.

As shown in Table 4.13, positive and significant correlation was found between n_p and l_p for both models. This implies that there was some tendency for reactions that had few alternative pathways to also have shorter average pathway lengths.

Table 4.13: Correlation between the number of alternative pathways and average pathway length for essential reactions in the *i*AF1260 and *i*NJ661 models. Only reactions with fewer than 500 alternative pathways are included. Pearson's r is given along with the p-value.

Model	r	p
*i*AF1260	0.44	$< 10^{-4}$
*i*NJ661	0.68	$< 10^{-4}$

It was suspected that n_p and l_p affected the superessentiality of reactions. Specifically, n_p was expected to be large and l_p to be small for highly superessential reactions, since a reaction with few and long alternative pathways should be harder to replace through randomization. To get an indication of whether this was indeed the case, correlation with superessentiality indices obtained from random viable metabolic networks generated from each of the two models was determined for n_p, l_p, and the ratio l_p/n_p. As can be seen from Table 4.14, it was found that I_{SE} correlated much more strongly and significantly with l_p/n_p than any of the two properties alone, indicating that a combination of the number of alternative pathways and the lengths of these pathways was indeed a key determinant of superessentiality.

Table 4.14: Correlations between superessentiality indices and properties of alternative metabolic pathways for essential reactions in the *i*AF1260 and *i*NJ661 models. Only reactions with fewer than 500 alternative pathways are included. The listed properties are the number of alternative pathways, n_p, the average length of alternative pathways, l_p, and the ratio of the first two properties, l_p/n_p.

Property	*i*AF1260		*i*NJ661	
	r	p	r	p
n_p	-0.28	0.036	-0.52	0.003
l_p	0.28	0.041	-0.27	0.130
l_p/n_p	0.57	$< 10^{-5}$	0.68	$< 10^{-4}$

Chapter 5

Conclusion

The reaction universe was found to contain 13,849 metabolic reactions, many more than what has been reported in previous studies. Topologically, it consisted mainly of a giant component with many of the same properties as smaller-scale metabolic networks. Integration of the reaction universe into microbial genome-scale metabolic reconstructions led to improved viability and robustness, indicating that the reaction universe contained sets of reactions capable of complementing and replacing reactions in the metabolic networks of microorganisms.

Hundreds of different reactions were identified as absolutely superessential, but most of these only were so in a single cellular context. No reactions were absolutely superessential in all investigated cellular contexts, meaning that no set of metabolic reactions that are always essential in any metabolism is likely to exist. Nonetheless, some specific reactions involved in peptidoglycan and riboflavin metabolism were irreplacable in more than 70 % of contexts and should be further explored as targets for novel broad-spectrum antimicrobial drugs.

The increased size of the reaction universe made it possible to randomize the reaction contents of metabolic networks to a greater degree than what has previously been achieved. However, several drawbacks of the method for metabolic network randomization were identified. Most importantly, small metabolic networks became less randomized than large ones and all randomized networks contained large fractions of blocked reactions. Future efforts should aim to improve the currently used method, specifically by reducing the fraction of blocked reactions in randomized networks and enabling equal degrees of randomization for networks of various sizes.

Many reactions were capable of being essential in all the cellular contexts that were explored through generation of random viable metabolic networks. This suggests that many essential reactions are context-general in the sense that, when they are essential, they are so due to factors that are common across organisms and environmental conditions. The same was found for reactions participating in synthetic lethal pairs. Elucidation of the causes of this context-generality could increase our understanding of general principles of metabolism.

Very similar distributions of superessentiality indices were obtained for all analyzed cellular contexts and positive correlations were found between the indices of different contexts. However, indices calculated from random viable metabolic networks generated from similar models correlated much more strongly than others. Positive correlations were also found between the superessentiality indices calculated in this study and those previously reported for *E. coli*, but much deviation was observed, primarily due to the larger reaction universe used in this study. The average superessentiality index revealed that some reactions were highly superessential regardless of context and ten of these reactions are specifically proposed as antimicrobial drug targets. The metabolic subsystems of these reactions, purine and histidine metabolism, seem to be underexplored for this purpose.

The effect of including synthetic lethality data in superessentiality indices was negligible. However, it was revealed that pairwise synthetic lethal interactions between metabolic reactions were organized in large network structures in which most reactions were located in a giant component. These synthetic lethality networks were highly clustered and single-scale and exhibited small-world properties. Evidence indicating network assortativity was also found. A potentially rewarding task for future research is the extraction of biological meaning from these networks. Specifically, the assignment of properties such as EC numbers and metabolic subsystems to nodes could reveal biologically interesting interaction patterns.

Application of the algorithm developed for identifying alternative metabolic pathways showed that more than 500 such pathways existed in the reaction universe for the majority of essential reactions in the metabolic networks of *E. coli* and *M. tuberculosis*. Reactions with fewer than 500 alternative pathways tended to have few and relatively short ones. This supports findings from analyses of superessentiality, namely that most essential reactions are easily replaced by alternatives while a few are very difficult to replace. Comparison to superessentiality indices suggested that the most important factors for reaction superessentiality were the number of alternative pathways and the lengths of these pathways.

BIBLIOGRAPHY

[1] Francis Crick. *Of Molecules and Men*. University of Washington Press, 1966.

[2] Bernhard Ø. Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, 2015.

[3] Albert-László Barabási and Zoltán N Oltvai. "Network biology: understanding the cell's functional organization." In: *Nature Reviews Genetics* 5.2 (Feb. 2004), pp. 101–13.

[4] Aarash Bordbar et al. "Constraint-based models predict metabolic and associated cellular functions." In: *Nature Reviews Genetics* 15.2 (Feb. 2014), pp. 107–20.

[5] Marc H.V. Van Regenmortel. "Reductionism and complexity in molecular biology". In: *EMBO reports* 5.11 (2004), pp. 1016–1020.

[6] Fulvio Mazzocchi. "Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory." In: *EMBO reports* 9.1 (2008), pp. 10–14.

[7] Zoltan Szallasi, Jörg Stelling, and Vipul Periwal, eds. *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*. The MIT Press, 2006.

[8] Hiroaki Kitano. "Systems biology: a brief overview". In: *Science* 295.5560 (2002), pp. 1662–1664.

[9] Denis Noble. *The Music of Life: Biology beyond genes*. OUP Oxford, 2008.

[10] Andrew R Joyce and Bernhard Ø Palsson. “The model organism as a system: integrating ’omics’ data sets.” In: *Nature Reviews Molecular Cell Biology* 7.3 (2006), pp. 198–210.

[11] Rafael U Ibarra, Jeremy S Edwards, and Bernhard O Palsson. “Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.” In: *Nature* 420 (2002), pp. 186–189.

[12] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. “Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.” In: *Nature Reviews Microbiology* 10.4 (Apr. 2012), pp. 291–305.

[13] JS Edwards and BO Palsson. “Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype”. In: *Journal of Biological Chemistry* 274.25 (June 1999), pp. 17410–17416.

[14] JS Edwards and BO Palsson. “The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities”. In: *Proceedings of the National Academy of Sciences* 97.10 (May 2000), pp. 5528–5533.

[15] AM Feist and BØ Palsson. “The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli”. In: *Nature Biotechnology* 26.6 (2008), pp. 659–667.

[16] AM Feist, MJ Herrgard, and Ines Thiele. “Reconstruction of biochemical networks in microorganisms”. In: *Nature Reviews Microbiology* 7.February (2009), pp. 129–143.

[17] Matthew A Oberhardt, Bernhard Ø Palsson, and Jason A Papin. “Applications of genome-scale metabolic reconstructions.” In: *Molecular Systems Biology* 5.320 (Jan. 2009), p. 320.

[18] Marvalee H. Wake. “Integrative Biology: Science for the 21st Century”. In: *BioScience* 58.4 (2008), p. 349.

[19] Daniel Machado et al. “Modeling formalisms in Systems Biology”. In: *AMB Express* 1.1 (Jan. 2011), p. 45.

