The Pennsylvania State University

The Graduate School

College of Engineering

COMPUTATIONAL TOOLS FOR GENOME-SCALE SYNTHETIC LETHALITY ANALYSIS AND METABOLIC MODELING OF MICROBIAL COMMUNITIES

A Dissertation in

Chemical Engineering

by

Ali Reza Zomorrodi

© 2012 Ali Reza Zomorrodi

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2012
The dissertation of Ali Reza Zomorrodi was reviewed and approved* by the following:

Costas D. Maranas  
Donald B. Broughton Professor of Chemical Engineering  
Dissertation Adviser  
Chair of Committee

Ali Borhan  
Professor of Chemical Engineering

Howard M. Salis  
Assistant Professor of Chemical Engineering

Reka Albert  
Professor of Physics and Biology

Andrew Zydney  
Professor of Chemical Engineering  
Head of the Department of Chemical Engineering

*Signatures are on file in the Graduate School.
Abstract

Genome-scale metabolic reconstructions are increasingly becoming available for a wide range of microorganisms. Given the inherent complexity of these reconstructions on one hand and their growing influence on biological, biotechnological and biomedical research on the other hand, it is timely to develop new analysis and modeling tools to improve their predictive ability, elucidate and quantify the full range of metabolic capabilities of the underlying microbial system, and provide guidance for metabolic engineering efforts.

The central theme of this dissertation is the development and deployment of efficient mathematical modeling approaches, and in particular optimization-based algorithms, for the curation, analysis and redesign of metabolic networks of single and multi-species microbial systems. First, an efficient optimization-based procedure, namely SL Finder, is introduced for the targeted enumeration of multi-gene (and by extension multi-reaction) synthetic lethals using genome-scale metabolic models. The complete identification of all double and triple gene and reaction synthetic lethals as well as some quadruple and higher order ones using the iAF1260 metabolic model of *E. coli* uncovered complex patterns of network robustness and gene/reaction utilization and interdependence thereby providing a bird’s-eye-view of the avenues available for redirecting metabolism. Subsequently, a systematic optimization-based protocol is presented for the curation of metabolic models using multi-gene deletion (i.e., synthetic lethal) experiments. By using the existing and developed curation procedures, 90 distinct modifications for the iMM904 metabolic model of *S. cerevisiae* along with several regulatory constraints were identified and vetted using literature sources. Incorporation of the suggested modifications led to substantial improvements in the prediction accuracy of the iMM904 model for essentiality and synthetic lethality data. Next, a comprehensive flux balance analysis (FBA) framework, called OptCom, is introduced for the metabolic modeling and analysis of microbial communities. In contrast to earlier FBA approaches that are based on optimization problems with a single objective function, OptCom relies on a multi-level and multi-objective optimization formulation to properly describe trade-offs between individual vs. community level fitness criteria. The applicability of OptCom is demonstrated by modeling three different microbial communities of varying complexities to uncover the inter-species interactions, identify the optimality levels of growth for the species involved, and examine the possibility of adding a new member to an existing microbial community.
In next part, the OptForce procedure is employed as a computational strain design tool to identify the minimal set of metabolic interventions leading to overproduction of L-serine in *E. coli*. The suggested interventions include not only straightforward upregulation of the terminal pathway but also non-intuitive manipulations distant from the target product. Finally, the focus is shifted from steady state flux balance analysis to kinetic and dynamic modeling of metabolic networks using the ensemble modeling (EM) approach. Here, an optimization-based algorithm is proposed to pro-actively identify gene/enzyme perturbations that maximally reduce the number of retained models in the ensemble after each round of model screening. The applicability of this procedure is demonstrated using a metabolic model of the central metabolism of *E. coli* and by successively identifying single, double and triple enzyme perturbations (i.e., knockouts, overexpressions or combinations thereof) that cause the maximum divergent flux predictions by the models in the ensemble. Overall, the wide array of mathematical tools presented in this dissertation highlights their utility for model-driven analysis and redesign of metabolic networks.
Table of Contents

List of Figures........................................................................................................................................viii
List of Tables........................................................................................................................................x
Acknowledgements.............................................................................................................................xi

Chapter 1. Genome-scale Gene/Reaction Essentiality and Synthetic Lethality Analysis.......................... 1
  1.1. Introduction..................................................................................................................................... 1
  1.2. Methods......................................................................................................................................... 4
    1.2.1. Exhaustive enumeration of SLs............................................................................................... 5
    1.2.2. Targeted enumeration of SLs................................................................................................... 7
  1.3. Results........................................................................................................................................... 10
    1.3.1. Synthetic lethal pairs............................................................................................................... 10
    1.3.1.1. Phenotypic classification..................................................................................................... 11
    1.3.1.2. Topological classification................................................................................................... 11
    1.3.1.3. Functional classification..................................................................................................... 13
    1.3.1.4. In vivo comparisons of the predicted results........................................................................ 15
    1.3.1.5. Model refinement suggestions............................................................................................. 16
    1.3.2. Synthetic lethal triples............................................................................................................. 20
    1.3.2.1. Synthetic lethal gene triples................................................................................................ 20
    1.3.2.2. Synthetic lethal reaction triples........................................................................................... 22
    1.3.3. Higher-order synthetic lethals................................................................................................. 25
    1.3.4. Degree of essentiality............................................................................................................... 28
  1.4. Discussion and conclusion............................................................................................................ 30

Chapter 2. Improving the iMM904 S. cerevisiae Metabolic Model Using Essentiality and Synthetic Lethality Data................................................................................................................. 34
  2.1. Introduction..................................................................................................................................... 34
  2.2. Methods......................................................................................................................................... 37
    2.2.1. The modified GrowMatch procedure..................................................................................... 38
    2.2.2. Test of the suggested hypotheses........................................................................................... 41
  2.3. Results........................................................................................................................................... 42
    2.3.1. Single gene perturbations......................................................................................................... 42
2.3.1.1. Resolution of NGG inconsistencies ........................................ 42
2.3.1.2. Resolution of GNG inconsistencies ........................................ 45
2.3.2. Double gene perturbations ...................................................... 47
2.3.2.1. Resolution of GSL inconsistencies ........................................ 49
2.3.2.2. Resolution of ESSL inconsistencies ....................................... 53
2.3.2.3. Resolution of SLG inconsistencies ........................................ 55
2.3.2.4. Resolution of SLES inconsistencies ....................................... 55
2.3.3. Auxotrophy inconsistencies ..................................................... 56
2.4. Discussion and conclusion ....................................................... 58

Chapter 3. OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities ............................................. 61
3.1. Introduction .................................................................................. 61
3.2. Methods ..................................................................................... 64
3.3. Results ....................................................................................... 67
3.3.1. Modeling a mutualistic microbial community .............................. 67
3.3.1.1. Comparing the OptCom predictions with experimental results ... 70
3.3.1.2. The role of hydrogen and formate in interspecies electron transfer .......................................................... 71
3.3.2. Assessing optimality levels in a phototrophic microbial community ... 72
3.3.2.1. Analysis of the daylight metabolism ...................................... 73
3.3.3. Elucidating trophic and electron accepting interactions in sub-surface anaerobic environments ...................................................... 76
3.3.3.1. Fumarate consumption by G. sulfurreducens ............................ 77
3.3.3.2. Uncovering the inter-species metabolite transfers in the community .......................................................... 78
3.3.3.3. Addition of a new member to the microbial community .......... 81
3.4. Discussion and conclusion ....................................................... 83

Chapter 4. Metabolic Interventions for Overproduction of L-serine in Escherichia coli Using the OptForce Procedure ............................................. 86
4.1. Introduction .................................................................................. 86
4.2. Methods ..................................................................................... 87
4.3. Results............................................................................................................. 90
  4.3.1. Targeted pathway...................................................................................... 90
  4.3.2. Identified FORCE sets for overproduction of L-serine.......................... 91
  4.4. Discussion and conclusion......................................................................... 92

Chapter 5. Optimization-driven Identification of Genetic Perturbations
Accelerating the Convergence of Model Parameters in Ensemble
Modeling of Metabolic Networks............................................................................ 94
  5.1. Introduction.................................................................................................... 94
  5.2. Methods......................................................................................................... 100
    5.2.1. Ensemble modeling of metabolic networks............................................ 100
    5.2.2. Identifying the most informative enzyme perturbations...................... 103
  5.3. Results........................................................................................................... 108
    5.3.1. Criterion selection for describing the most informative
    enzyme perturbations....................................................................................... 108
    5.3.2. The metabolic model............................................................................. 109
    5.3.3. Perturbation results............................................................................... 111
      5.3.3.1. Single enzyme perturbations........................................................... 111
      5.3.3.2. Double enzyme perturbations......................................................... 120
      5.3.3.3. Triple enzyme perturbations............................................................ 121
  5.4. Discussion and conclusion.......................................................................... 122

References............................................................................................................ 123

Appendix A. Optimization formulation and solution procedure for OptCom...... 141
  A.1. Mathematical description of OptCom.......................................................... 141
  A.2. Determining the optimality levels using Descriptive OptCom................. 145

Appendix B. Details of the metabolic model used in the case study of Chapter 5... 148
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Topological and functional classification of clusters of SL gene pairs.</td>
<td>12</td>
</tr>
<tr>
<td>1.2</td>
<td>Reaction centric view and gene reaction association of cluster J in Figure  1.1.</td>
<td>15</td>
</tr>
<tr>
<td>1.3</td>
<td>Topological classification of motifs in SL gene triples.</td>
<td>21</td>
</tr>
<tr>
<td>1.4</td>
<td>Topological classification of motifs in SL reaction triples.</td>
<td>24</td>
</tr>
<tr>
<td>1.5</td>
<td>Venn diagram of the number of genes and reactions participating in SLs of order one, two and three.</td>
<td>26</td>
</tr>
<tr>
<td>1.6</td>
<td>Pictorial view of all SLs for reaction GLYCL (glycine cleavage system).</td>
<td>27</td>
</tr>
<tr>
<td>1.7</td>
<td>Percentage of degree of essentiality in different parts of the network arranged by the COG [39] functional classifications.</td>
<td>29</td>
</tr>
<tr>
<td>1.8</td>
<td>Color-coded representation of the reactions in central metabolism according to their degree of essentiality.</td>
<td>30</td>
</tr>
<tr>
<td>2.1</td>
<td>Accuracy of the iMM904 model before and after modifications using single gene perturbations.</td>
<td>43</td>
</tr>
<tr>
<td>2.2</td>
<td>Different types of mismatches between <em>in silico</em> predictions and <em>in vivo</em> observations for double gene perturbations.</td>
<td>48</td>
</tr>
<tr>
<td>2.3</td>
<td>Resolution of a GSL through reaction suppressions.</td>
<td>52</td>
</tr>
<tr>
<td>2.4</td>
<td>Change of GPR for reaction PRPPS.</td>
<td>53</td>
</tr>
<tr>
<td>2.5</td>
<td>A summary of all modifications to the iMM904 model using the single and double gene perturbations.</td>
<td>58</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic illustration of OptCom.</td>
<td>66</td>
</tr>
<tr>
<td>3.2</td>
<td>Pictorial illustration of the customized OptCom for various types of interactions.</td>
<td>68</td>
</tr>
<tr>
<td>3.3</td>
<td>Comparison of the predicted metabolic activities during the syntrophic growth with experimental data.</td>
<td>69</td>
</tr>
<tr>
<td>3.4</td>
<td>Optimality levels for the SYN and FAP guilds and their effect on the total community biomass.</td>
<td>75</td>
</tr>
</tbody>
</table>
3.5 Comparison of the predicted fluxes by OptCom with estimates in the proposed model of [177].

4.1 Flux map used to define the phenotypic space of a base strain [225].

4.2 Metabolic pathways showing the three step conversion of 3-phosphoglycerate to L-serine.

4.3 Metabolic interventions suggested by OptForce.

5.1 Comparison of variance and departure from steady state as descriptors for the most informative enzyme perturbations

5.2 Metabolic model used for simulations and the corresponding flux distribution for the wild-type strain

5.3 Optimal enzyme perturbations when pts is used as the basis to compute the variance.

5.4 Optimal enzyme perturbations when EX_succ is used as the basis to compute the variance.

5.5 Optimal enzyme perturbations when EX_r5p is used as the basis to compute the variance.

5.6 Optimal enzyme perturbations when pts and EX_succ are used as the basis to compute the variance.

5.7 Optimal enzyme perturbations when pts and EX_r5p are used as the basis to compute the variance.

5.8 Optimal enzyme perturbations when EX_succ and EX_r5p are used as the basis to compute the variance.

5.9 Optimal enzyme perturbations when pts, EX_succ and EX_r5p are used as the basis to compute the variance.

A.1 Graphical representation of the variables and equations defined in the OptCom formulation.
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Number of essential genes and genes involved in SL gene pairs for different COG [41] functional classes.</td>
<td>14</td>
</tr>
<tr>
<td>1.2</td>
<td>Direct and Indirect experimental evidence for predicted SL gene pairs and their auxotrophic characteristics.</td>
<td>17</td>
</tr>
<tr>
<td>1.3</td>
<td>Mismatches between the predicted SL pairs and experimental data for single gene knockouts.</td>
<td>19</td>
</tr>
<tr>
<td>1.4</td>
<td>Model refinements for iAF1260 suggested by SL gene pair analysis.</td>
<td>20</td>
</tr>
<tr>
<td>1.5</td>
<td>Comparison of the CPU times for finding different orders of SL reactions using the exhaustive and targeted enumeration approaches.</td>
<td>23</td>
</tr>
<tr>
<td>1.6</td>
<td>Frequency of participation of reactions in multiple SL triples.</td>
<td>25</td>
</tr>
<tr>
<td>2.1</td>
<td>Accuracy of microbial metabolic models vs iMM904 model.</td>
<td>36</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of the optimal enzyme perturbations when different sets of reactions are used as the basis to compute the variance</td>
<td>112</td>
</tr>
<tr>
<td>B.1</td>
<td>List of reactions and inhibition information</td>
<td>148</td>
</tr>
<tr>
<td>B.2</td>
<td>Expanded reaction names</td>
<td>149</td>
</tr>
<tr>
<td>B.3</td>
<td>Expanded metabolite names</td>
<td>150</td>
</tr>
</tbody>
</table>
Acknowledgements

My utmost gratitude goes to my adviser, Dr. Costas D. Maranas, for his continuous support, excellent guidance and constant encouragement from the inception of my doctoral research to completion of this dissertation. His broad and deep scientific knowledge along with his scholarly way of student advising has been exceptionally inspiring all his advisees including myself. His continuous emphasis on improving technical writing, presentation and communication skills has played a crucial role towards achieving my academic and professional career goals. Indeed, I am indebted to him more than anybody else for teaching me how to conduct high quality research, and I will carry the honor of working with a world-class researcher like him throughout the course of my professional life. I would also like to extend special thanks to Drs. Ali Borhan, Patrick Cirino, Howard Salis and Reka Albert for agreeing to serve on my doctoral committee, and in particular Dr. Salis for allowing me to conduct experiments in his lab. Many thanks go to Costas Maranas’ past and present group members, particularly Dr. Patrick Suthers (whom I was pleased to collaborate with on synthetic lethality project) and Dr. Anthony Burgard for invaluable guidance and suggestions, as well as Dr. Madhukar Dasika, Dr. Vinay Satish Kumar, Dr. Sridhar Ranganathan and Rajib Saha for insightful discussions. I would like to convey my deepest gratitude to my wife, Assieh, whose dedication, love, encouragement and persistent confidence in me rendered my graduate life smooth and manageable. It was simply not possible to get this journey through without her company. I would also like to extend my extreme gratitude to my beloved daughter, Tara, whom my wife and I are expecting in a few days. The excitement of her joining us brought indescribable happiness, joy and hope to the last year of my graduate life at Penn State. Finally, I cannot put to words the unwavering love, continuous support and prays of my parents, Fatemeh and Hossein, throughout my life and in all my pursuits. There are no two other people in the world that I can respect more.
Chapter 1

Genome-scale Gene/Reaction Essentiality and Synthetic Lethality Analysis

This chapter has been previously published in modified form in Molecular Systems Biology [1] (The first two authors contributed equally). Patrick F. Suthers served as a collaborator in this project.