[20] Michael Madigan et al. *Brock Biology of Microorganisms 13th Edition*. Benjamin Cummings, 2012.

[21] E Yoko Furuya and Franklin D Lowy. “Antimicrobial-resistant bacteria in the community setting.” In: *Nature Reviews Microbiology* 4.1 (2006), pp. 36–45.

[22] Nienke Van De Sande-Bruinsma et al. “Antimicrobial drug use and resistance in Europe”. In: *Emerging Infectious Diseases* 14.11 (2008), pp. 1722–1730.

[23] David J Payne et al. “Drugs for bad bugs: confronting the challenges of antibacterial discovery.” In: *Nature Reviews Drug discovery* 6.1 (2007), pp. 29–40.

[24] WHO. *Antimicrobial Resistance: Global Report on Surveillance 2014*. Tech. rep. 2014.

[25] Kim Lewis. “Platforms for antibiotic discovery.” In: *Nature Reviews Drug discovery* 12.5 (2013), pp. 371–87.

[26] Fred C Tenover. *Mechanisms of antimicrobial resistance in bacteria.* June 2006.

[27] Michael A Kohanski, Daniel J Dwyer, and James J Collins. “How antibiotics kill bacteria: from targets to networks.” In: *Nature Reviews Microbiology* 8.6 (2010), pp. 423–435.

[28] Aditya Barve, João Frederico Matias Rodrigues, and Andreas Wagner. “Superessential reactions in metabolic networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.18 (May 2012), E1121–30.

[29] Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, 2009.

[30] Jan Schellenberger et al. “Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.” In: *Nature Protocols* 6 (2011), pp. 1290–1307.

[31] Albert Lehninger, David L. Nelson, and Michael M. Cox. *Lehninger Principles of Biochemistry*. Fifth Edition. W. H. Freeman, June 2008.

[32] M Kanehisa and S Goto. “Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28 (2000), pp. 27–30.

[33] Minoru Kanehisa et al. “Data, information, knowledge and principle: Back to metabolism in KEGG”. In: *Nucleic Acids Research* 42 (2014).

[34] MW Covert et al. “Metabolic modeling of microbial strains in silico”. In: *Trends in Biochemical Sciences* (2001).

[35] Ines Thiele and Bernhard Ø Palsson. “A protocol for generating a high-quality genome-scale metabolic reconstruction.” In: *Nature Protocols* 5 (2010), pp. 93–121.

[36] Nathan D Price et al. “Genome-scale microbial in silico models: the constraints-based approach.” In: *Trends in Biotechnology* 21.4 (Apr. 2003), pp. 162–9.

[37] Lincoln Stein. “Genome annotation: from sequence to biology”. In: *Nature Reviews Genetics* 2.July (2001).

[38] Christopher S Henry et al. “High-throughput generation, optimization and analysis of genome-scale metabolic models.” In: *Nature Biotechnology* 28 (2010), pp. 977–982.

[39] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. “What is flux balance analysis?” In: *Nature Biotechnology* 28.3 (Mar. 2010), pp. 245–8.

[40] MR Andersen, ML Nielsen, and J Nielsen. “Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger.” In: *Molecular Systems Biology* 4.178 (Jan. 2008), p. 178.

[41] Adam M Feist et al. “A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information”. In: *Molecular Systems Biology* 3.121 (Jan. 2007), p. 121.

[42] NC Duarte, MJ Herrgard, and BØ Palsson. “Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model”. In: *Genome Research* 14.7 (July 2004), pp. 1298–309.

[43] Joshua A Lerman et al. “In silico method for modelling metabolism and gene product expression at genome scale.” In: *Nature Communications* 3.May (Jan. 2012), p. 929.

[44] Edward J O’Brien et al. “Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction.” In: *Molecular Systems Biology* 9.693 (Jan. 2013), p. 693.

[45] Joanne K Liu et al. “Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale.” In: *BMC Systems Biology* 8.1 (Sept. 2014), p. 110.

[46] Daniel Machado and Markus Herrgard. “Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.” In: *PLoS Computational Biology* 10.4 (Apr. 2014), e1003580.

[47] MW Covert, EM Knight, and JL Reed. “Integrating high-throughput and computational data elucidates bacterial networks”. In: *Nature* 429.May (2004), pp. 2–6.

[48] Ali Navid and Eivind Almaas. “Genome-level transcription data of Yersinia pestis analyzed with a New metabolic constraint-based approach”. In: *BMC Systems Biology* 6.1 (Jan. 2012), p. 150.

[49] Patrick F Suthers et al. “A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189.” In: *PLoS Computational Biology* 5.2 (Mar. 2009), e1000285.

[50] I Thiele et al. “Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single-and double-deletion mutants”. In: *Journal of bacteriology* (2005).

[51] Adam M Feist et al. “Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri.” In: *Molecular Systems Biology* 2 (2006).

[52] A Navid and E Almaas. “Genome-scale reconstruction of the metabolic network in Yersinia pestis, strain 91001”. In: *Molecular BioSystems* (2009).

[53] Monica L Mo, Bernhard O Palsson, and Markus J Herrgå rd. “Connecting extracellular metabolomic measurements to intracellular flux states in yeast.” In: *BMC systems biology* 3 (2009), p. 37.

[54] Cristiana Gomes de Oliveira Dal’Molin et al. “AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis.” In: *Plant Physiology* 152.2 (2010), pp. 579–589.

[55] Natalie C Duarte et al. “Global reconstruction of the human metabolic network based on genomic and bibliomic data.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104.6 (Feb. 2007), pp. 1777–82.

[56] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. “Genome-scale models of microbial cells: evaluating the consequences of constraints.” In: *Nature Reviews Microbiology* 2 (2004), pp. 886–897.

[57] Markus W Covert, Iman Famili, and Bernhard O Palsson. “Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology?” In: *Biotechnology and Bioengineering* 84.7 (Dec. 2003), pp. 763–72.

[58] Robert Schuetz, Lars Kuepfer, and Uwe Sauer. “Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.” In: *Molecular Systems Biology* 3.119 (2007), p. 119.

[59] Robert Schuetz, Nicola Zamboni, and Mattia Zampieri. “Multidimensional optimality of microbial metabolism”. In: *Science* May (2012), pp. 601–604.

[60] J Stelling et al. “Robustness of cellular functions”. In: *Cell* 118 (2004), pp. 675–685.

[61] J Arjan G M de Visser et al. “Perspective: Evolution and detection of genetic robustness.” In: *Evolution* 57.9 (2003), pp. 1959–1972.

[62] Duncan S. Callaway et al. “Network Robustness and Fragility: Percolation on Random Graphs”. In: *Physical Review Letters* 85.25 (Dec. 2000), pp. 5468–5471.

[63] R Albert, H Jeong, and AL Barabási. “Error and attack tolerance of complex networks”. In: *Nature* (2000).

[64] William G Kaelin. “The concept of synthetic lethality in the context of anticancer therapy.” In: *Nature Reviews Cancer* 5.9 (Sept. 2005), pp. 689–98.

[65] João F. Matias Rodrigues and Andreas Wagner. “Evolutionary plasticity and innovations in complex metabolic reaction networks”. In: *PLoS Computational Biology* 5 (2009).

[66] João F Matias Rodrigues and Andreas Wagner. “Genotype networks, innovation, and robustness in sulfur metabolism.” In: *BMC Systems Biology* 5.1 (2011), p. 39.

[67] Areejit Samal et al. “Genotype networks in metabolic reaction spaces.” In: *BMC Systems Biology* 4 (2010), p. 30.

[68] Andreas Wagner, Vardan Andriasyan, and Aditya Barve. “The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism”. In: *Journal of Molecular Biochemistry* 3 (2014), pp. 2–13.

[69] Andreas Wagner. “Genotype networks shed light on evolutionary constraints”. In: *Trends in Ecology and Evolution* 26.11 (Nov. 2011), pp. 577–84.