1.1. Introduction

Robustness is an inherent property of metabolic networks enabling living systems to maintain their cellular functions in response to genetic and environmental perturbations [2]. The study of metabolic robustness in response to genetic perturbations is usually associated with the concepts of gene essentiality and lethality alluding to whether an organism can survive single or multiple gene deletions. Essential genes consist of genes whose individual deletion is lethal (i.e., no biomass formation) under a specific environmental condition (e.g., glucose minimal medium). By analogy, synthetic lethals (SL) refer to pairs of non-essential genes whose simultaneous deletion is lethal [3,4]. Here, we extend the concepts of essentiality and synthetic lethality to reactions whereby elimination of a single or a pair of reactions, respectively, precludes biomass formation.
Synthetic gene lethality can arise for a variety of reasons. For example, two gene protein products can be interchangeable with respect to an essential function (isozymes), act in the same essential pathway (with each mutation decreasing the flux through that pathway), or operate in two separate pathways with redundant or complementary essential functions [4,5,6]. The study of synthetic lethality plays a pivotal role in elucidating functional associations between genes and gene function predictions [7]. For example, SL screens have been used to identify new genes involved in morphogenesis [8,9], vacuolar protein transport [10], DNA damage [11], spindle migration [12] and in many other studies [13,14]. In the context of human genetics, gene lethality studies have been implicated in cancer therapies and the development of new pharmaceuticals [6,15,16,17].

The traditional method for identifying SL interactions relies on mutant screens [18], however, in recent years we have witnessed rapid progress in the development of high-throughput SL screens. In one of the first efforts, Tong et al [19], developed a genome-scale method for the construction of double mutants termed synthetic genetic array (SGA) analysis and applied it to the yeast genome. Later, Ooi et al [20] introduced a systematic technique called lethality analysis by microarray (SLAM), which takes advantage of molecular bar codes to detect lethality. Other efforts in this direction include development of an improved technology called diploid-based synthetic lethality analysis on microarrays (dSLAM) that exploits heterozygous diploid Yeast Knockouts (YKOs) to detect genome-wide lethality [21] and more recently a technique termed GIANT coli for high-throughput generation of double mutants in E. coli based on F factor-driven conjugation [22]. In spite of these advances in large-scale screening techniques, the comprehensive mapping of all SL pairs remains a labor-intensive task. For example, for the well-studied genetic system of Saccharomyces cerevisiae, even using SGA [19,23] only about 4% of the total estimated interactions under a single growth condition have been queried. This task becomes even more taxing when considering multiple growth conditions [24].

The availability of genome-scale metabolic models of organisms has provided the foundation for the development of computational frameworks to rapidly predict the effect of multiple genetic manipulations on the strain growth phenotype under different media.
For example, by applying flux balance analysis (FBA) to the iFF708 metabolic network model of *S. cerevisiae* [25], Segre and DeLuna [26] calculated the maximal rates of biomass production of all single and double gene knockouts in comparison to the wild-type strain to assess the spectrum of epistatic interactions. Plaimas *et al* [27] proposed using a machine learning strategy to distinguish between essential and non-essential reactions in *E. coli* by characterizing an enzyme based on its local network topology, gene homologies, co-expression and flux balance analysis. While most studies focused so far on *S. cerevisiae* and *E. coli* [2,24,28] there is nonetheless a number of reports studying the *in silico* lethality for other organisms. For example, the reconstructed metabolic network of *Helicobacter pylori* was used to carry out single and double mutation studies based on flux balance analysis [29]. Alternatively, Wunderlich and Mirny [30] introduced a network topological measure termed synthetic accessibility and showed that just the topology of the metabolic network of both *E. coli* and *S. cerevisiae* is sufficient to predict the viability of knockout strains with an accuracy comparable to flux balance analysis. Similarly, other studies explore metabolic network essentiality and lethality using the topological concept of missing alternatives (MA) in reaching one or more nodes in the network [31,32].

The vast majority of *in vivo* and *in silico* studies have concentrated on perturbing/deleting a single gene or a gene pair at a time. Thus, these analyses might fail to assess the full range of robustness and functional organization of the metabolic networks afforded by higher-order interactions and redundancies. Extending the concept of lethality for not just gene pairs but triples, quadruples, etc. can capture multi-gene/reaction interdependencies. The challenge in exhaustively identifying higher-order SLs lies in the combinatorial complexity of the underlying mathematical problem. Efforts towards addressing this challenge include the work of Deutscher *et al* [33] who conducted an *in silico* multiple knockout investigation of the iFF708 [25] yeast metabolic network. They cataloged gene sets that provide mutual functional backup of up to eight interacting genes. In a subsequent study, Deutscher *et al* [34] developed a computational approach based on ideas from game theory for multiple knockout analysis in *S. cerevisiae* to elucidate insights into the localization of metabolic functions. Alternatively, Behre *et al* [35] extended their previous study on single knockouts [36] by introducing a
generalized framework for analyzing structural robustness of metabolic networks based on the concept of elementary flux modes. They applied this framework to metabolic networks describing amino acid metabolism in both *E. coli* and human hepatocytes, and for the central metabolism in human erythrocytes. Yeast is the preferred system for the analysis of genetic interactions [7] due to its short non-coding regions, a genome containing less than 7% introns [37] and its existence in both haploid and diploid states. Consequently, most of research focused on investigating lethality in *S. cerevisiae*, rather than on other model microorganisms such as *E. coli*. All studies using *E. coli* are limited to only a sample of pairwise interactions.

In this chapter, we present a comprehensive map of SL gene and reaction pairs for genome-scale models. We move beyond SL pairs to exhaustively identify SL triples and some higher-order interactions among genes or reactions. Limited by the absence of customized algorithms, most existing in silico multiple knockout studies use brute-force searches [33] or focus on limited parts of metabolism [35]. We overcome these challenges by introducing a bilevel optimization framework that utilizes flux balance analysis to completely identify all multi-reaction/gene lethals for genome-scale models. This framework is applied to the iAF1260 model of *E. coli* K12 [38] for aerobic growth on minimal glucose medium. We contrast the predicted SLs against experimental data and provide a number of model refinement possibilities. We elucidate all SL gene and reaction triples. We also introduce the concept of degree of essentiality to unravel the contribution of each reaction in “buffering” cellular functionalities. This study provides a complete analysis of gene and reaction essentiality and lethality for the latest *E. coli* iAF1260 and ushers the computational means for performing similar analyses for other genome-scale models. Furthermore, by exhaustively elucidating all model growth predictions in response to multiple gene knock-outs it provides a many-fold increase in the number of genetic perturbations that can be used to assess the performance of in silico metabolic models.

### 1.2. Methods

Two separate optimization-based procedures for the enumeration of all SLs in the gene and reaction levels are described. The first one relies on the exhaustive biomass
formation capability evaluation of all single, double, triple etc. combinations of gene and/or reaction deletions. This method becomes computationally prohibitive when searching for higher-order SLs ($n > 2$). Therefore, an alternative much more efficient and targeted method, relying on bilevel optimization, is described that identifies all SL combinations without relying on the exhaustive enumeration of all possible gene/reaction eliminations. To reduce the search space in both of these approaches, a flux coupling analysis [39] was performed as a pre-processing step to allow the removal of only one representative of each fully coupled reaction set.

The analysis of synthetic lethality for the metabolic networks requires the introduction of the following sets:

$I = \{i \mid i = 1,2,\ldots,N\}$ = set of metabolites

$J = \{j \mid j = 1,2,\ldots,M\}$ = set of reactions

$K = \{k \mid k = 1,2,\ldots,G\}$ = set of genes

where, $N$, $M$ and $G$ denote the total number of metabolites, reactions and genes in the network, respectively. Upon imposing metabolite balances across the entire metabolic model under steady-state conditions we obtain:

$$\sum_{j} s_{ij}v_{j} = 0, \quad \forall \ i \in I$$  \hspace{1cm} (1.1)

where, $s_{ij}$, represents the stoichiometric coefficient of the metabolite $i$, in reaction $j$, and $v_{j}$, denotes the flux of reaction $j$. Next, we describe the exhaustive and targeted SL enumeration approaches in detail.

### 1.2.1. Exhaustive enumeration of SLs

The computational prediction of synthetic lethality hinges upon the calculation of the maximum biomass formation in the presence of the gene or reaction deletions implied by the examined synthetic lethal. We chose 1% of the maximum theoretical biomass yield as the cutoff for computationally predicted growth. We found that the prediction of *in silico* lethality was not particularly sensitive on the selected biomass formation cutoff value. For example, when considering single gene mutants only eight mutants (which is only about 0.6% of all genes in the model) involve biomass formation values between 1%-50%. We are aware of the definition of *in silico* synthetic lethality in the [33] study.
However, we believe that using a conservative universal cutoff of 1% for all mutants safeguards against the possibility of misclassifying as synthetic lethal mutants gene mutants that could be viable. In other words, we want to minimize the occurrence of false positives perhaps at the expense of missing some synthetic lethals. If $D$ is the set of reactions that is set to zero either directly or as a consequence of the GPR associations implied by the gene deletions then the problem of determining the maximum biomass formation can be formulated as the following linear program (LP):

Maximize $v_{\text{biomass}}$  \[\text{[MaxBiomass]}\]  (1.2)

Subject to:

\[\sum_{j} s_{ij} v_{j} = 0 \quad \forall \; i \in I\]  (1.1)

\[LB_{j} \leq v_{j} \leq UB_{j} \quad \forall \; j \in J\]  (1.3)

\[v_{d} = 0 \quad d \in D \subseteq J\]  (1.4)

\[v_{\text{glucose}} \geq v_{\text{uptake limit glucose}}\]  (1.5)

\[v_{\text{oxygen}} \geq v_{\text{uptake limit oxygen}}\]  (1.6)

\[v_{\text{ATPM}} = v_{\text{ATPM maintenance}}\]  (1.7)

\[v_{j} \in \mathbb{R}_{+}^{n} \quad \forall \; j \in J\]

Here, $v_{\text{biomass}}$ denotes the biomass flux while $v_{\text{uptake limit glucose}}$, $v_{\text{uptake limit oxygen}}$ and $v_{\text{ATPM maintenance}}$ denote the minimum required glucose and oxygen uptake rates and the non-growth associated ATP for maintenance, respectively. The values of the upper and lower bounds, $UB_{j}$ and $LB_{j}$, in equation (1.3) were chosen as not to exclude any physiologically relevant metabolic flux values. The upper bound for all reactions was set to 1,000. The lower bound was set to zero for irreversible reactions and to -1,000 for reversible reactions. For any external carbon containing metabolite, the maximum transport rate into the cell was set to 20 mmol gDW$^{-1}$h$^{-1}$. For the remaining source exchange fluxes the lower bound was set to -1000 mmol gDW$^{-1}$h$^{-1}$ [38]. Glucose minimal conditions were modeled by restricting the glucose uptake rate at 10 mmol gDW$^{-1}$ h$^{-1}$ and the oxygen uptake rate at 20 mmol gDW$^{-1}$ h$^{-1}$. The non-growth associated ATP maintenance was fixed at 8.39 gDW$^{-1}$h$^{-1}$ [38].
The elucidation of reaction SL using formulation MaxBiomass is straightforward as the membership in set D (the set of reaction deletions) is known a priori. In the case, of gene deletions, information gleaned from the GPR association relations needs to be encoded to elucidate the effect of gene deletions onto reaction deletions accounting for isozymes, multi-meric enzymes and combination thereof. Formulation MaxBiomass is iteratively solved to identify SL for genes or reactions involving a pre-specified number n of deletions (where n if the order of SL sought-after). This exhaustive evaluation becomes computationally intractable for higher-order (higher than two) SLs (see Table 1.5) motivating the development of the following targeted enumeration procedure.

1.2.2. Targeted enumeration of SLs

The proposed targeted enumeration procedure relies on the solution of a bilevel optimization formulation that identifies n simultaneous gene/reaction deletions suppressing biomass formation. This bilevel formulation identifies the set of n gene/reaction deletions that minimizes the maximum biomass formation potential of the network. If the minimal value of the maximum biomass is found to be below the imposed cutoff (i.e., one percent of maximum biomass) then the corresponding combination of n gene/reaction deletions forms a SL. It is important to note that this obviates the need to explore exhaustively all deletion combinations as the bilevel formulation “homes in” in only the biomass negating combinations.

The mathematical description of the bilevel formulation to determine SL reactions requires the definition of binary variable $y_j$ that encodes which reactions are deleted:

$$y_j = \begin{cases} 
0, & \text{if reaction } j \text{ is eliminated} \\
1, & \text{if reaction } j \text{ is active} 
\end{cases}, \quad \forall \ j \in J \quad (1.8)$$

This allows us to put forth the following min-max bilevel optimization formulation (SL Finder):
Formulation (SL Finder) is a min-max mixed integer linear program (MILP). The inner problem adjusts the fluxes to achieve maximum biomass production, subject to network stoichiometry, reaction deletions imposed by the outer problem and other possible growth and environmental constraints. The outer problem on the other hand, aims at finding synthetic reaction eliminations that lower the maximum biomass production below the imposed cutoff. Here we split the reversible fluxes into forward and backward reaction steps. To solve the bilevel formulation shown above the inner maximization is recast as a set of constraints by appending to the formulation the list of constraints corresponding to the dual of the inner problem and setting the primal objective function equal to the dual as introduced before in [40]:

Minimize \( v_{\text{biomass}} \)

\[
\begin{align*}
\text{Minimize} & \quad v_{\text{biomass}} \\
\text{subject to} & \quad \sum_j s_j v_j = 0 \quad \forall \ i \in I \\
& \quad v_j \leq UB_j y_j \quad \forall \ j \in J \\
& \quad v_j \geq 0 \quad \forall \ j \in J \\
& \quad \sum_j (1 - y_j) \leq n \\
& \quad y_j \in \{0,1\} \quad \forall \ j \in J
\end{align*}
\]

Minimize \( v_{\text{biomass}} \)

\[
\begin{align*}
\sum_j s_j v_j &= 0 \quad \forall \ i \in I \\
v_j &= UB_j y_j \quad \forall \ j \in J
\end{align*}
\]

\[
\begin{align*}
v_{\text{biomass}} &= \sum_j \mu_j UB_j y_j + v_{\text{ATPM}}^{\text{maintenance}} \mu_{\text{ATPM}} \\
\sum_j s_j v_j &= 0 \quad \forall \ i \in I \\
v_j &= UB_j y_j \quad \forall \ j \in J
\end{align*}
\]
\[ v_{glucose} \leq v_{uptake \ limit}^{\ glucose} \quad (1.12) \]
\[ v_{oxygen} \leq v_{uptake \ limit}^{\ oxygen} \quad (1.13) \]
\[ V_{ATPM} = v_{maintenance} \quad (1.7) \]
\[ \sum_j \lambda_j s_j + \mu_j \geq 0 \quad \forall \ j \in J - \{biomass,ATPM\} \quad (1.14) \]
\[ \sum_i \lambda_i s_i^{biomass} + \mu_{biomass} \geq 1 \quad (1.15) \]
\[ \sum_i \lambda_i s_i^{ARPM} + \mu_{ATPM} \geq 0 \quad (1.16) \]
\[ \sum_j (1 - y_j) \leq n \quad (1.17) \]
\[ v_j, \mu_j \geq 0 \quad \forall \ j \in J - \{ATPM\} \]
\[ \lambda_i, \mu_{ATPM} \in R \quad \forall \ i \in I \]
\[ y_j \in \{0,1\} \quad \forall \ j \in J \]

Here, \( \lambda_i, \mu_j \) and \( \mu_{ATPM} \) are the dual variables associated with the stoichiometric constraints (equation 1.1), inequalities in equation (1.11) and the constraint for non-growth associated ATP maintenance (equation 1.7), respectively. Equations (1.13) and (1.14) replace (1.6) and (1.8), respectively, because of the split of the reversible fluxes into forward and backward reaction steps. Note that the nonlinear term \( \mu_j y_j \) in equation (1.10) are exactly linearized as follows:

\[ \alpha_j = \mu_j y_j \quad (1.18) \]
\[ \mu_j^{\ min} y_j \leq \alpha_j \leq \mu_j^{\ max} y_j \quad (1.19) \]
\[ \mu_j - \mu_j^{\ max} (1 - y_j) \leq \alpha_j \leq \mu_j - \mu_j^{\ min} (1 - y_j) \quad (1.20) \]

where \( \mu_j^{\ min} \) and \( \mu_j^{\ max} \) are the lower and upper bounds on the dual variable \( \mu_j \).

If the optimal objective function value of the above optimization problem is less than the imposed cut-off, then the reactions for which \( y_j = 0 \), are reported as a SL of degree \( n \). All alternative SL reaction sets of size \( n \) are successively obtained by excluding the previously identified SLs using integer cuts and resolving the bilevel formulation. For example, if reactions \( j_1, j_2, \ldots, j_n \) are found to form a synthetic lethal set of size \( n \), we can exclude this solution and obtain the next one by appending the following constraint to the
formulation that ensures that at least one of the reactions forming the previously identified SL is active.

\[ y_{j_1} + y_{j_2} + ... + y_{j_n} \geq 1 \quad j_1, j_2, ..., j_n \in J \quad (1.21) \]

Note that while searching for the set of all SL reactions of a particular order \( n \), we need to preclude the removal of all reactions forming lower order SLs. This is accomplished by appending constraints of the following form to the outer problem of the bilevel optimization program for all reactions \( j_1, j_2, \ldots, j_p \) forming a SL set of size \( p \) (\( p < n \)):

\[ y_{j_1} + y_{j_2} + ... + y_{j_p} \geq 1 \quad j_1, j_2, ..., j_p \in J \quad \forall \ p < n \quad (1.22) \]

It is important to note that the elimination of certain reactions can prevent equation (1.7) from being satisfied, which precludes the identification of some SLs. Merging the required non-growth ATP for maintenance into the biomass equation resolves this problem.

The formulation (SL Finder) introduced above can be modified to find the set of all SL genes by adding a set of constraints to the outer problem describing GPR associations. To this end, a binary variable \( w_k \), representing if a gene \( k \) should be deleted is defined as follows:

\[ w_k = \begin{cases} 0, & \text{if gene } k \text{ is deleted} \\ 1, & \text{if gene } k \text{ is active} \end{cases}, \quad \forall \ k \in K \quad (1.23) \]

The impact of gene deletions on reaction eliminations through GPR relationships can be mathematically described and incorporated into the model by using appropriate equations relating the binary variables \( w_k \) and \( y_j \).

**1.3. Results**

**1.3.1. Synthetic lethal pairs**

By testing the impact of the removal of one gene at a time on the feasibility of biomass formation, we identified a total of 188 essential genes (15% of total metabolic genes) and five essential non-gene associated reactions (out of a total of 155 present in the model) for the \( E. \ coli \) iAF1260 model [38] when incorporating the list of reactions suppressed under aerobic glucose medium conditions. These results are in agreement
with the list reported by [38]. The relatively small fraction of genes that are essential alludes to the build-in robustness of *E. coli* metabolism to single-gene deletions, implying that a higher-order gene essentiality analysis is indeed needed to adequately assess metabolic network redundancy. By using the exhaustive enumeration procedure described in the methods section we identified 83 genes and 4 non-gene associated reactions involved in 86 SL pairs (~ 0.01% of total possible pairs) as shown in Figure 1.1. All these SL pairs are next analyzed in detail in terms of their phenotypic, topological and functional impact.

1.3.1.1. Phenotypic classification

The identified SL pairs are phenotypically classified into two types. The first type includes the ones that yield auxotroph strains that can be rescued through the supply of missing nutrients (i.e., amino acids or other compounds), whereas the second type includes those that lack essential functionalities that cannot be restored by adding extra components to the growth medium. Of the 86 predicted SL pairs, 53 (~ 62%) of them were found to yield auxotroph strains *in silico* that can be restored through supplementation. For example, an *E. coli* strain lacking both *asnA* (b3744) and *asnB* (b0674) can be rescued *in silico* through the supplementation of the growth medium by asn-L (L-asparagine). Note that the names and abbreviations of all reactions and metabolites follow those in iAF1260 [38]. On the other hand, disruption of *modA* (b0763) and *cysA* (b2422) results in a strain that cannot be rescued through the addition of the missing compound mobd (molybdate) as the gene disruptions eliminate MOBDabcpp (molybdate periplasm transport through ABC system). Alternatively, a double mutant strain lacking the gene pair *folA* (b0048) and *folM* (b1606) is unable to grow on supplemented medium, since it can neither produce nor uptake the precursor metabolite thf (5,6,7,8-tetrahydrofolate).

1.3.1.2. Topological classification

By representing all genes forming SL pairs as nodes connected by an edge a variety of different topological motifs emerge (see Figure 1.1). These include disjoint pairs, stars and highly connected subgraphs. Disjoint pairs are motifs representing gene pairs that either code for isozymes of an essential reaction or for two separate reactions that form
an essential reaction pair. We found a total of 19 disjoint gene pairs and four disjoint pairs containing non-gene associated reactions (see Figure 1.1). Almost all (i.e., 17) of the SLs encode for isozyme pairs. For example, both $\text{cysK}$ (b2414) and $\text{cysM}$ (b2421) enable the same reaction CYSS (cysteine synthase) that is required for cysteine anabolism.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{	extbf{Figure 1.1. Topological and functional classification of clusters of SL gene pairs.} Three types of network motifs are present: disjoint pairs, stars ($1$-connected motifs) and $k$-connected motifs (highly-connected subgraphs). Genes are color-coded in accordance to the COG \cite{41} functional categorization.}
\end{figure}

Star motifs are clusters with a single gene (i.e., hub gene) connected to all other genes. An important implication of these clusters is that unless the hub gene is present all “satellite” genes need to be functional for biomass formation feasibility. Star motifs are 1-connected graphs as biomass formation is preserved by simply retaining the functionality of a single gene (i.e., the hub gene). We identified a total of six star clusters involving 32 genes and two non-gene associated reactions organized in 30 SL pairs. For example, in cluster F (see Figure 1.1), the hub is a non-gene associated reaction $\text{FE3abcpp}$ (Fe(III) transport via the ABC system [periplasm to cytoplasm]). All members
of this cluster are directly or indirectly involved with iron III transport from the extracellular environment to the periplasm or from the periplasm to the cytoplasm.

Highly connected subgraphs, formally known as \( k \)-connected motifs with \( k > 1 \) [42], describe clusters that unlike star clusters (\( k = 1 \)) require the functionality of more than one gene for biomass formation to be feasible. We identified four such \( k \)-connected clusters that contained a total of 22 genes participating in 33 SL pairs. In all four clusters many multi-protein enzymes/isoenzymes were present. The largest cluster of this type (i.e., J) consists of 7 nodes and 14 edges with genes coding for four reactions involved in serine, glycine and folate metabolism (see Figure 1.2B). The underlying reasons for the complicated connectivity can be deduced by redrawing this cluster using reactions instead of genes (see Figure 1.2). As shown in Figure 1.2A, both GLYCL and GHMT2r form SLs with two other reactions. When this figure is expanded to show the gene-reaction associations (see Figure 1.2B), the reason for the essentiality connections between the corresponding genes becomes more clearly discernible.

1.3.1.3. Functional classification

We investigated the membership of SL gene pairs to clusters of orthologous groups (COGs) ontology [41] as illustrated using different colors in Figure 1.1. Table 1.1 lists the number of genes involved in each category. As shown in Figure 1.1, SL genes participate in diverse parts of metabolism though predominantly in amino acid, nucleotide and inorganic transport and metabolism. A comparison with the COG functional classification of essential genes (Table 1.1) reveals that a large number of essential genes are also involved in amino acid and nucleotide transport and metabolism as a consequence of the pivotal role of these pathways in contributing biomass components. However, unlike SLs, only a small portion of the essential genes are involved in inorganic ion transport and metabolism. In contrast, only a few genes in SL pairs belong to coenzyme transport and metabolism.

When analyzing the COG functional classifications, shown in Figure 1.1, a number of trends are revealed. We find that most lethal pairs involve genes that belong to the same COG. Notably, all genes in categories G (carbohydrate transport and metabolism), M (cell wall/membrane/envelope biogenesis) and L (replication, recombination and
repair) follow the pattern of intra-category lethality with no exceptions. Using the gene-reaction-protein (GPR) associations, we deduce that these gene pairs almost always encode isozymes catalyzing essential reactions. Conversely, most lethal pairs whose genes belong to different functional groups form complex clusters. It has been previously noted that two functionally distant genes can cause synthetic lethality because a gene deletion not only causes the loss of function of the primary function but also creates a cascade of compensatory cellular responses possibly affecting many pathways [12]. These inter-category connections are thus indicative of the need to bring to bear different parts of metabolism to enable the production of all biomass precursors. This is quite apparent for category C (energy production and conversion) for which all but two of the genes form inter-category gene essential pairs. Interestingly, the majority of the genes in this category form SLs with genes from category F (nucleotide transport and metabolism) alluding to the interdependence of nucleotide and energy (such as ATP and GTP) metabolism in supporting crucial aspects of metabolism.

Table 1.1. Number of essential genes and genes involved in SL gene pairs for different COG [41] functional classes.

<table>
<thead>
<tr>
<th>COG functional class</th>
<th>COG Abbreviation</th>
<th># of essential genes</th>
<th># of genes involved in SL pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid transport and metabolism</td>
<td>E</td>
<td>55</td>
<td>28</td>
</tr>
<tr>
<td>Nucleotide transport and metabolism</td>
<td>F</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>Inorganic ion transport and metabolism</td>
<td>P</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Energy production and conversion</td>
<td>C</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Cell wall/membrane/envelope biogenesis</td>
<td>M</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Carbohydrate transport and metabolism</td>
<td>G</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Coenzyme transport and metabolism</td>
<td>H</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>Replication, recombination and repair</td>
<td>L</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Lipid transport and metabolism</td>
<td>I</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Post-translational modification, protein turnover, chaperons</td>
<td>O</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
<td>Q</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>General function prediction only</td>
<td>R</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Translation, ribosomal structure and biogenesis</td>
<td>J</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Signal transduction mechanism</td>
<td>T</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Defense mechanisms</td>
<td>V</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
We searched for experimental evidence to examine the validity of the in silico predicted SL pairs. Direct experimental evidence was found in the literature (see Table 1.2) for eleven such SLs. All of these SLs could be rescued by nutrient supplementation: five with amino acids alone, five with other metabolites, and one with a combination of amino acids and other nutrients. One such auxotrophic example is the predicted SL (aroK, aroL). Lobner-Olesen and Marinus [43] reported that an E. coli strain deficient in aroK (b3390) and aroL (b0388) requires aromatic amino acid supplementation to grow. The conversion of shikimic acid to its phosphorylated derivative, shikimate 3-phosphate is an essential step in the synthesis of aromatic amino acids in E. coli and is catalyzed by the two isozymes shikimic acid kinase (SK) I and II, encoded by aroK (b3390) and aroL (b0388), respectively. In another example, the cysteine supplementation requirement of an E. coli strain lacking both cysteine synthase genes cysK (b2414) and cysM (b2421), was observed experimentally by Sairo et al. [44].