[70] Aditya Barve and Andreas Wagner. “A latent capacity for evolutionary innovation through exaptation in metabolic systems.” In: *Nature* 500 (2013), pp. 203–6.

[71] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Second edition. Springer Texts in Statistics. Springer-Verlag, New York, 2004.

[72] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[73] Jun Dong and Steve Horvath. "Understanding network concepts in modules." In: *BMC Systems Biology* 1 (2007), p. 24.

[74] Linton C. Freeman. "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3 (1978), pp. 215–239.

[75] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-law distributions in empirical data". In: *SIAM Review* 51.4 (2007), p. 43.

[76] Eivind Almaas. "Biological impacts and context of network theory." In: *The Journal of Experimental Biology* 210 (2007), pp. 1548–1558.

[77] D J Watts et al. "Collective dynamics of 'small-world' networks." In: *Nature* 393.6684 (1998), pp. 440–2.

[78] Albert-Laszlo Barabasi and Reka Albert. "Emergence of scaling in random networks". In: *Science* 310.1980 (1999), p. 11.

[79] Réka Albert and Albert Laszlo Barabasi. "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1 (2002), pp. 47–97.

[80] Erzsébet Ravasz and Albert-László Barabási. "Hierarchical organization in complex networks." In: *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 67.2 Pt 2 (2003), p. 026112.

[81] *Python.org*. URL: https://www.python.org/ (visited on 12/11/2014).

[82] Ali Ebrahim et al. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python." In: *BMC Systems Biology* 7.1 (Jan. 2013), p. 74.

[83] *The openCOBRA Project*. URL: http://opencobra.github.io/ (visited on 12/11/2014).

[84] *Documentation for COBRApy*. URL: http://cobrapy.readthedocs.org/en/latest/ (visited on 12/11/2014).

[85] *Extensible markup language (XML)*. URL: http://www.w3.org/XML/ (visited on 05/08/2015).

[86] Claudine Chaouiya et al. “SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools.” In: *BMC Systems Biology* 7 (Jan. 2013), p. 135.

[87] *SBML*. URL: http://sbml.org/ (visited on 05/08/2015).

[88] Benjamin J. Bornstein et al. “LibSBML: An API library for SBML”. In: *Bioinformatics* 24.6 (2008), pp. 880–881.

[89] *Gurobi Optimization, Inc.* 2015. URL: http://www.gurobi.com (visited on 05/08/2015).

[90] *MATLAB Documentation*. URL: http://se.mathworks.com/help/matlab/ (visited on 06/04/2015).

[91] *Graph-tool: Efficient network analysis with Python*. URL: https://graph-tool.skewed.de/ (visited on 06/04/2015).

[92] Melissa S Cline et al. “Integration of biological networks and gene expression data using Cytoscape”. In: *Nature Protocols* 2.10 (2007), pp. 2366–2382.

[93] Yassen Assenov et al. “Computing topological parameters of biological networks”. In: *Bioinformatics* 24.2 (2008), pp. 282–284.

[94] *Vilje*. URL: https://www.hpc.ntnu.no/display/hpc/Vilje (visited on 06/06/2015).

[95] Thomas Bernard et al. “Reconciliation of metabolites and biochemical reactions for metabolic networks”. In: *Briefings in Bioinformatics* 15.1 (2014), pp. 123–135.

[96] Mathias Ganter et al. “MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks”. In: *Bioinformatics* 29.6 (2013), pp. 815–816.

[97] M D Humphries, K Gurney, and T J Prescott. “The brainstem reticular formation is a small-world, not scale-free, network.” In: *Proceedings of the Royal Society B: Biological Sciences* 273.1585 (2006), pp. 503–511.

[98] *Power-law distributions*. URL: http://tuvalu.santafe.edu/~aaronc/powerlaws/ (visited on 06/07/2015).

[99] A Wagner and DA Fell. “The small world inside large metabolic networks”. In: *Proceedings of the Royal Society B: Biological Sciences* 268.1478 (Sept. 2001), pp. 1803–10.

[100] H Jeong et al. “The large-scale organization of metabolic networks”. In: *Nature* 760.1990 (2000), pp. 651–654.

[101] Mikael Huss and P Holme. “Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks.” In: *IET systems biology* 1.5 (2007), pp. 280–285.

[102] S Alexander Riemer, René Rex, and Dietmar Schomburg. “A metabolite-centric view on flux distributions in genome-scale metabolic models.” In: *BMC Systems Biology* 7 (2013), p. 33.

[103] Elaine R Lee, Kenneth F Blount, and Ronald R Breaker. “Roseoflavin is a natural antibacterial compound that binds to FMN riboswitches and regulates gene expression.” In: *RNA Biology* 6.2 (2009), pp. 187–194.

[104] Danielle B. Pedrolli et al. “The antibiotics roseoflavin and 8-demethyl-8-amino-riboflavin from Streptomyces davawensis are metabolized by human flavokinase and human FAD synthetase”. In: *Biochemical Pharmacology* 82.12 (2011), pp. 1853–1859.

[105] Simone Langer et al. “Flavoproteins are potential targets for the antibiotic roseoflavin in Escherichia coli”. In: *Journal of Bacteriology* 195.18 (2013), pp. 4037–4045.

[106] Ron Caspi et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic Acids Research* 42 (2014).

[107] Maren Lang, Michael Stelzer, and Dietmar Schomburg. “BKM-react, an integrated biochemical reaction database.” In: *BMC Biochemistry* 12.1 (2011), p. 42.

[108] Akhil Kumar, Patrick F Suthers, and Costas D Maranas. “MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases.” In: *BMC Bioinformatics* 13.1 (Jan. 2012), p. 6.

[109] Anne Morgat et al. “Updates in Rhea — a manually curated resource of biochemical reactions”. In: *Nucleic Acids Research* 43.October 2014 (2015), pp. D459–D464.

[110] Albert Bolhuis and Janice R. Aldrich-Wright. “DNA as a target for antimicrobials”. In: *Bioorganic Chemistry* 55 (2014), pp. 51–59.

[111] Masayuki Matsushita and Kim D. Janda. “Histidine kinases as targets for new antimicrobial agents”. In: *Bioorganic and Medicinal Chemistry* 10.4 (2002), pp. 855–867.

[112] Jennifer L Reed et al. “An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)”. In: *Genome Biology* 4.9 (2003), pp. 1–12.

[113] Mark D. Humphries and Kevin Gurney. “Network ’small-world-ness’: A quantitative method for determining canonical network equivalence”. In: *PLoS ONE* 3.4 (2008).

[114] LAN Amaral et al. “Classes of small-world networks.” In: *Proceedings of the National Academy of Sciences of the United States of America* 97.21 (2000), pp. 11149–11152.

[115] Rajib Saha et al. “Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803”. In: *PLoS ONE* 7.10 (2012).

[116] Juan Nogales, Bernhard Ø Palsson, and Ines Thiele. “A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory.” In: *BMC Systems Biology* 2 (2008), p. 79.

[117] Neema Jamshidi and Bernhard Ø Palsson. “Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets.” In: *BMC Systems Biology* 1 (2007), p. 26.

[118] You Kwan Oh et al. “Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data”. In: *Journal of Biological Chemistry* 282.39 (2007), pp. 28791–28799.

Appendix A

Example of a Python script using COBRApy

The following simple Python script, based on code available online [84], exemplifies the use of COBRApy. A model in SBML format is imported and a new reaction with new metabolites is added to the model. The reaction identifier and equation is printed along with the growth rate of the updated model. Finally, the updated model is exported to an SBML file.

```
from cobra import Model, Reaction, Metabolite
from cobra.io.sbml import create_cobra_model_from_sbml_file

# Import model from SBML file
cobra_model = create_cobra_model_from_sbml_file('model.xml')

# Create new reaction and set some properties
reaction = Reaction('my_new_reaction')
reaction.name = '3 oxoacyl acyl carrier protein synthase n C140'
reaction.subsystem = 'Cell Envelope Biosynthesis'
reaction.lower_bound = 0
reaction.upper_bound = 1000
reaction.objective_coefficient = 0

# Create new metabolites
ACP_c = Metabolite('ACP_c', formula='C11H21N2O7PRS',
        name='acyl-carrier-protein', compartment='c')
omrsACP_c = Metabolite('3omrsACP_c', formula='C25H45N2O9PRS',
        name='3-Oxotetradecanoyl-acyl-carrier-protein', compartment='c')
co2_c = Metabolite('co2_c', formula='CO2', name='CO2', compartment='c')
malACP_c = Metabolite('malACP_c', formula='C14H22N2O10PRS',
```

```
        name='Malonyl-acyl-carrier-protein', compartment='c')
h_c = Metabolite('h_c', formula='H', name='H', compartment='c')
# Get metabolite that already exists in model
ddcaACP_c = cobra_model.metabolites.get_by_id('ddcaACP_c')

# Add metabolites to reaction
reaction.add_metabolites({malACP_c: -1.0,
                          h_c: -1.0,
                          ddcaACP_c: -1.0,
                          co2_c: 1.0,
                          ACP_c: 1.0,
                          omrsACP_c: 1.0})

# Print the identifier and equation of the newly added reaction
print reaction.id, reaction.reaction

# Optimize and print growth rate
cobra_model.optimize(solver='gurobi')
print 'Growth rate:', cobra_model.solution.f

# Write modified model to SBML file
write_cobra_model_to_sbml_file(cobra_model, 'model.out.xml')
```

Appendix B

Parameters for metabolic network randomization

The number of reaction swaps to perform before sampling the first randomized networks and between sampling of two randomized networks when using the randomization procedure described in Chapter 3.5 was determined through testing. First, the model containing the largest number of metabolic reactions among the models from which random viable metabolic networks were to be generated, the *i*AF1260 *E. coli* model [41], was randomized 40 times, performing 50,000 reaction swaps each time and sampling networks every 1,000 swaps. The metabolic reaction contents of all sampled networks were analyzed by dividing reactions into three categories: blocked, essential, and those that were also present in the original network. Reactions in the first two categories were further classified based on whether they belonged to the third category or not. As shown in Figure 4.7, it was found that 50,000 swaps was enough for the mean number of reactions in all investigated categories to stabilize around a steady-state value and thus sufficient for randomization.

Once the number of reactions in all categories have reached their steady-state values and the first randomized network has been sampled, fewer swaps should be necessary between subsequent networks. It was decided that an appropriate value would be the number of swaps necessary to allow all metabolic reactions in a network to be candidates for swapping at least once. To get an approximation of this value, metabolic reactions were selected at random from the *i*AF1260 model until all had been selected at least once. This was repeated 100 times and the number of random reaction se-

lections performed was recorded each time. Based on the results, which are shown in the histogram in Figure B.2, 10,000 reaction swaps was chosen as an appropriate value.

It is worth noting that both of the chosen parameter values are larger than the ones that have previously been employed by for example Samal et al. [67] or Barve, Rodrigues, and Wagner [28].

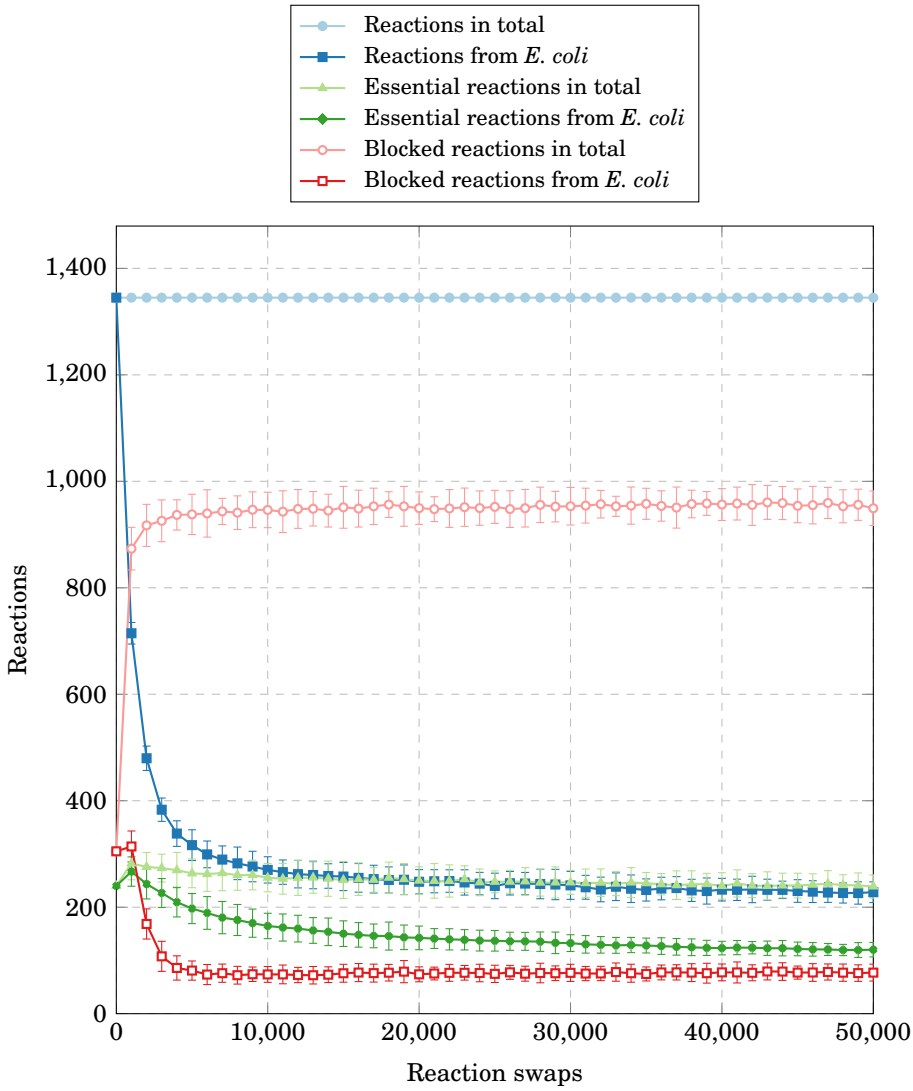


Figure B.1: Mean number of metabolic reactions in different categories plotted against the number of reaction swaps performed for 40 metabolic networks randomized from the *i*AF1260 model. The plots show the total number of metabolic reactions, the total number of reactions from *E. coli*, the total number of essential reactions, the number of essential reactions from *E. coli*, the total number of blocked reactions, and the number of blocked reactions from *E. coli*. Each data point is a mean value taken over all 40 networks. The error bars indicate two standard deviations. After 50,000 swaps, the number of reactions in all categories had stabilized around a steady-state value, indicating network randomization.

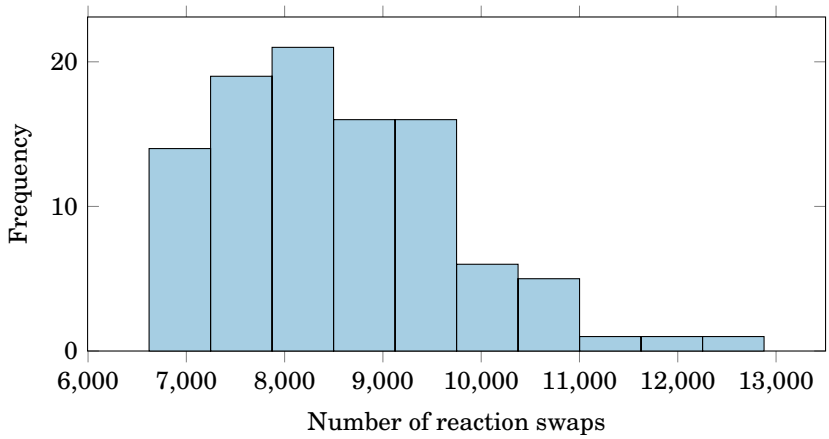


Figure B.2: Histogram of the number of reactions swaps needed for all metabolic reactions to be candidates for swapping at least once when randomizing the *i*AF1260 model. The sample size is 100.