Another five of the experimentally verified SLs yielded strains that were auxotroph for compounds other than amino acids (see Table 1.2). For example, Wild et al. [45] demonstrated that alanine racamase activity in E. coli is due to two distinct genes, alr (b4053) and dadX (b1190) and found that the double alr and dadX mutant (a predicted SL) is dependent on external d-Ala for growth. Similarly, McCoy and Maurelli [46]
reported the dependence of an *E. coli* strain, deficient in both *ddlA* (b0381) and *ddlB* (b0092) on an exogenous supply of D-alanyl-D-alanine (D-Ala-D-Ala) dipeptide for growth. Finally, Zhao and Winkler [47] observed that the *tktA* (b2935) and *tktB* (b2465) double mutant, predicted to be an SL, is devoid of two transketolase isoenzymes and requires pyridoxine (vitamin B₆) as well as all aromatic amino acids and vitamins for growth.

In addition, we also uncovered five other cases for which experimental evidence indirectly supports the lethality of the identified SL pairs (see Table 1.2). An example of this type is the predicted SL involving *ndk* (b2518) and *adk* (b0474). These two genes code for the nucleoside diphosphate kinase (Ndk) and adenylate kinase (Adk) activities, respectively, that catalyze two reactions involved in ADP synthesis [48]. It has been reported that the presence of Adk alone is able to restore the normal growth rates of mutant strains of *E. coli* lacking Ndk [48], implying that the simultaneous disruption of both *adk* and *ndk* would be lethal for *E. coli*. Similarly, it has been shown that MalY, encoded by *malY* (b1622) is able to compensate for the methionine requirement of *metC* mutants for growth [49], in agreement with the lethality of disrupting both of these two genes. Interestingly, all but one of the SLs that can be rescued by supplementation form disjoint pairs (see Table 1.2 and Figure 1.1). One possible reason for this is that disjoint pairs (unlike stars and k-connected motifs) tend to correspond to isozymes, which are much more likely to have been experimentally characterized. Overall, the presence of direct and indirect experimental evidence for some of the predicted SLs alludes to the reliability of the *iAF1260* model and SL predictions.

### 1.3.1.5. Model refinement suggestions

Comparisons of *in silico* predictions and *in vivo* observations for single gene essentiality data [50] were used before to drive the process of metabolic model refinement [51]. We believe that extending this workflow to include SL pairs, triplets, etc. will provide additional layers of model validation and opportunities for correction. We have already identified 27 *in silico* SLs that are inconsistent with *in vivo* SL data. They fall into two different groups. The first one includes predicted SLs that contain one or more essential genes whereas the latter contains predicted SL that are in agreement
with *in vivo* SL data but imply incorrect supplementation rescue (i.e., auxotrophy) scenarios.

### Table 1.2. Direct and Indirect experimental evidence for predicted SL gene pairs and their auxotrophic characteristics.

<table>
<thead>
<tr>
<th>SL gene pair</th>
<th>Topology</th>
<th>Experimental evidence</th>
<th>Growth supplementation of mutant strain</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ddlA</em> (b0381), <em>ddlB</em> (b0092)</td>
<td>Disjoint pair</td>
<td>[46]</td>
<td>(\text{D-alanyl-D-alanine (D-Ala-D-Ala) dipeptide})</td>
</tr>
<tr>
<td><em>dadX</em> (b1190), <em>alr</em> (b4053)</td>
<td>Disjoint pair</td>
<td>[45]</td>
<td>D-Ala</td>
</tr>
<tr>
<td><em>gutQ</em> (b2708), <em>yrbH</em> (b3197)</td>
<td>Disjoint pair</td>
<td>[52]</td>
<td>D-arabinose 5-phosphate</td>
</tr>
<tr>
<td><em>acnA</em> (b1276), <em>acnB</em> (b0118)</td>
<td>Disjoint pair</td>
<td>[53]</td>
<td>Glutamate</td>
</tr>
<tr>
<td><em>tktA</em> (b2935), <em>tktB</em> (b2465)</td>
<td>Disjoint pair</td>
<td>[47]</td>
<td>Aromatic amino acids and Vitamin B6</td>
</tr>
<tr>
<td><em>cynT</em> (b0339), <em>YadF</em> (b0126)</td>
<td>Disjoint pair</td>
<td>[54]</td>
<td>High CO₂ concentration</td>
</tr>
<tr>
<td><em>metE</em> (b3829), <em>metH</em> (b4019)</td>
<td>Disjoint pair</td>
<td>[55] and [56]</td>
<td>Methionine</td>
</tr>
<tr>
<td><em>cysK</em> (b2414), <em>cysM</em> (b2421)</td>
<td>Disjoint pair</td>
<td>[44]</td>
<td>Cysteine</td>
</tr>
<tr>
<td><em>argF</em> (b0273), <em>argI</em> (b4254)</td>
<td>Disjoint pair</td>
<td>[57]</td>
<td>Arginine</td>
</tr>
<tr>
<td><em>aroL</em> (b0388), <em>aroK</em> (b3390)</td>
<td>Disjoint pair</td>
<td>[43]</td>
<td>Aromatic amino acids</td>
</tr>
<tr>
<td><em>purT</em> (b1849), <em>purN</em> (b2500)</td>
<td>Cluster A</td>
<td>[58]</td>
<td>Purine</td>
</tr>
<tr>
<td><em>ubiX</em> (b2311), <em>ubiD</em> (b3843)</td>
<td>Disjoint pair</td>
<td>[59] and [60]</td>
<td>-</td>
</tr>
<tr>
<td><em>metC</em> (b3008), <em>malY</em> (b1622)</td>
<td>Disjoint pair</td>
<td>[49]</td>
<td>-</td>
</tr>
<tr>
<td><em>aroE</em> (b1692), <em>YdiB</em> (b3281)</td>
<td>Disjoint pair</td>
<td>[61] and [62]</td>
<td>-</td>
</tr>
<tr>
<td><em>adk</em> (b0474), <em>ndk</em> (b2518)</td>
<td>Cluster F</td>
<td>[48]</td>
<td>-</td>
</tr>
<tr>
<td><em>nrdA/2234/5 nrdE/F</em> (b2675/6)</td>
<td>Cluster I</td>
<td>[63] and [64]</td>
<td>-</td>
</tr>
</tbody>
</table>
The majority of the inconsistent SL predictions (i.e., first group) involve at least one member reported as essential in vivo [38,65,66]. As indicated in Table 1.3, there are 25 essential genes involved in 44 of the predicted SL pairs. Kumar and Maranas [50] recently showed that three of these essential genes (see Table 1.3) are most likely misclassified as alluded by their marginal essentiality scores. Of the remaining 22 genes, we find that six of them form SL pairs with genes that are not expressed under aerobic glucose minimal conditions [67]. Therefore, the essentiality prediction for these six genes in vivo can be recapitulated by appending appropriate regulatory constraints to the model that restricts gene expression for seven genes under aerobic glucose conditions (see Table 1.4). In support of this, knockout mutants for some of these six genes have been rescued through overexpressing their SL partner(s) in E. coli. For example, there is genetic evidence regarding complementation of nrdA (b2334) or nrdB (b2335) mutants of E. coli with nrdE (b2675) or nrdF (b2676) overexpressed on a plasmid [63,64]. In another example, ydiB (b1692) and aroE (b3281) encode YdiB and its paralog AroE, respectively, which are members of the quinate/shikimate 5-dehydrogenase family that functions in the essential shikimate pathway [61]. This relationship implies that only the simultaneous disruption of both of these genes would be lethal. However, aroE (b3281) is reported to be essential (Table 1.3). This outcome may arise from inability of the low specific activity of YdiB to compensate for the deletion of AroE unless amplified, as reported by Michel et al [62]. We note that no regulation rules were introduced in [67] for nrdE, nrdF or ydiB.

Four of the 22 genes that are reported to be essential (Table 1.3), (i.e., pdxH (b1638), ubiB (b3835), adk (b0474), prsA (b1207)) form lethal pairs with four non-gene associated reactions. In addition, ubiB, adk, and prsA are always essential under rich medium conditions, whereas pdxH is essential in glucose minimal medium but not in rich medium. These results imply that the model is missing regulatory restrictions for the non-gene associated reactions listed above. We suggest that OPHHX3, R1PK and PDX5PO2 should be suppressed under the examined experimental conditions (Table 1.4). In particular, the non-gene associated reactions OPHHX3 and R1PK are most likely inactive under aerobic conditions [68,69], whereas PDX5PO2 is likely to be inactive under
Table 1.3. Mismatches between the predicted SL pairs and experimental data for single gene knockouts.

<table>
<thead>
<tr>
<th>Gene (Blattner no.)</th>
<th>Topology</th>
<th>Experimental condition for which is essential</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyrH (b0171)</td>
<td>Disjoint pair</td>
<td>Always</td>
</tr>
<tr>
<td>ubiD (b3843)</td>
<td>Disjoint pair</td>
<td>Always</td>
</tr>
<tr>
<td>pdxH (b1638)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>ubiB (b3835)</td>
<td>Disjoint pair</td>
<td>Always</td>
</tr>
<tr>
<td>folA (b0048)</td>
<td>Disjoint pair</td>
<td>Always</td>
</tr>
<tr>
<td>yadF (b0126)</td>
<td>Disjoint pair</td>
<td>Always</td>
</tr>
<tr>
<td>metE (b3829)</td>
<td>Disjoint pair</td>
<td>Glucose</td>
</tr>
<tr>
<td>metL (b3940)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>thrA (b0002)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>metC (b3008)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>aroE (b3281)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>glnA (b3870)</td>
<td>Disjoint pair</td>
<td>Shared</td>
</tr>
<tr>
<td>eno (b2779)</td>
<td>Cluster C</td>
<td>Always</td>
</tr>
<tr>
<td>gapA (b1779)</td>
<td>Cluster C</td>
<td>Always</td>
</tr>
<tr>
<td>pgk (b2926)</td>
<td>Cluster C</td>
<td>Always</td>
</tr>
<tr>
<td>guaB (b2508)</td>
<td>Cluster D</td>
<td>Shared</td>
</tr>
<tr>
<td>prsA (b1207)</td>
<td>Cluster E</td>
<td>Always</td>
</tr>
<tr>
<td>adk (b0474)</td>
<td>Cluster E</td>
<td>Always</td>
</tr>
<tr>
<td>entD (b0583)</td>
<td>Cluster F</td>
<td>Always</td>
</tr>
<tr>
<td>nrdA (b2234)</td>
<td>Cluster H</td>
<td>Always</td>
</tr>
<tr>
<td>nrdB (b2235)</td>
<td>Cluster H</td>
<td>Always</td>
</tr>
<tr>
<td>glyA (b2551)</td>
<td>Cluster J</td>
<td>Shded</td>
</tr>
<tr>
<td>lpd (b0116)</td>
<td>Cluster J</td>
<td>Glucose</td>
</tr>
<tr>
<td>serA (b2913)</td>
<td>Cluster J</td>
<td>Shded</td>
</tr>
<tr>
<td>serB (b4388)</td>
<td>Cluster J</td>
<td>Shded</td>
</tr>
</tbody>
</table>

* All listed genes are reported as essential based on experimental data on glucose MOPS medium [65] and analyzed by [38]. Glycerol minimal medium data were derived and analyzed by [70]. All conditions were aerobic.

* Abbreviations. *Always*: always essential under rich medium, *Glucose*: essential on glucose minimal medium conditions only, *Shared*: essential on both glucose and glycerol minimal media.

* At least one of the genes forming a pair with these genes is not expressed under aerobic glucose conditions based on data from [67], for expression level cutoff of 300.

* Classified as non-essential based on the analysis of the glucose minimal medium data of [65] by Satish Kumar and Maranas [50].

We found only two cases of mismatches with experimental results concerning auxotrophy: (aroL, aroK) and (tktA, tktB). The SL pair (tktA, tktB) is auxotrophic for aromatic amino acids and requires the addition of pydxn (pyridoxine) in the medium [47]. In contrast, the *in silico* predictions found that it remained a SL even in a rich medium, as it is unable to produce the biomass precursor pydx5p (pyridoxal 5'-phosphate). Pyridoxine is a direct precursor to pydx5p, but inspection of the transport reactions...
contained in iAF1260 reveals that no pyridoxine transport reaction is present. Thus, we resolved the in vivo/in silico conflict for this SL through the addition of a pyridoxine uptake pathway to the model. Interestingly, adding this uptake pathway also leads to the corrected prediction that pdxH (b1638) is non-essential in rich medium after implementing the regulatory adjustments. Table 1.4 summarizes all suggested iAF1260 model modifications.

Table 1.4. Model refinements for iAF1260 suggested by SL gene pair analysis.

<table>
<thead>
<tr>
<th>Modification</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suppress ubiX (b2311)</td>
<td>Cannot complement ubiD (b3843); not expressed&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Suppress malY (b1622)</td>
<td>Cannot complement metC (b3008); not expressed</td>
</tr>
<tr>
<td>Suppress ydB (1692)</td>
<td>Cannot complement aroE (b3281); not expressed</td>
</tr>
<tr>
<td>Suppress ygeS (b2866)</td>
<td>Cannot complement guaB (b2508); not expressed</td>
</tr>
<tr>
<td>Suppress ygeT (b2867)</td>
<td>Cannot complement guaB (b2508); not expressed</td>
</tr>
<tr>
<td>Suppress ygeU (b2868)</td>
<td>Cannot complement guaB (b2508); not expressed</td>
</tr>
<tr>
<td>Suppress ndrE&lt;sup&gt;b&lt;/sup&gt; (b2675) and ndrF (b2676)</td>
<td>Cannot complement ndrA (b2234) and ndrB (b2235); not expressed</td>
</tr>
<tr>
<td>Suppress PDX5PO2</td>
<td>Cannot complement pdxH (b1638)</td>
</tr>
<tr>
<td>Suppress OPHHX3 under aerobic conditions</td>
<td>Cannot complement ubiB (b3835)</td>
</tr>
<tr>
<td>Suppress R1PK under aerobic conditions</td>
<td>Cannot complement prsA (b1207)</td>
</tr>
<tr>
<td>Suppress cnk (b0910)</td>
<td>Cannot complement pyrH (b0171); Hypothesis</td>
</tr>
<tr>
<td>Suppress ydgB (b1606)</td>
<td>Cannot complement folA (b0048); Hypothesis</td>
</tr>
<tr>
<td>Suppress cynT (b0339)</td>
<td>Cannot complement yadF (b0126); Hypothesis</td>
</tr>
<tr>
<td>Suppress metH (b4019)</td>
<td>Cannot complement metE (b3829); Hypothesis</td>
</tr>
<tr>
<td>Change HSDy GPR&lt;sup&gt;a&lt;/sup&gt; relationship from OR to AND</td>
<td>not expressed under aerobic glucose conditions based on data from [67], for an expression level cutoff of 300.</td>
</tr>
<tr>
<td>Suppress puuA (b1297) cannot complement glnA (b3870)</td>
<td>Hypothesis</td>
</tr>
<tr>
<td>Suppress ppsA (b1702)</td>
<td>Cannot complement eno (b2779) or gapA (b1779); Hypothesis</td>
</tr>
<tr>
<td>Suppress gcVP (b2903), gcVH (b2904), gcVT (b2905)</td>
<td>gcVP (b2903), gcVH (b2904), gcVT (b2905) cannot complement serA (b2913), serB (b4388), glyA (b2551); Hypothesis</td>
</tr>
</tbody>
</table>

<sup>a</sup> All modifications are for aerobic glucose conditions unless specified otherwise. Abbreviations: GPR: gene-protein-reaction association.

<sup>b</sup> The expression level of ndrE from data in [67] was only slightly above the expression level cutoff (300).

<sup>c</sup> not expressed under aerobic glucose conditions based on data from [67], for an expression level cutoff of 300.

### 1.3.2. Synthetic lethal triples

#### 1.3.2.1. Synthetic lethal gene triples

The concept of synthetic (pair) lethality can be extended to SL triples where the simultaneous deletion of three genes is lethal. When searching for SL triples, all essential genes and SL pairs are excluded from consideration to eliminate trivial results. We
identified 193 SL gene triples involving 114 genes and 15 non-gene associated reactions. Of these predicted SL triples, 111 (~57%) found to yield auxotroph strains in silico that can restore growth through the supplementation of the growth medium and the rest result in strains that cannot be rescued even in a supplemented medium.

Similarly to SL gene pairs, a variety of different topological motifs emerge when all SL gene triples are depicted. Note that we pictorially represented them using a triangle with the three members making up the SL triple depicted as edge connected nodes (see Figure 1.3). Figure 1.3 shows a number of disjoint triples and $k$-connected clusters of

---

**Figure 1.3. Topological classification of motifs in SL gene triples.** Nodes set in roman type are non-gene associated reactions. Note that that all the reaction abbreviations follow those in iAF1260 [38].
different size. An example of a disjoint triple is cluster A where two genes mgtA (b4242) and corA (b3816) form a SL triple with a non-gene associated reaction (i.e., Mg2t3_2pp (magnesium (Mg^{2+}) transport in/out via proton antiport (periplasm)). All the components of this cluster are responsible for magnesium transport under different mechanisms. Cluster H is an example of a 1-connected cluster, where the presence of at least one gene (i.e., pitA or pitB) can prevent lethality. As seen in Figure 1.3, unlike SL gene pairs, only a small number of SL gene triples participate in disjoint triples. Instead the majority of them form complex $k$-connected clusters (e.g., clusters K, L and M). We employed the mixed-integer optimization formulation proposed by Burgard et al [72] to identify the minimum required set of genes and non-gene associated reactions in each of these clusters to prevent lethality. Surprisingly, for clusters K and L we found that the minimal sets contained only a single member (i.e., $k=1$). For example, by maintaining only the activity of purT (b1849) in cluster K or the activity of either purU (b1232) or purN (b2500) in cluster L we can prevent lethality. Unlike clusters K and L, cluster M has 14 alternative minimal sets each containing 9 members (i.e., $k=9$) that need to be active to prevent lethality.

1.3.2.2. Synthetic lethal reaction triples

The application of the exhaustive enumeration procedure described in the methods section for single reaction deletions led to the identification of 277 essential reactions (~13.5% of the total number of reactions). Note that we did not allow any of the 304 exchange reactions and the 29 spontaneous reactions in the iAF1260 model to participate in any SL. After excluding all essential, exchange, spontaneous and 981 blocked reactions we first considered using the exhaustive enumeration procedure (see methods section) on the 792 remaining reactions to identify what pair or triple combinations of reaction eliminations negate biomass formation. For the case of pairs, we found 96 SL reaction pairs. However, applying this approach to identify all SL reaction triples would have required exhaustively exploring 83 million triple combinations. To avoid this computational burden we developed a targeted enumeration procedure (see methods section) relying on a bilevel optimization procedure to identify all synthetic reaction triples without having to explicitly test all 83 million triple combinations. It identified a
total of 243 SL triples involving 163 reactions. Table 1.5 depicts the CPU times for the targeted enumeration procedure versus the exhaustive enumeration procedure revealing orders of magnitude in improvement.

Table 1.5. Comparison of the CPU times for finding different orders of SL reactions using the exhaustive and targeted enumeration approaches. The CPU times (single 3GHz) are approximate.

<table>
<thead>
<tr>
<th>Order of SLs</th>
<th>Exhaustive enumeration</th>
<th>Targeted enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Possible combinations</td>
<td>SL %</td>
</tr>
<tr>
<td>Single</td>
<td>~ 2050</td>
<td>13.5</td>
</tr>
<tr>
<td>Double</td>
<td>~ 313,000</td>
<td>0.03</td>
</tr>
<tr>
<td>Triple</td>
<td>~ 8.3×10^7</td>
<td>2.9×10^{-4}</td>
</tr>
<tr>
<td>Quadruple</td>
<td>~ 1.6×10^{10}</td>
<td>N/D</td>
</tr>
</tbody>
</table>

N/D: not determined

Similarly to gene SLs, elimination of these reaction triples can yield auxotroph strains capable of restoring growth through the supply of missing nutrients or strains that lack essential functionalities that cannot be rescued by adding extra components to the growth medium. Of the 243 predicted essential reaction triples, as many as 202 (83%) were found to yield in silico auxotroph strains that can be rescued through supplementation. For example, elimination of PGK (phosphoglycerate kinase), TALA (transaldolase) and TPI (triose-phosphate isomerase) results in a strain that can be rescued (according to the model) through the supplementation of the growth medium by murein5px4p (two disacharide linked murein units, pentapeptide crosslinked tetrapeptide). In contrast, eliminating AGMHE (ADP-D-glycero-D-manno-heptose epimerase), RPE (ribulose 5-phosphate 3-epimerase) and TALA (transaldolase) yields a strain, which cannot produce in silico all biomass precursors even for a supplemented medium since it is unable to make or uptake the precursor metabolite pydx5p (pyridoxal 5'-phosphate).

Reaction SL triples are also pictorially shown as triangles with the three reactions depicted as edge connected nodes (see Figure 1.4). This graphical representation reveals that only a small number of reactions (i.e., 34) form small clusters (see Figure 1.4) while most of them (i.e., 129) are joined together into the highly connected large cluster I (i.e., giant component). For example, all three reactions in cluster A shown in Figure 1.4 are responsible for magnesium transport under different mechanisms. Interestingly, by
looking at GPR associations we find that this cluster maps exactly to cluster A of Figure 1.3 as the two reactions MG2uabcpp and MG2tpp are coded for by the genes mgtA (b4242) and corA (b3816), respectively. The giant component (cluster I) consists of 129 reactions forming 222 SL triples. It is $k$-connected with $k$ equal to 23. We employed the mixed-integer optimization formulation proposed by Burgard et al [72] to identify 9 alternative reaction sets, each with 23 reactions (i.e., $k = 23$) that allow for all biomass components formation. These 9 minimal reaction sets spanned 29 different reactions, with 17 of them present in all 9 alternative minimal sets.

**Figure 1.4.** Topological classification of motifs in SL reaction triples. Note that all the reaction abbreviations follow those in iAF1260 [38].

We analyzed further the reaction SL triples by determining the number of SL triples in which each reaction participates (see Table 1.6). We find a wide range of participation for different reactions. Most of the reactions (i.e., 85%) appear in seven or fewer triples, whereas only 12 reactions participate in more than 10 triples. Notably, TPI (triose-phosphate isomerase) is the most highly triple-participating reaction, with membership in 35 different SL triples. Not surprisingly, all these twelve high triple-participating reactions appear in the complex cluster of SL triples (cluster I).
Table 1.6. Frequency of participation of reactions in multiple SL triples.

<table>
<thead>
<tr>
<th># of SL triples</th>
<th># of reactions (Rxn abb.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>1 (RPI)</td>
</tr>
<tr>
<td>12</td>
<td>1 (PPS)</td>
</tr>
<tr>
<td>17</td>
<td>1 (PGI)</td>
</tr>
<tr>
<td>20</td>
<td>1 (PGM)</td>
</tr>
<tr>
<td>27</td>
<td>3 (GAPD, PGK, TALA)</td>
</tr>
<tr>
<td>33</td>
<td>1 (RPE)</td>
</tr>
<tr>
<td>34</td>
<td>3 (ATPS4rpp, FTHFD, GARFT)</td>
</tr>
<tr>
<td>35</td>
<td>1 (TPI)</td>
</tr>
</tbody>
</table>

All abbreviations follow those in [38].

Interestingly, almost half of these reactions belong to glycolysis, whereas the rest of them (except ATPS4rpp) are involved in pentose phosphate pathway (PPP), folate metabolism and purine and pyrimidine biosynthesis. Among the other interesting patterns are two highly participating in SLs reactions FTHFD and GARFT that are coded for by the two highly participating in SLs genes of cluster L in Figure 1.3, i.e., purU (b1232) and purN (b2500). Similar observations for cluster J of Figure 1.1 and clusters A of Figures 1.3 and 1.4 indicate that gene synthetic lethality can be explained by looking at the corresponding reaction synthetic lethality.

1.3.3. Higher-order synthetic lethals

We identified a number of SL quadruples for the iAF1260 model using aerobic minimal glucose medium conditions. Upon excluding spontaneous, exchange, blocked and essential reactions 229 SL reaction quadruples with 137 reactions involved were identified. Using the targeted enumeration approach we were able to elucidate even some higher order SL interactions, such as SL reaction quintuples. For example, the set of reactions F6PA (fructose 6-phosphate aldolase), FBA (fructose-bisphosphate aldolase), GLCptspp (D-glucose transport via PEP:Pyr PTS [periplasm]), GLCt2pp (D-glucose
transport in via proton symport [periplasm]) and RPI (ribose-5-phosphate isomerase) is a SL quintuple.

Unlike essential reactions that cannot participate in any SLs, it is possible for a gene/reaction involved in SL pairs to also participate in one or more SL triples or even higher order SLs. Figure 1.5 shows in the form of a Venn diagram the number of genes/reactions that participate in various orders of synthetic lethals. We note that 183 non-essential, “non-blocked” genes participate in at least one SL pair or triple. Interestingly, we see that fourteen genes participating in at least one SL pair also participate in some of the SL triples. In the case of SL reactions, we identified 39 that participate in both SL pairs and triples. For example, the glycine cleavage system (GLYCL), which is involved in the degradation of glycine to ammonia and CO₂, participates in four SL pairs, ten SL triples and nine SL quadruples (see Figure 1.6). We also identified up to 19 SL quintuples for GLYCL. This implies that the deletion of any of the reactions depicted in Figure 1.6 can be compensated by GLYCL alone. Alternatively, the removal of GLYCL would render PGCD, PSP_L, PSERT and GHMT2r essential. Notably, all the reactions present in Figure 1.6 (except PSERT) are associated with the genes of cluster J of Figure 1.1. Among these reactions GHMT2r converts glycine to serine, whereas, PGCD, PSERT and PSP_L serve as the first, second and third committed steps in the serine biosynthesis pathway, respectively. Therefore, the primary reason for the synthetic lethality of GLYCL with any of these four reactions is the requirement for serine production directly or through conversion from glycine.

Figure 1.5. Venn diagram of the number of genes and reactions participating in SLs of order one, two and three.
Figure 1.6. Pictorial view of all SLs for reaction GLYCL (glycine cleavage system). Double, solid and dashed lines depict SL pairs, triples and quadruples, respectively. The reaction abbreviations follow those in iAF1260 [38]. Note that for ease of presentation reactions that occur in both SL triples and SL quadruples are repeated (marked with a star) and we show neither the SL quadruples containing the coupled fluxes nor any of the SL quintuples.

Furthermore, the elimination of GAPD or PGK with either ENO or PGM prevents the production of metabolite 3pg (3-phospho-D-glycerate), thereby blocking the serine biosynthesis pathway. Finally, the removal of GLYCL with any combination of two from GLUCYS, GLYAT, AACTOOR and GTHRDabc2pp prevents the formation of the biomass precursor metabolites coa (co-enzyme A), amet (S-Adenosyl-L-methionine) and sheme (sirohem). This comprehensive synthetic lethality analysis of GLYCL demonstrates that by looking at higher-order SLs one can unravel non-intuitive biomass component deficiencies that may not be apparent from a visual inspection of the metabolic map.

A similar pattern emerged in the identified set of SL reaction quadruples. Most of the highly participating reactions in SL triples also appear with high frequency in the list of identified SL quadruples. For example, TPI (triose-phosphate isomerase), the reaction that participated in the most number of SL triples, also appears most often in the SL
quadruples (i.e., participates in 55 SL quadruples). Notably, we found many instances of more than one highly participating member of SL triples occurring together in the SL quadruples. For instance, out of 55 SL quadruples found for TPI, 21 contained PGM (phosphoglycerate mutase). The reason these reactions appear in many SLs of different orders is that they serve as key branch points of central metabolism. The simultaneous removal of multiple branch points will require flux rerouting through other bypass reactions or latent pathways [73] for the production of essential biomass precursors such as amino acids.