Appendix C

Visualization of the reaction universe

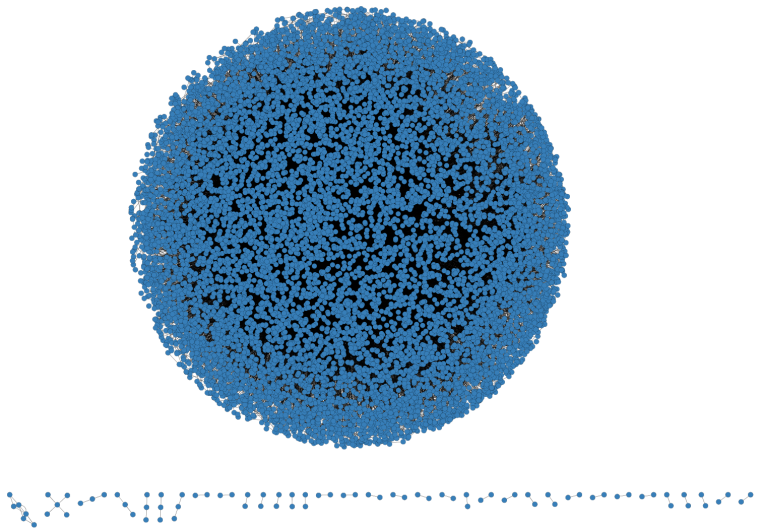


Figure C.1: Visualization of the graph representation of the reaction universe. The nodes are metabolites and two nodes are connected by an edge if they were a reactant-product pair in at least one reaction in the reaction universe. The network consists of one giant component, in which virtually all nodes are found, and 35 smaller ones containing only a few nodes each.

Appendix D

Information about models

Table D.1 lists model names, organism names, organism domains, and references for all models that were used in this study. Table D.2 lists the number of compartments, reactions, metabolic reactions, and metabolites in these models before and after merging with the reaction universe. Table D.3 gives the number of different reactions in each compartment for the multicompartment models from which random viable metabolic networks were generated. Figure D.1 gives a visualization of differences in biomass compositions and growth media between models. Exact biomass and growth media information may be obtained through MetaNetX.org [96] or the model publications that are listed in Table D.1.

Table D.1: Model names, organism names, domains, and references for all the models that were used in this study.

Model name	Organism name	Domain	Reference
*i*AF1260	*Escherichia coli*	Bacteria	[41]
*i*AF692	*Methanosarcina barkeri*	Archaea	[51]
*i*Cyt773	*Cyanothece* sp.	Bacteria	[115]
*i*IT341	*Helicobacter pylori*	Bacteria	[50]
*i*JN746	*Pseudomonas putida*	Bacteria	[116]
*i*JR904	*Escherichia coli*	Bacteria	[112]
*i*MM904	*Saccharomyces cerevisiae*	Eukaryota	[53]
*i*ND750	*Saccharomyces cerevisiae*	Eukaryota	[42]
*i*NJ661	*Mycobacterium tuberculosis*	Bacteria	[117]
*i*YO844	*Bacillus subtilis*	Bacteria	[118]
Opt158879.1	*Staphylococcus aureus*	Bacteria	[38]
Opt171101.1	*Streptococcus pneumoniae*	Bacteria	[38]
Opt208964.1	*Pseudomonas aeruginosa*	Bacteria	[38]
Opt243277.1	*Vibrio cholerae*	Bacteria	[38]
Opt71421.1	*Haemophilus influenzae*	Bacteria	[38]
Opt83332.1	*Mycobacterium tuberculosis*	Bacteria	[38]
Opt83333.1	*Escherichia coli*	Bacteria	[38]
Opt85962.1	*Helicobacter pylori*	Bacteria	[38]
Opt99287.1	*Salmonella typhimurium*	Bacteria	[38]
Seed100226.1	*Streptomyces coelicolor*	Bacteria	[38]
Seed122586.1	*Neisseria meningitidis*	Bacteria	[38]
Seed160488.1	*Pseudomonas putida*	Bacteria	[38]
Seed169963.1	*Listeria monocytogenes*	Bacteria	[38]
Seed176299.10	*Agrobacterium tumefaciens*	Bacteria	[38]
Seed177416.3	*Francisella tularensis*	Bacteria	[38]
Seed190304.1	*Fusobacterium nucleatum*	Bacteria	[38]
Seed190485.4	*Xanthomonas campestris*	Bacteria	[38]
Seed190650.1	*Caulobacter crescentus*	Bacteria	[38]
Seed192222.1	*Campylobacter jejuni*	Bacteria	[38]
Seed196627.4	*Corynebacterium glutamicum*	Bacteria	[38]
Seed211586.8	*Shewanella oneidensis*	Bacteria	[38]
Seed220668.1	*Lactobacillus plantarum*	Bacteria	[38]
Seed224324.1	*Aquifex aeolicus*	Bacteria	[38]
Seed242231.4	*Neisseria gonorrhoeae*	Bacteria	[38]
Seed243230.1	*Deinococcus radiodurans*	Bacteria	[38]
Seed243231.1	*Geobacter sulfurreducens*	Bacteria	[38]
Seed243274.1	*Thermotoga maritima*	Bacteria	[38]
Seed257313.1	*Bordetella pertussis*	Bacteria	[38]
Seed272560.3	*Burkholderia pseudomallei*	Bacteria	[38]
Seed272562.1	*Clostridium acetobutylicum*	Bacteria	[38]
Seed300852.3	*Thermus thermophilus*	Bacteria	[38]
Seed309807.19	*Salinibacter ruber*	Bacteria	[38]
Seed99287.1	*Salmonella typhimurium*	Bacteria	[38]

Table D.2: Number of compartments, reactions, metabolic reactions, and metabolites for all models before and after merging with the reaction universe. Merging leaves the number of compartments unchanged. Metabolic reactions were defined as nonboundary, nonbiomass reactions whose metabolites were all in the same compartment. Metabolites and reactions were counted multiple times if they occured in multiple compartments.

Model name	Comp.	Reactions		Metabolic reactions		Metabolites	
		Before	After	Before	After	Before	After
*i*AF1260	3	2,383	29,025	1,345	27,987	1,672	21,590
*i*AF692	2	690	14,095	528	13,933	631	10,711
*i*Cyt773	5	944	56,340	726	56,122	812	42,514
*i*IT341	2	554	14,089	392	13,927	487	10,692
*i*JN746	3	1,044	28,155	739	27,850	908	21,322
*i*JR904	2	1,075	14,320	723	13,968	764	10,754
*i*ND750	8	1,271	97,577	837	97,143	1,068	74,296
*i*MM904	8	1,569	97,817	979	97,227	1,227	74,387
*i*NJ661	2	1,022	14,220	831	14,029	824	10,739
*i*YO844	2	1,252	14,584	760	14,092	994	11,016
Opt158879.1	2	1,220	14,386	979	14,145	1,073	10,743
Opt171101.1	2	891	14,257	721	14,087	841	10,702
Opt208964.1	2	1,603	14,607	1,126	14,130	1,343	10,877
Opt243277.1	2	1,308	14,400	1,029	14,121	1,124	10,770
Opt71421.1	2	1,053	14,288	861	14,096	959	10,717
Opt83332.1	2	1,079	14,238	938	14,097	999	10,699
Opt83333.1	2	1,770	14,742	1,219	14,191	1,386	10,900
Opt85962.1	2	855	14,226	685	14,056	839	10,701
Opt99287.1	2	1,746	14,711	1,230	14,195	1,386	10,888
Seed100226.1	2	1,246	14,310	1,067	14,131	1,123	10,715
Seed122586.1	2	957	14,189	838	14,070	921	10,683
Seed160488.1	2	1,401	14,431	1,120	14,150	1,220	10,768
Seed169963.1	2	1,182	14,368	955	14,141	1,027	10,743
Seed176299.10	2	1,348	14,358	1,137	14,147	1,188	10,743
Seed177416.3	2	879	14,207	732	14,060	848	10,692
Seed190304.1	2	882	14,281	685	14,084	871	10,725
Seed190485.4	2	1,185	14,283	1,021	14,119	1,095	10,711
Seed190650.1	2	1,023	14,208	884	14,069	972	10,692
Seed192222.1	2	847	14,146	742	14,041	818	10,669
Seed196627.4	2	1,010	14,253	844	14,087	988	10,707
Seed211586.8	2	1,247	14,318	1,051	14,122	1,045	10,723
Seed220668.1	2	1,017	14,366	791	14,140	918	10,726
Seed224324.1	2	790	14,179	676	14,065	776	10,677
Seed242231.4	2	923	14,193	798	14,068	885	10,684
Seed243230.1	2	1,090	14,282	905	14,097	989	10,715
Seed243231.1	2	940	14,224	827	14,111	898	10,685
Seed243274.1	2	917	14,245	774	14,102	889	10,696
Seed257313.1	2	1,192	14,297	1,014	14,119	1,052	10,715
Seed272560.3	2	1,442	14,451	1,163	14,172	1,264	10,775
Seed272562.1	2	1,102	14,336	893	14,127	1,030	10,733
Seed300852.3	2	1,013	14,276	858	14,121	940	10,711
Seed309807.19	2	1,045	14,236	928	14,119	993	10,688
Seed99287.1	2	1,605	14,556	1,255	14,206	1,283	10,797

Table D.3: Number of metabolic reactions per randomized intracellular compartment for models that were used to generate random viable metabolic networks. The shown compartments are cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion, nucleus, periplasm, peroxisome, and vauole. Dashes indicate compartments that are not present in a model.

Model name	Number of metabolic reactions							
	Cytop.	ER	Golgi	Mitoc.	Nucleus	Perip.	Peroxis.	Vacuole
*i*AF1260	1,153	–	–	–	–	192	–	–
*i*AF692	528	–	–	–	–	–	–	–
*i*Cyt773	721	–	–	–	–	1	–	–
*i*IT341	392	–	–	–	–	–	–	–
*i*JN746	718	–	–	–	–	21	–	–
*i*JR904	723	–	–	–	–	–	–	–
*i*MM904	706	11	6	173	16	–	64	3
*i*ND750	598	5	7	152	14	–	58	3
*i*NJ661	831	–	–	–	–	–	–	–
*i*YO844	760	–	–	–	–	–	–	–

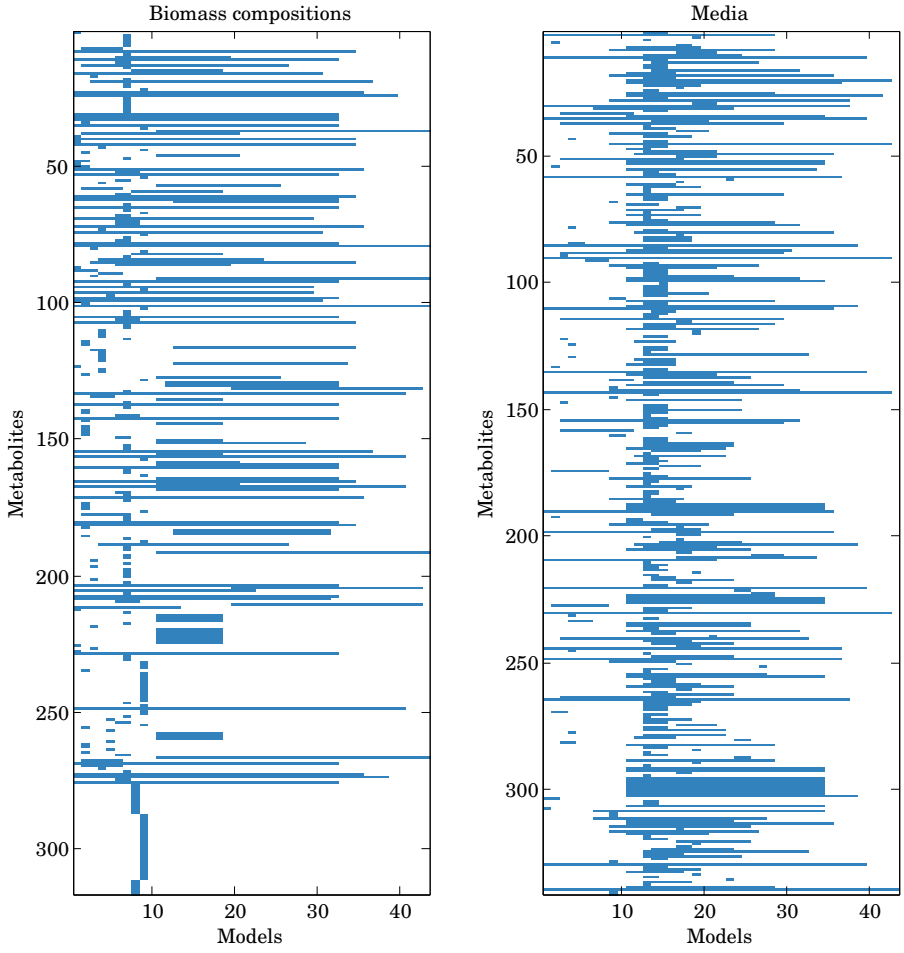


Figure D.1: Visualizations of biomass compositions and growth media for all models. Metabolites are listed along the vertical axis and models along the horizontal one. A blue mark indicates that a metabolite occured in the biomass composition or growth medium of a model. The first ten models correspond to the ten models from which random viable metabolic networks were generated.

Appendix E

Essential and synthetic lethal reactions by compartment

Tables E.1 and E.2 give the compartmental distributions of the different essential reactions and reactions participating in synthetic lethal reaction pairs, respectively, that were identified in random viable metabolic networks generated from multicompartment models. Tables E.3, E.4, E.5, and E.6 give the compartmental distributions of the different synthetic lethal reaction pairs that were identified in random viable metabolic networks generated from the multicompartment models *i*AF1260, *i*JN746, *i*MM904, and *i*ND750, respectively.

Table E.1: Compartmental distribution of the different essential reactions that were identified in random viable metabolic networks generated from multicompartment models.

Model name	Cytoplasm	ER	Mitochondrion	Nucleus	Periplasm	Peroxisome
*i*AF1260	5,108	–	–	–	1,041	–
*i*JN746	4,936	–	–	–	28	–
*i*MM904	4,855	4	2,034	7	–	10
*i*ND750	5,033	2	1,789	2	–	4

Table E.2: Compartmental distribution of the different reactions participating in synthetic lethal pairs that were identified in random viable metabolic network generated from multicompartment models.

Model name	Cytoplasm	ER	Mitochondrion	Nucleus	Periplasm	Peroxisome
*i*AF1260	5,185	–	–	–	1,053	–
*i*JN746	4,847	–	–	–	13	–
*i*MM904	4,735	9	1,402	3	–	2
*i*ND750	4,670	3	943	3	–	4

Table E.3: Compartmental distribution of the different synthetic lethal reaction pairs that were identified in random viable metabolic networks generated from the *i*AF1260 model.

	Cytoplasm	Periplasm
Cytoplasm	172,490	
Periplasm	8,212	1,104

Table E.4: Compartmental distribution of the different synthetic lethal reaction pairs that were identified in random viable metabolic networks generated from the *i*JN746 model.

	Cytoplasm	Periplasm
Cytoplasm	136,026	
Periplasm	15	4

Table E.5: Compartmental distribution of the different synthetic lethal reaction pairs that were identified in random viable metabolic networks generated from the *i*MM904 model.