1.3.4. Degree of essentiality

In order to quantify the degree of dispensability of a gene or reaction in a metabolic network with respect to biomass formation we introduce the concept of degree of essentiality (DOE). This metric is defined as the size of the smallest SL that the gene or reaction is a member of. Therefore, essential genes or reactions have a DOE of one while genes or reactions that participate in SL pairs (and perhaps in higher order SLs) have a DOE of two. It should be noted that the DOE metric for genes is akin to the ‘k-robustness’ term introduced by Deutscher et al [33]. As noted above, we determined the DOE of up to three for all genes and reactions and the DOE of up to four for all reactions of central metabolism active under aerobic glucose conditions. The distribution of DOE for genes and reactions present in different COG classifications [41] is shown Figure 1.7. Figure 1.7 shows that genes and reactions in different COGs have quite different DOE statistics. Figure 1.7b pictorially delineates the percent reaction participation in each DOE across all COGs and reveals the differing buffering capacity of each functional category for biomass formation [33].

Next, we focus our attention to the DOE results for reactions participating in the central metabolism spanning glycolysis, pentose phosphate pathway (PPP), TCA cycle and anaplerotic reactions. Figure 1.8 illustrates the color-coded degree of essentiality of all reactions in central metabolism up to DOE of four. We can see that the majority of reactions in central metabolism have a DOE of two, three or more. This is most likely due to the presence of multiple diverging and converging branches in pathways of central
metabolism. The relatively small fraction of essential reactions was expected, as previous reports noted that conserved metabolic pathways such as the TCA cycle or glycolysis generally contain few essential reactions [74,75]. Most of the reactions in the PPP have a DOE of two whereas glycolysis reactions involve DOEs of three or more. No reaction in glycolysis or PPP has a DOE of one, while four of the five reactions with an essentiality degree of one belong to the TCA cycle. Eliminating any of these four essential reactions will prevent the formation of the same list of precursor metabolites including sheme (sitroheme), pheme (protoheme) and murein5px4p (two disacharide linked murein units, pentapeptide crosslinked tetrapeptide). All four of these reactions are subsequent steps in TCA cycle converting oxaloacetate to 2-ketoglutarate. Application of flux coupling analysis [39] showed that three of them are fully coupled (i.e., ACONTa (aconitase [half-reaction B, Isocitrate hydro-lyase]), ACONTb (aconitase [half-reaction B, Isocitrate hydro-lyase]) and ICDHyr (isocitrate dehydrogenase)). It is important to note that reactions operating in opposite directions can have different DOEs. Such examples include reaction pairs FBP and PFK as well as PPC and PPCK.
We presented a comprehensive *in silico* study of gene and reaction synthetic lethality in *E. coli* K12 based on the recent genome-scale metabolic model *iAF1260*. This computationally intensive goal was made possible by developing an efficient procedure.

1.4. Discussion and conclusion

Figure 1.8. Color-coded representation of the reactions in central metabolism according to their degree of essentiality.
for the targeted enumeration of all SL interactions relying on bilevel optimization. Unlike earlier efforts that relied on incomplete sampling to elucidate partial lists of higher-order SLs [34,35], here we provided complete enumerations of high-order SL interactions for both gene and reaction centric representations. The network organization of the elucidated SLs, recapitulated modular lethality relationships consistent with previous observations in yeast [26]. By coloring network nodes using the COG classification surprising compensatory interactions between seemingly unrelated reaction classes were revealed. Earlier efforts for the in silico elucidation of SLs [30,31,32] did not anticipate lethality caused by the inability to meet the non-growth associated ATP maintenance requirement. Notably, in this study we found over 120 SLs that were triggered by this deficiency.

By contrasting literature data on gene essentiality and synthetic lethality against predicted SLs, 27 instances of inconsistencies (false-positive SLs) were identified. Similar examples of false-positive predictions can be also found for reaction SLs. For example, reaction RPI (ribose-5-phosphate isomerase) was found to be essential in vivo [76], however according to the model iAL1260, it forms synthetic lethal pairs with RPE (ribulose 5-phosphate 3-epimerase), TALA (transaldolase), TKT1 and TKT2 (transketolase) and also participates in a number of higher-order SLs (see Table 1.4). These erroneous model predictions are due to the fact that genome-scale metabolic model reconstructions tend to over rather than under predict the metabolic capabilities of the organism. This arises from the inclusion in the model of functionalities that are not active at a sufficient level (e.g., due to regulation) to ensure biomass formation. The list of SLs reported in this study should therefore be interpreted as a conservative depiction of synthetic lethality, as we recognize that many additional synthetic lethals are likely present that stoichiometric models alone cannot reveal (i.e., false-negatives).

We exploited the identified in vivo/in silico mismatches to suggest a number of iAF1260 model modifications (see Table 1.4). This is motivated by the observation that false negative/positive model predictions provide opportunities for not only model improvement but also re-evaluation of experimental data [29]. Many of these model corrections do not involve the addition or removal of reactions to iAF1260 but rather the incorporation of regulatory constraints (i.e., condition-dependent presence of different
reactions). Unfortunately, existing experimental data on SLs account for only a small portion of the predicted SL thus limiting the potential for model improvement. This calls for the development of combinatorial methods for the rapid generation and screening of SLs such as the recently developed GIANT-coli technology that allows for the high-throughput generation of double mutants in *E. coli* [22].

One of the key challenges in correctly interpreting the obtained predictions is the delineation of the true function of isozymes and complementation under the studied conditions. For example, the experimental evidence for the predicted SL (*ubiX, ubiD*) is conflicting. Some analyses propose that the *ubiD* knockout is lethal [38,65,66], whereas the data in Covert *et al* [67] suggest that *ubiX* is not expressed under the examined conditions. In contrast, Gulmezian *et al* [60] recently observed that both UbiX and UbiD are required for decarboxylation, especially during logarithmic growth, implying that both genes are essential. These inconsistencies among *in silico* predictions and *in vivo* observations call for more nuanced model modifications that are dependent on not only conditions but also on growth phase. In another example, *metL* (b3940) and *thrA* (b0002) form a disjoint SL pair and are isozymes for the aspartate kinase activity. Surprisingly, both genes are reported to be essential (Table 1.4) suggesting that the OR operator must be changed into an AND operator in the gene-protein-reaction association. Therefore, simply knowing that a gene is essential is not always sufficient. In some cases, elucidating the functional reason(s) (e.g., loss of sufficient aspartate kinase activity, etc.) for this essentiality is needed before properly correcting the model. As indicated in Table 1.3, the classification of *entD* in cluster F as essential is likely erroneous. This observation is supported by the fact that eliminating FE3abcpp to make *entD* essential would also render another twelve genes essential that are known to be non-essential. In addition, the essentiality score of *entD* is very close to the threshold (see Table 1.3). Finally, predictions on whether the disruption of a particular SL results in an auxotroph strain can likewise be exploited for model correction as we demonstrated for SL (*tktA, tktB*).

The introduction of the concept of degree of essentiality enabled the quantitative assessment of the dispensability of any gene or reaction in the metabolic network. Moreover, by queering the derived list of SLs, one can examine how the removal of a
gene/reaction affects the dispensability characteristics of other genes/reactions. Results for GLYCL (see Figure 1.6) revealed surprising compensatory interactions with reactions in seemingly unrelated pathways. The elucidation of SL and DOE in human metabolism has implications in the identification of drug targets. For example, it has been suggested that the SLs partners of missing enzymatic functionalities of tumors cells would be promising drug targets [16,17]. The idea is that healthy cells would remain unaffected due to the ability to compensate for the drug-suppressed functionality, whereas tumor cells, with the missing enzymatic functionality, would not. On another front, SL predictions could be used to pinpoint multi-gene disease mappings [77] and identify combinations of genes most likely to interact in disease phenotypes [24].

The procedure proposed herein can be used to rapidly predict the growth phenotypes of multiple knockout mutants for a variety of other organisms such as *S. cerevisiae, B. subtilis* and *H. pylori* in various media. The effect of different conditions (i.e., alternate carbon substrates, aerobic vs. anaerobic, etc.) and/or certain regulatory constraints on the membership of the SL sets can be assessed in a straightforward manner by adjusting appropriate model constraints based on exchange reaction usage and gene/reaction availability.
Chapter 2

Improving the iMM904 S. cerevisiae Metabolic Model Using Essentiality and Synthetic Lethality Data

This chapter has been previously published in modified form in BMC Systems Biology [78].

2.1. Introduction

Saccharomyces cerevisiae is the first eukaryote whose genome was fully sequenced and annotated [79]. It has since then been the focus of many genome-scale reconstruction efforts. Forster et al [25] reconstructed the first multi-compartment genome-scale metabolic model for yeast (i.e., iFF708). The model accounted for 708 open reading frames (ORFs) (~10.5% of all ORFs) and 1,175 metabolic reactions. The metabolic reactions for this model were compartmentalized between cytosol and mitochondria. Transport mechanisms between these two compartments as well as between the environment and cytosol were included in the model. Soon thereafter, the iND750 model was introduced, which included five additional compartments (i.e., peroxisome, nucleus, golgi apparatus, vacuole and endoplasmic reticulum) by re-assessing the localization of gene products [80]. In a parallel study [14], the predictive capability of iFF708 was improved through a number of modifications in the biomass equation and the removal of
blocked reactions (iLL672 model). Subsequently, another version of the yeast metabolic model with an improved description of the lipid metabolism containing 800 genes and 1,446 reactions (i.e., iN800) was introduced by Nookaew et al [81]. These improvements to the yeast model culminated with the iMM904 model [82] that increased the size of iND750 to 904 genes and 1,412 reactions. All these metabolic reconstruction efforts were carried out largely independently of one another using different data sources or in some cases different interpretations of the same literature evidence. This lack of consistency motivated the yeast systems biology community to consolidate all available metabolic models into a single consensus reconstruction and annotation model [83]. This model has been updated regularly since it was published and the latest version (i.e., Yeast 4.0) contains, in decompartmentalized form, 1102 reactions and 924 proteins [84].

Despite all these efforts there still exists a gap in the quality between the available metabolic reconstructions for yeast and corresponding models for microbial metabolism. Table 2.1 summarizes the quality metrics for the E. coli [38], M. genitalium [85], B. subtilis [86], P. putida [87], H. pylori [29] and Salmonella Typhimurium [88] latest genome-scale models when growth predictions are contrasted against experimental data for single gene deletions. As shown in this table, the accuracy (i.e., the fraction of correctly predicted lethal knockouts) of the iMM904 metabolic model of the yeast is significantly worse than any one of corresponding microbial models. This is partly caused by uncertainty in enzyme localization and inter-compartment metabolite transport in yeast [89]. This implies that a draft model reconstruction followed by even a detailed manual curation may not be a sufficient strategy to bring a eukaryotic genome-scale model to the same quality level as a microbial one. Here we explore the extent of model correction that can be brought about by systematically comparing the model predictions for single and multiple gene deletions with available experimental data.

The established standard [89] for testing the accuracy of genome-scale metabolic models is to contrast the predicted growth phenotype of single mutant strains with the available experimental data under various growth conditions [38,66,82,90]. These comparisons result in four different outcomes [50]: GG or NGNG when both model and experimental data either imply growth (G) or no growth (NG) for the mutant strain, NGG
Table 2.1. Accuracy of microbial metabolic models vs iMM904 model. Comparison of the fraction of correctly predicted lethal knockouts (i.e., specificities) of microbial genome-scale metabolic models against the yeast iMM904 model for single gene mutation experiments.

<table>
<thead>
<tr>
<th>Microorganism</th>
<th>Name of the metabolic model</th>
<th>Specificity (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>iMM904</td>
<td>38.8</td>
<td>[82] and this study</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>iAF1260</td>
<td>73.4</td>
<td>[38]</td>
</tr>
<tr>
<td><em>Mycoplasma genitalium</em></td>
<td>iPS189</td>
<td>79.0</td>
<td>[85]</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>iBSU1103</td>
<td>89.3</td>
<td>[86]</td>
</tr>
<tr>
<td><em>Pseudomonas putida</em></td>
<td>-</td>
<td>74.5</td>
<td>[87]</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em></td>
<td>iT341 GSM/GPR</td>
<td>73.0</td>
<td>[29]</td>
</tr>
<tr>
<td><em>Salmonella Typhimurium</em></td>
<td>iMA945</td>
<td>66.7</td>
<td>[88]</td>
</tr>
</tbody>
</table>

when the model predicts that the gene deletion is lethal but the experiment shows that it is viable, and finally GNG when the model predicts that the mutant strain would be viable but *in vivo* observations show a lethal effect. Reed *et al* [70] proposed a systems analysis approach to restore growth for NGGs for a variety of growth media through the addition of appropriate metabolic and transport functions to the model. In another study, Satish Kumar *et al* [91] introduced GapFind and GapFill, to first identify metabolites that cannot be produced or consumed in the model under any uptake conditions and then bridge these gaps by adding missing functionalities to the model. Subsequently, another procedure termed GrowMatch [50] was proposed to reconcile both NGG and GNG growth prediction inconsistencies across different substrates (see Orth and Palsson [92] for a review). Notably, in the GrowMatch procedure [50], the GNG mismatches are corrected by modifying the metabolic model so as to convert them to NGNGs. Alternatively, Motter *et al* [93,94] suggested to reconcile GNGs by identifying candidate gene knockouts that can restore the growth of the mutants that were initially non-viable *in vivo*.

Recent studies have suggested that making use of not only single gene deletion information but also synthetic lethal pairs or (higher orders) can provide an additional layer for curation and validation of metabolic models. Harrison *et al* [95] showed that the investigation of falsely predicted synthetic lethals of *S. cerevisiae* can help to improve functional annotation. Furthermore, recent research by our group [1] demonstrated that mismatches between both growth and auxotrophy phenotypes of synthetic lethal predictions and *in vivo* observations can be used to provide 19 model correction hypotheses for the iAF1260 model of *E. coli*. In this paper, we modify and deploy
automated approaches [50,91] for resolving falsely predicted single and multiple gene deletions under aerobic minimal and a complex (YP) medium, through the generation of appropriate model-correction hypotheses for the iMM904 model of the yeast. We chose iMM904 for this study because it contains biomass composition and reaction reversibility information making it suitable for performing FBA computations. Following the workflow presented in this paper we pinpointed 90 model corrections and identified 30 missing regulatory constraints with supporting literature evidence that almost doubled the prediction accuracy of the growth phenotype of single and double gene deletions under aerobic minimal and YP media for the iMM904 model. Examples of literature evidence include interaction of two proteins to support a modification in GPR associations, presence of a specific compound in the cell wall to corroborate its inclusion in the biomass equation, or gene expression data to confirm suppression of a gene under a given condition. The majority of modifications proposed here (i.e., 86% of them) remain relevant even for the latest update of the community yeast model (Yeast 4.0) [84].

2.2. Methods

We applied the SL Finder procedure developed by Suthers et al [1] to the iMM904 model of S. cerevisiae [82] comprised of 904 genes and 1,412 reactions to determine the set of essential genes/reactions as well as different orders of synthetic lethals. All simulations were performed for aerobic growth on two different media with glucose as the sole carbon source: minimal medium and a complex (yeast-extract peptone or YP) medium. The in silico minimal medium contains ammonium, sulfate and phosphate as nitrogen, sulfur and phosphor sources, respectively, as well as necessary salts (such as Na and K). In addition to all these components the in silico YP medium contains all 20 amino acids and all four nucleotides [80]. All simulations were performed for a strain auxotrophic for methionine, leucine, histidine and uracil to closely mimic the conditions used in experimental studies [14]. This auxotrophy was simulated by deleting genes (his3Δ leu2Δ met15Δ ura3Δ) and supplementation of the growth medium with the missing nutrients at non-limiting but low levels [82]. In addition, trace amounts of other essential compounds that are present in the experimental growth medium including 4-aminobenzoate, biotin, inositol, nicotinate, panthothenate and thiamin were added to the
in silico media [82]. Consistent with Suthers et al [1] we chose 1% of the maximum theoretical biomass yield as the viability threshold for computational identification of lethal knockouts. The upperbound for all reactions was set to 1000, whereas, the lowerbound was set to 0 for irreversible reactions and -1000 for reversible reactions. The maximum rate of the glucose uptake was set to 10 mmol gDW\(^{-1}\)h\(^{-1}\) and the aerobic condition was modeled by limiting the oxygen uptake rate to 2 mmol gDW\(^{-1}\)h\(^{-1}\) [82]. The uptake rate for all other source metabolites was also set to 1000 mmol gDW\(^{-1}\)h\(^{-1}\). The flux of non-growth associated ATP maintenance was fixed at 1 mmol gDW\(^{-1}\)h\(^{-1}\) [82]. We employed the GrowMatch procedure [50] to reconcile the growth phenotype discrepancies for the NGGs, and modified it to fix GNGs (in addition to using its original form). All these algorithms were adapted for resolving the inconsistencies associated with double gene perturbations. In the following we provide a brief overview of GrowMatch and describe in detail its modified version.

2.2.1. The modified GrowMatch procedure

The GrowMatch procedure relies on two separate procedures to resolve NGG and GNG growth prediction inconsistencies. The NGGs are fixed one-by-one by minimally perturbing the original metabolic model using three mechanisms including (i) relaxation of the irreversibility constraints on reactions in the model, (ii) addition of new reactions from external databases such as KEGG [96] to the model and (iii) allowing for direct import/export of metabolites into/out of the cell, and for multi-compartment models [50], addition of transport reactions between compartments and cytosol. Alternatively, the GNG mismatches are corrected by identifying the minimal set of suppression constraints for reactions or transport mechanisms that lower the maximum biomass yield of the network below a pre-specified viability threshold. The suggested modifications by GrowMatch are referred to as global if they do not conflict with any correct model predictions, and they are called conditional otherwise.

Here, we modified the GrowMatch procedure to identify the minimal number of suppressed genes, rather than reactions, that negate the biomass formation below the viability threshold. This can be done by defining two sets of binary variables one for genes and one for reactions and then relating these two binary variables in such a way
that it reflects the specific GPR associations [1]. For the reactions with no gene associations a fictitious gene coding for that reaction can be assumed. An alternative way to avoid introducing a new set of binary variables, and as a result reduce the computational burden further, is to directly correlate the binary variables for genes to the reaction fluxes. To this end, we first, define the following sets:

\[ I = \{ i | i = 1,2,\ldots,N \} = \text{set of metabolites} \]
\[ J = \{ j | j = 1,2,\ldots,M \} = \text{set of reactions} \]
\[ K = \{ k | k = 1,2,\ldots,G \} = \text{set of genes} \]

where, \( N \), \( M \) and \( G \) denote the total number of metabolites, reactions and genes in the network, respectively. To simulate the gene knockouts, a binary variable \( w_k \), representing if a gene \( k \) should be deleted is defined as following:

\[
w_k = \begin{cases} 
0, & \text{if gene } k \text{ is deleted} \\
1, & \text{if gene } k \text{ is active} 
\end{cases} \quad \forall \ k \in K
\]  

(2.1)

The impact of gene knockouts on reactions through the GPR relationships can be mathematically described by using appropriate constraints relating \( w_k \) to the reaction fluxes, \( v_j \). Let \( LB_j \) and \( UB_j \) represent the lowerbound and upperbound on a reaction \( j \), respectively. Different cases for the GPR relationships can then be considered:

(i) A single gene \( k \) codes for the enzyme catalyzing a reaction \( j \). This can be easily incorporated into our mathematical framework by using the following constraint:

\[
w_k \cdot LB_j \leq v_j \leq w_k \cdot UB_j
\]  

(2.2)

(ii) Two genes \( k_1 \) and \( k_2 \) form a single multi-protein enzyme to enable a reaction \( j \). This case, which is recast as a logic AND relationship between the genes \( k_1 \) and \( k_2 \), can be enforced by the following set of constraints:

\[
\begin{cases} 
w_{k_1} \cdot LB_j \leq v_j \leq w_{k_1} \cdot UB_j \\
w_{k_2} \cdot LB_j \leq v_j \leq w_{k_2} \cdot UB_j
\end{cases}
\]  

(2.3)

Note that based on these equations, if at least one of \( w_{k_1} \) or \( w_{k_2} \) is zero, then the flux through reaction \( j \) is forced to zero. This set of constraints can be easily generalized for multi-protein enzymes with more than two genes involved.

(iii) The genes \( k_1 \) and \( k_2 \) provide isozymes catalyzing a reaction \( j \). This case, indicates a logic OR relationship between the genes \( k_1 \) and \( k_2 \) to enable the reaction \( j \), and can be mathematically expressed by the following constraints:
\[
\begin{align*}
&(w_{k_1} + w_{k_2})LB_j \leq v_j \leq (w_{k_1} + w_{k_2})UB_j \\
&LB_j \leq v_j \leq UB_j
\end{align*}
\] (2.4)

Note that according to these equations, \( v_j \) is forced to zero only if both genes are knocked out, i.e., \( w_{k_1} = w_{k_2} = 0 \). For the case, where both genes are present \( w_{k_1} = w_{k_2} = 1 \), the second constraint is more binding and will restrict \( v_j \) to fall within its defined lower and upperbound. This approach can be readily generalized for GPRs containing more than two genes related with OR.

\((iv)\) More than two genes with a combination of AND and OR relationships are required to enable a reaction \( j \). These cases can be mathematically enforced through an appropriate combination of equations (2.3) and (2.4). As an example, if three genes \( k_1, k_2 \) and \( k_3 \) are correlated as \( (k_1 \text{ AND } k_2) \text{ OR } (k_1 \text{ AND } k_3) \) to code for the enzyme catalyzing a reaction \( j \), then the constraints simulating this relationship can be written as following:

\[
\begin{align*}
&w_{k_1} LB_j \leq v_j \leq w_{k_1} UB_j \\
&(w_{k_2} + w_{k_3}) LB_j \leq v_j \leq (w_{k_2} + w_{k_3}) UB_j
\end{align*}
\] (2.5)

The modified GrowMatch formulation to resolve the GNGs (or GSLs) can now be formulated as following:

\[
\begin{align*}
\text{Minimize} & \quad v_{\text{biomass}} \\
\text{s.t.} & \quad v_{\text{biomass}} \\
\text{s.t.} & \quad \sum_j s_{ij} v_j = 0 \quad \forall \ i \in I \\
& \quad \text{Appropriate GRP eqns} \quad \forall \ j \in J \\
& \quad v_{\text{glucose}} \leq v_{\text{uptake limit}} \\
& \quad v_{\text{oxygen}} \leq v_{\text{uptake limit}} \\
& \quad v_{\text{ATPM}} = v_{\text{maintenance}} \\
& \quad LB_j \leq v_j \leq UB_j \quad \forall \ j \in J \\
& \quad \sum_j (1 - w_k) \leq n \\
& \quad w_k \in \{0,1\} \quad \forall \ k \in K
\end{align*}
\]
where, $s_{ij}$ represents the stoichiometric coefficient of the metabolite $i$, in reaction $j$, $v_{\text{biomass}}$ denotes the biomass flux while $v_{\text{uptake limit, glucose}}$, $v_{\text{uptake limit, oxygen}}$ and $v_{\text{maintenance}}$ denote the minimum required glucose and oxygen uptake rates and the non-growth associated ATP for maintenance, respectively. The parameter $n$, represents the allowable number of knock-outs. This bilevel optimization problem can be solved similarly to GrowMatch and SL Finder [1,50] through writing the dual of the inner problem.

2.2.2. Test of the suggested hypotheses

All the model correction strategies provided by GrowMatch serve as hypotheses that need to be tested to confirm their applicability. Different methods were used to test the validity of each type of modification. Similarly to [50], relaxation of the irreversibility constraints on existing reactions in the model can be checked by using three independent methods. In the first step we check the reversibility of that reaction in iAF1260 model of E. coli [38] or metabolic models of other organisms. We then query other databases such as MetaCyc [97] about the reversibility of the desired reaction to find out if it is reversible in other organisms. Finally, we examine the reversibility of reactions by computing the value of free Gibbs energy change, $\Delta G$ [98].

The validity of the added transport reactions to the model are examined by searching the literature to find potential clues about the presence of the suggested transport mechanisms or by querying the databases such as MetaCyc for possibility of presence of those mechanisms in other multi-cellular organisms. The hypotheses for adding new reactions from external databases such as KEGG to the model are tested by performing the bi-directional BLAST between the enzymes catalyzing those reactions and the yeast genome. Consistent with [50,91] we assumed a BLAST expectation value cutoff of $10^{-13}$ as the basis to define high sequence similarity. Finally, the gene suppressions are validated by analysis of the gene expression data as well as searching the literature for available evidence. All the global modifications for which we did not find any supporting evidence using the methods mentioned above, were not incorporated into the model and were just added to the list of modifications with no corroborating evidence.
2.3. Results

2.3.1. Single gene perturbations

Analysis of the impact of single gene or reaction deletions on the growth phenotype (for the iMM904 model) revealed 106 essential genes and 163 essential reactions for growth using a minimal medium as well as 57 essential genes and 92 essential reactions for the YP medium. We contrasted the predicted essential genes in both media to the experimental single gene deletion data reported by Kuepfer et al [14] to pinpoint any model inconsistencies. A summary of the model/experiment (mis)matches is given in Figure 2.1A and B. We note that the two metrics specificity and sensitivity in this figure represent the fraction of correctly predicted lethal and viable mutants, respectively. Comparing the accuracy of the iMM904 model with iFF708 (specificity = 68.2%) [25,99] for growth using YP medium, we find that there is a significant reduction in specificity. This reduction is primarily due to uncertainties associated with the assignment of functionalities to compartments as well as the expansion of the model with less studied reactions, as pointed out earlier [89]. We applied the computational procedures described in the Methods section to reconcile these false predictions with the experimental growth. We note that all reaction and metabolite abbreviations throughout this manuscript are based on the iMM904 model.

2.3.1.1. Resolution of NGG inconsistencies

As denoted in Figures 2.1A and B, we identified 19 NGGs for the minimal and 14 for the YP medium. The first step towards resolution of NGGs in the GrowMatch procedure is to look for alternate genes in yeast capable of carrying the same function but absent from the model or specific gene-protein-reaction (GPR) associations [50]. A bidirectional protein-protein BLAST (i.e., BLASTp) search against the S. cerevisiae genome revealed that seven of the genes involved in NGGs under the minimal medium and three under the YP medium share significant sequence similarity over the entire length of protein (i.e., forward and backward BLASTp expectation value of less than 10⁻¹³) with other ORFs in yeast. We however, rejected three such resolution strategies in the minimal medium and one in the YP medium due to contradicting experimental evidence.
Figure 2.1. Accuracy of the iMM904 model before and after modifications using single gene perturbations. The number of false and correct predictions of the iMM904 model before (A and B) and after (C and D) applying the modifications for single gene mutations under minimal and YP media, respectively. Note that specificity = \#NGNG/(\#NGNG + \#GNG), sensitivity or true viable rate (TVR) = \#GG/(\#GG + \#NGG) and false viable rate (FVR) = \#GNG/(\#GNG + \#NGNG).

Alternative correction strategies for the resolved NGGs were also explored. Two NGG inconsistencies under the minimal medium and one under the YP medium were resolved by relaxing the irreversibility of at least two reactions. For example, the deletion of gene ADKI (YDR226W), which is involved in nucleoside salvage pathway, becomes non-lethal by treating any of the reactions ADPT (Adenine phosphoribosyltransferase), HXPRT (Hypoxanthine phosphoribosyltransferase (Hypoxanthine)), UPPRT (Uracil phosphoribosyltransferase) or GUAPRT (Guanine phosphoribosyltransferase) as reversible. The reversibility of these reactions was further supported based on the value of free Gibbs energy change, \(\Delta G\), or previous reports in literature [100,101,102,103]. Notably, relaxing the irreversibility constraint on any of these reactions enables the
production of prpp (5-phospho-alpha-D-ribose 1-diphosphate), which serves as a precursor for many biomass components.