	Cytoplasm	ER	Mitochondrion	Nucleus	Peroxisome	Vacuole
Cytoplasm	136,958					
ER	22	0				
Mitochondrion	8,492	0	2,576			
Nucleus	51	0	1	0		
Peroxisome	32	0	3	0	0	
Vacuole	2	0	0	0	0	0

Table E.6: Compartmental distribution of the different synthetic lethal reaction pairs that were identified in random viable metabolic networks generated from the *i*ND750 model.

	Cytoplasm	Mitochondrion	Nucleus	Peroxisome	Vacuole
Cytoplasm	128,687				
Mitochondrion	4,444	1,124			
Nucleus	4	0	0		
Peoxisome	13	3	0	0	
Vacuole	30	0	0	0	0

APPENDIX F

COMBINED INDICES

Figure F.1 shows rank plots illustrating the small differences observed between superessentiality indices and combined indices for all models from which random viable metabolic networks were generated.

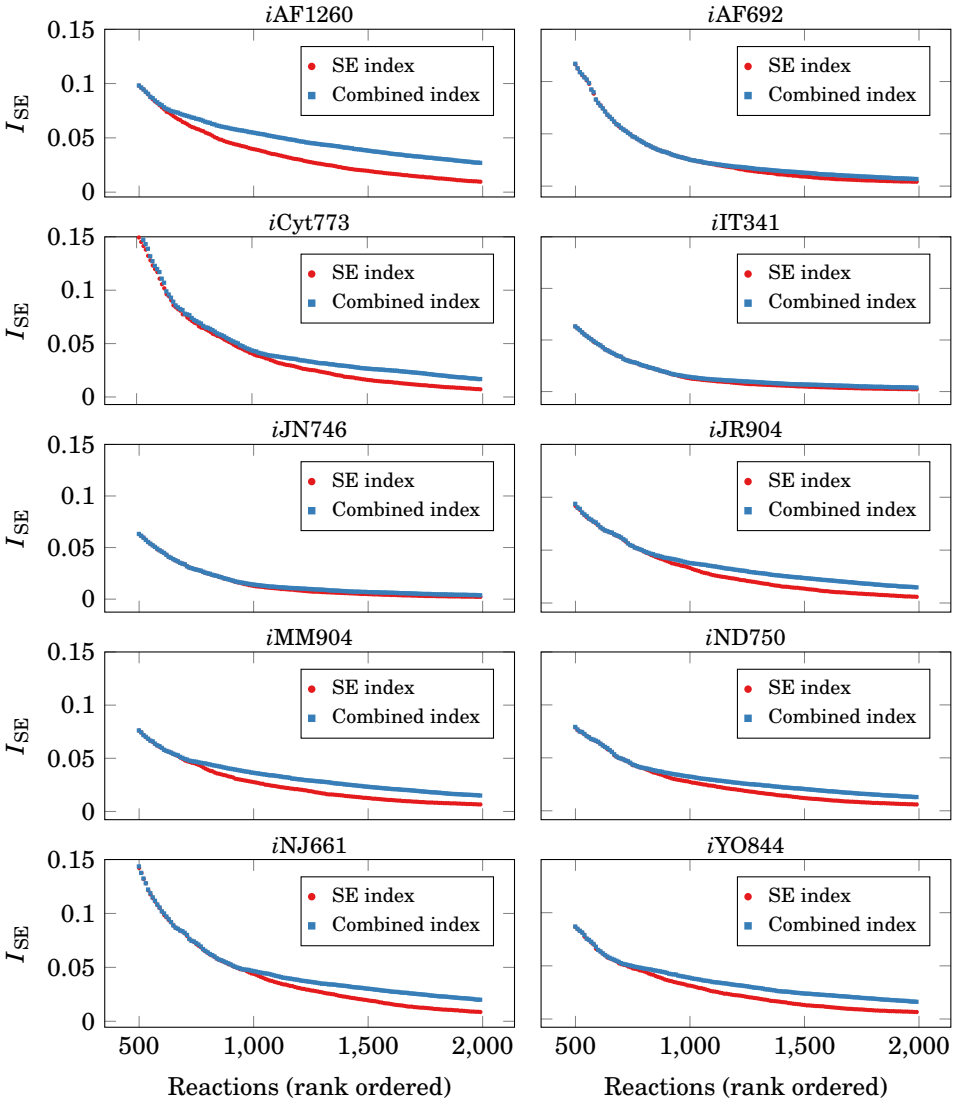
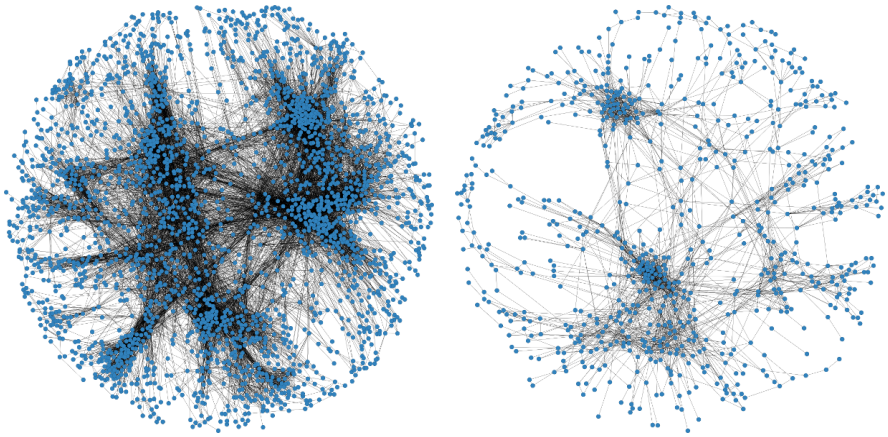


Figure F.1: Rank plots illustrating the small differences observed between super-essentiality and combined indices for all models from which random viable metabolic networks were generated. The combined index of a reaction is the sum of its super-essentiality and "super-synthetic lethality" indices. For each model, indices were determined from reaction essentiality and synthetic lethality in 5,000 random viable metabolic networks. Only indices for cytoplasmic reactions are included and only the indices of reactions ranked between 500 and 2,000 are shown. For reactions with higher indices, no difference was discernible.

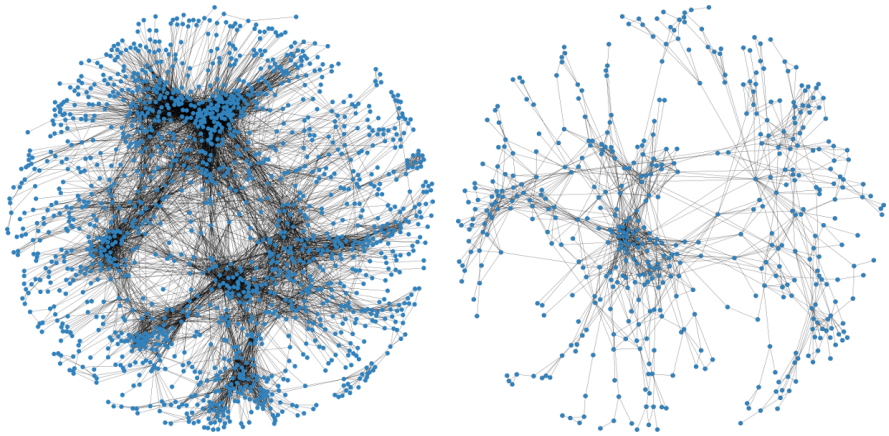
Appendix G

Visualizations of synthetic lethality networks

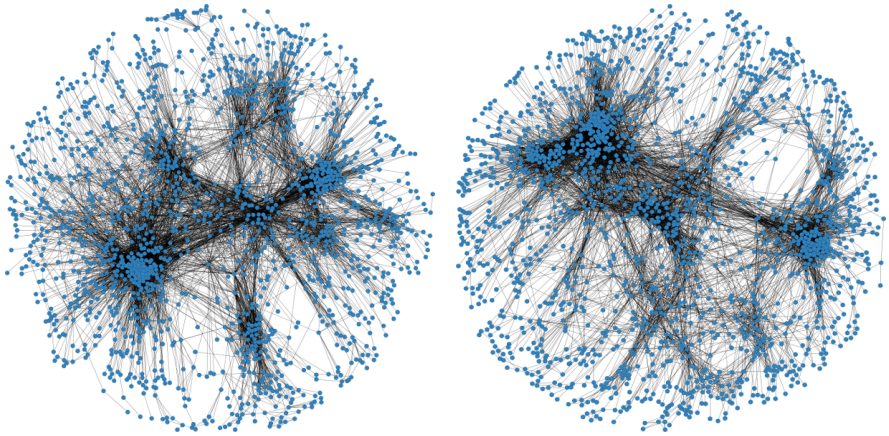


(a) *i*AF1260 (b) *i*AF692

Figure G.1: Visualizations of the synthetic lethality networks obtained from the *i*AF1260 and *i*AF692 models.

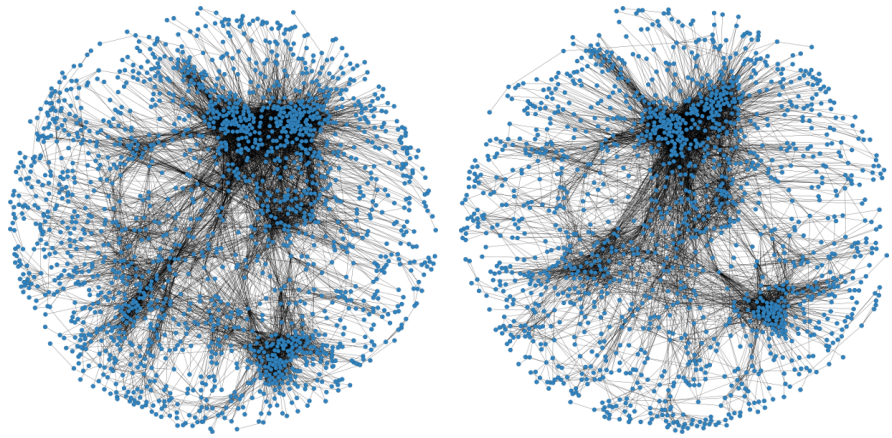


(a) *i*Cyt773 (b) *i*IT341

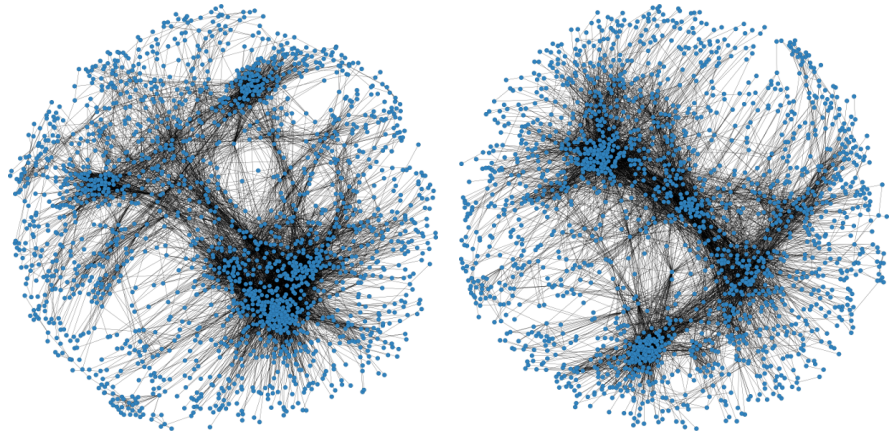


(c) *i*JN746 (d) *i*JR904

Figure G.2: Visualizations of the synthetic lethality networks obtained from the *i*Cyt773, *i*IT341, *i*JN746, and *i*JR904 models.



(a) *i*MM904 (b) *i*ND750



(c) *i*NJ661 (d) *i*YO844

Figure G.3: Visualizations of the synthetic lethality networks obtained from the iMM904, *i*ND750, *i*NJ661 and *i*YO844 models.

Appendix H

Parameters for synthetic lethality networks

Table H.1 lists the values that were obtained for the parameters that were used to determine whether the synthetic lethality networks had small-world properties. The means of the exponential distributions that were fitted to the node degree distributions are given in Table H.2 along with coefficients of determination. The slopes and intersections of the lines that were fitted to the average clustering and neighborhood connectivity distributions are given in Tables H.3 and H.4, respectively, also with coefficients of determination. Table H.5 lists maximum likelihood power law fits for the degree distributions of synthetic lethality networks

Table H.1: The parameters of the small-world criteria γ, λ, and S, for all synthetic lethality networks.

Model name	γ	λ	S
*i*AF1260	15.59	1.28	12.15
*i*AF692	10.64	1.52	7.00
*i*Cyt773	11.32	1.40	8.10
*i*IT341	9.26	1.26	7.35
*i*JN746	10.16	1.30	7.83
*i*JR904	8.96	1.30	6.90
*i*MM904	7.49	1.37	5.48
*i*ND750	10.71	1.25	8.58
*i*NJ661	11.27	1.31	8.62
*i*YO844	9.58	1.30	7.35

Table H.2: Parameters and coefficients of determination for the exponential distributions that were fitted to the degree distributions of synthetic lethality networks. The distribution means, μ, are listed along with coefficients of determination, R^2.

Model name	μ	R^2
*i*AF1260	16.83	0.989
*i*AF692	15.84	0.996
*i*Cyt773	26.97	0.989
*i*IT341	6.07	0.997
*i*JN746	28.42	0.986
*i*JR904	25.50	0.988
*i*MM904	28.66	0.984
*i*ND750	27.90	0.983
*i*NJ661	32.82	0.981
*i*YO844	27.64	0.988

Table H.3: Slopes, a, and intersections, b, of the lines that were fitted to the average clustering coefficient distributions of synthetic lethality networks. The fitted lines have the form $y = a \log x + b$.

Model name	a	b	R^2
*i*AF1260	-0.342	0.903	0.903
*i*AF692	-0.172	0.579	0.563
*i*Cyt773	-0.309	0.799	0.913
*i*IT341	-0.271	0.622	0.725
*i*JN746	-0.277	0.765	0.873
*i*JR904	-0.269	0.729	0.891
*i*MM904	-0.290	0.778	0.902
*i*ND750	-0.255	0.708	0.881
*i*NJ661	-0.241	0.740	0.857
*i*YO844	-0.270	0.751	0.813

Table H.4: Slopes, a, and intersections, b, of the lines that were fitted to the average neighborhood connectivity distributions of synthetic lethality networks. The fitted lines have the form $y = ax + b$.

Model name	a	b	R^2
*i*AF1260	0.072	36.55	0.180
*i*AF692	0.311	11.01	0.709
*i*Cyt773	0.137	26.63	0.414
*i*IT341	0.259	8.53	0.627
*i*JN746	0.149	29.37	0.425
*i*JR904	0.158	26.88	0.441
*i*MM904	0.179	30.27	0.525
*i*ND750	0.222	25.63	0.544
*i*NJ661	0.205	36.95	0.501
*i*YO844	0.132	29.57	0.483

Table H.5: Maximum likelihood power law fits for the degree distributions of synthetic lethality networks. The scaling parameter, α, and the minimum x value for power law behavior are listed along with a p value that indicates the proabibility that the distribution follows a power law. No good fits were found.

Model name	α	$x_{\min}$	p
*i*AF1260	3.50	40	0.000
*i*AF692	3.50	29	0.000
*i*Cyt773	3.50	66	0.000
*i*IT341	3.50	12	0.080
*i*JN746	3.50	67	0.000
*i*JR904	3.50	64	0.016
*i*MM904	3.50	72	0.000
*i*ND750	3.50	68	0.000
*i*NJ661	3.50	88	0.005
*i*YO844	3.50	69	0.000