Upon incorporation of these two global modifications in the model, we pursued the resolution of additional NGGs by adding transport reactions or reactions from the KEGG database [96] (see Methods). We were able to reconcile the growth inconsistencies for two NGGs in the minimal medium by this method. For example, *ADE8* (YDR408C), which codes for reaction GARFTi (phosphoribosylglycinamide formyltransferase) was fixed by adding reaction R06974 (glycinamide ribonucleotide transformylase) to the model. GARFTi that serves as a step in the *de novo* purine nucleotide biosynthetic pathways is the only reaction in the model responsible for producing the essential metabolite fgam (N2-Formyl-N1-(5-phospho-D-riboyl)glycinamide). Addition of reaction R06974 to the model provides an alternative way of producing fgam that can compensate for the absence of GARFTi. By performing the BLAST bi-directional test we identified an ORF in the yeast genome (i.e., *ADE2*) with very high sequence similarity (forward E-value = 3×10^{-24} and backward E-value = 3×10^{-24}) with the enzyme phosphoribosylglycinamide formyltransferase 2 (EC 2.1.2.-) catalyzing R06974 in *Methanococcus jannaschii*, which supports the presence of this reaction in yeast.

Except for two cases (i.e., *RIB1* and *RIB4*), all other NGGs identified for the YP medium are also present and therefore corrected when considering the minimal medium. Both *RIB1* (YBL033C) and *RIB4* (YOL143C) are fixed not by modifying the *iMM904* model but by adding missing compounds to the *in silico* YP medium description. Both of these genes are involved in riboflavin (vitamin B2) biosynthesis, which is a biomass precursor. Because yeast extract is reported to be a major source of vitamin B2 [104], we decided to add riboflavin to the list of components in the *in silico* YP medium. This renders the deletion of *RIB1* and *RIB4* non-lethal. Notably, the importance of correctly describing the medium composition in growth phenotype predictions has been raised in previous studies [99]. Overall, by using the GrowMatch procedure along with literature searches we were able to fix thirteen NGGs under the minimal medium and eleven under the YP medium (see Figure 2.1).
2.3.1.2. Resolution of GNG inconsistencies

As shown in Figure 2.1A and B, a total of 137 GNG inconsistencies in both minimal and YP media was identified with 128 of them jointly present. The distribution of these genes across different functional classes of metabolism revealed that most of them are involved in tRNA charging, oxidative phosphorylation and extracellular transport. Resolution of these GNGs by GrowMatch generally involves suppression of incorrectly present alternative production routes of biomass components in the model. The identified GNGs are divided into three major categories based on how they affect the flux distribution in the network [50]. The first category is comprised of genes coding for isozymes alluding that the deletion of these genes should not affect the reaction flux. A straightforward resolution strategy in this case is to suppress the other gene(s) whose product serves as isozyme for the coded reaction. An alternative hypothesis is to modify the corresponding GPR relationship so as to recast the deletion of that gene as lethal. These resolution strategies would be viable only if the coded reaction(s) are essential or form synthetic lethal(s) according to the model. For example, gene *ACO1* (YLR304C), identified as a GNG under minimal medium, is an isozyme (with *ACO2*) for mitochondrial aconitase (*ACONTm*) and also independently catalyzes cytosolic aconitaase (*ACONT*) according to the iMM904 model. Notably, reactions, *ACONT* and *ACONTm* form a synthetic lethal pair under the minimal medium according to the iMM904 model. We did not find any evidence indicating that *ACO2* is suppressed under aerobic glucose condition. Instead, we found that the protein coded for by *ACO2* (Aco2p) is a putative mitochondrial aconitase isozyme, whose function has been assigned only based on the sequence similarity with Aco1p [105,106]. Therefore, we decided to remove *ACO2* from the GPR for *ACONTm* thus rendering the deletion of *ACO1* as lethal. This modification is consistent with the latest update of the community model [84]. Notably, another possibility is that *ACO2* is correctly assigned to *ACONTm*, but *ACO1* is involved in other unaccounted in the model essential functions (e.g., mitochondrial DNA maintenance [107]) in addition to its aconitase activity. We manage to fix five GNGs under the minimal medium and three under the YP medium by using this procedure.

The second category of GNGs contains genes that code for blocked reactions, i.e., the reactions that cannot carry any flux. This implies that even though the deletion of
these genes will not affect the flux distribution \textit{in silico}, their knockout is lethal \textit{in vivo}. Resolution of these inconsistencies involves first unblocking the coded reactions and next rendering its deletion lethal under the examined conditions. Notably, 52 of such these GNGs code for reactions that are always blocked in the model mainly due to the presence of a no-consumption metabolite. One can thus resolve the inconsistency by exploring whether the no-consumption metabolite is a component of the biomass equation. Interestingly, we found that 21 of such these GNGs code for tRNA charging reactions that can be converted to NGNG consistencies by including the charged and uncharged tRNA molecules in place of the corresponding amino acids in the biomass equation. This strategy has been previously used in the reconstruction of the \textit{Salmonella} metabolic model \cite{88}. For example, the GNG prediction for \textit{WRSI} (YOL097C), which codes for the blocked reaction TRPTRS (Tryptophanyl tRNA synthetase), was corrected by including charged tRNA (trptrna) and uncharged tRNA (trantrp) as a reactant and product of the biomass reaction, respectively. This modification simultaneously unblocks reaction TRPTRS and renders it essential.

In many cases, including the no-consumption metabolite of a blocked reaction encoded by a GNG into the biomass equation was not a viable option as it also converted some GG consistencies to NGG mismatches (i.e., a conditional modification). Using the GapFind procedure \cite{91} we found that such problem metabolites are all upstream no-consumption, implying that they can be fixed if their corresponding root (or one of its downstream) no-consumption metabolite(s) is resolved. Therefore, we tested whether these GNGs can be fixed globally, if the root or one of the downstream no-consumption metabolites is included as a component of the biomass equation. An additional seven GNGs corresponding to blocked reactions were fixed by this method but were not added to the list of proposed model modifications due a lack of experimental validation. We note that the incomplete description of the biomass equation has been implicated as an important source of false model predictions in earlier studies \cite{14,80,99}. Therefore, we exercised great caution when modifying the biomass equation to fix an inconsistency. Corrections were accepted only if solid corroborating literature evidence was found.

Two other GNGs were identified to correspond to a reaction that is blocked only under minimal conditions: Genes \textit{MPH2} (YDL247W) and \textit{MPH3} (YJR160C) encode
isozymes catalyzing the transport of maltose from the extra-cellular environment to cytosol (reaction MALT2 in iMM904). This reaction is blocked since no maltose is present in the minimal medium. Therefore, we explored whether these two genes are also involved in the transport of D-glucose (reaction GLCt1). A BLAST bi-directional analysis revealed that both of these genes have very high sequence similarity with glucose transport genes. Interestingly, we found that previous studies have reported on the involvement of MPH2 and MPH3 in mediating residual glucose uptake [108].

The third category of GNGs contains genes whose deletion affects the flux distribution in the network. To rectify these mismatches we employed the original GrowMatch procedure introduced in [50] and its modified version proposed in this study (see Methods) to identify the minimal number of genes/reactions whose suppression lower the biomass formation below the pre-specified viability threshold. We performed this analysis by allowing for up to two simultaneous gene/reaction suppressions for each GNG leading to the correction of 18 inconsistencies under the minimal and 19 under the YP medium. Twelve of these fixed GNGs in the minimal medium and fourteen in the YP were excluded from the list of corrected mismatches as we did not find any supporting evidence in the literature. Overall, we fixed 33 GNGs in the minimal and 28 in the YP medium by global modifications that do not conflict with any correctly predicted growth phenotype for single gene mutants. We note that 27 GNGs that appeared in both media were corrected by the same mechanisms.

Upon incorporating into the model only the corrections for NGGs and GNGs for which we found literature evidence, the accuracy of the iMM904 model was substantially improved. The number of correctly predicted lethal knockouts (i.e., specificity) out of a total of 224 was increased from 87 to 120 for the minimal medium and from 45 to 73 (out of a total of 182) for the YP medium. The corresponding false viability rate (FVR) was decreased from 61.16% to 46.43% for the minimal medium and from 75.27% to 59.89% for the YP medium. These results are summarized in Figures 2.1C and D.

2.3.2. Double gene perturbations

We employed the SL-Finder procedure [1] to identify the set of all synthetic lethal (SL) gene pairs under both minimal and YP media based on iMM9904 model of the
yeast. This analysis led to identification of 97 SL pairs in the minimal medium and 42 in the YP medium. Model SL prediction inconsistencies were identified by contrasting against the available experimentally identified SLs. As shown in Figure 2.2, this comparison reveals a number of additional ways that model and experiment may differ in their predictions. Notably, the “no growth” phenotype in this case could be due to either essentiality (ES) or synthetic lethality (SL) of the gene deletions. For example, GES and GSL inconsistencies refer to cases where the in silico deletion of a gene pair is not lethal (i.e., Growth) but in vivo they are lethal due to gene essentiality or synthetic lethality (i.e., ESsential or Synthetic Lethal). Similarly, ESG and ESSL represent mismatches where the single deletion of one of the genes in silico is lethal (i.e., ESsential), however, their simultaneous deletion in vivo results in either a viable strain (i.e., Growth) or a lethal phenotype (i.e., Synthetic Lethal), respectively. Finally, SLG and SLES denote inconsistencies where the model implies that only the double gene mutation is lethal (i.e., Synthetic Lethal) but experimental observations support either growth (G) or lethality of any of the two single gene deletions (i.e., ESsential), respectively.

![Figure 2.2](image.png)

**Figure 2.2.** Different types of mismatches between in silico predictions and in vivo observations for double gene perturbations. The abbreviations G, ES and SL in this figure refer to Growth, Essential and Synthetic Lethal, respectively. Here, ‘No Growth’ can be due to either essentiality or synthetic lethality of single or double gene deletions.
In the following we describe how the restoration of each type of these inconsistencies can be used to improve the predictive capability of the iMM904 model. In analogy to the definition of specificity for single gene mutations, we define the specificity for double gene perturbations as the fraction of correctly predicted synthetic lethals (SLSL), i.e., \( \text{Specificity} = \frac{\#\text{SLSL}}{\#\text{SLSL} + \#\text{ESSL } + \#\text{GSL}} \). We do not comment on the resolution of GES and ESG mismatches, as they are completely equivalent to GNG and NGG inconsistencies. In addition, we note that all the model corrections are based on the currently available incomplete list of experimentally identified SLs.

2.3.2.1. Resolution of GSL inconsistencies

The GSL inconsistencies denote cases where the model implies growth under the simultaneous deletion of two genes while experimental results show a lethal effect. A total of 104 GSLs in the minimal medium and 98 in the YP medium were identified. We note that GSLs for double perturbations can be treated similarly to the GNG mismatches for single perturbations with the only difference that “no growth” in GNGs is due to the essentiality of single genes whereas in GSLs it is the result of synthetic lethality.

The strategy for fixing GSLs involved in coding for isozymes is to repress the other genes coding for that isozyme under the examined condition or modify the GPR associations. When both genes code for isozymes for the same reaction this resolution strategy is viable only if that reaction is essential \textit{in silico}. Alternatively, when two genes are associated with two (or more) different reactions they must form a synthetic lethal. We note that the gene repressions for this group of GSLs can be inferred either manually by inspection of the GPR associations for each inconsistency (as we did for GNGs), or automatically, by using the modified version of the GrowMatch (see Methods). As an example, the modified GrowMatch procedure suggests that the GSL \((ASNI,ASN2)\) can be fixed by suppressing gene YML096W. By reviewing the GPR relationships we found that genes \textit{ASNI} (YPR145W) and \textit{ASN2} (YGR124W) together with YML096W provide isozymes for catalyzing the essential reaction ASNS1 (asparagine synthase (glutamine-hydrolysing)). Therefore, suppressing gene YML096W will render the simultaneous deletion of \textit{ASNI} and \textit{ASN2} lethal. Note, however, that we did not find any evidence in
support of YML096W suppression under aerobic minimal conditions, but we found that this gene is coding for a putative protein of unknown function and has been assigned to ASN1 only based on the sequence similarity [109,110,111] prompting us to remove YML096W from the GPR association to ASN1. This example demonstrates that the identified gene suppressions by GrowMatch may allude to incorrect assignment of genes to reactions.

Resolution of GSL inconsistencies coding for the blocked reactions is more complicated than for GNGs. Different cases need to be examined: (1) If both genes appear in the GPR relationship for the same blocked reaction then we revert to the method discussed for fixing GNGs. (2) If the two genes are coding for two (or more) different reactions that are all blocked because of the same root no-consumption metabolite then exploring the addition of the (common) downstream, or the corresponding root problem metabolite, to the biomass equation as a reactant may fix the inconsistency. (3) If the two genes are coding for reactions that are blocked because of different root no-consumption metabolites or one of the genes codes for a non-blocked reaction then we cannot reconcile the growth inconsistency by simply modifying the biomass reaction. To resolve such GSLs, the GapFill procedure is applied first to correct the root no-consumption metabolites by adding a consumption or export pathway to the model, thereby unblocking the coded reactions. With this modification, these GSLs can now be treated as those where deletion of each single gene will change the flux distribution in silico. We could only fix one GSL inconsistency by modifying the biomass equation: $KRE6$ (YPR159W) and $SKN1$ (YGR143W) that provide isozymes for the blocked reaction 16GS (1,6 β-glucan synthase) were fixed by including 16BDgln (1,6 β-D-glucan) a root no-consumption metabolite in the biomass equation as a reactant. This modification is corroborated by the previous reports showing that 1,6 β-D-glucan is a key component of the yeast cell wall [112,113,114].

Reconciling GSLs where disruption of each single gene affects the in silico flux distribution, involves identifying missing regulatory constraints on genes/reactions by using the GrowMatch procedure and its modified version. We followed this analysis by allowing for up to two simultaneous gene/reaction suppressions for each GSL and considered only the global modifications. For example, genes $ZWF1$ (YNL241C) and
RPE1 (YJL121C) are both involved in the pentose-phosphate pathway and form a GSL (under minimal medium). The deletion of either ZWF1 or RPE1 will change the flux distribution in silico. Application of the GrowMatch to fix this inconsistency suggests suppressing reaction G6PDH2er (glucose-6-phosphate dehydrogenase [endoplasmic reticulum]) which is identical to the reaction catalyzed by Zwf1p in the cytosol (i.e., G6PDH2). This suggests that glucose-6-phosphate dehydrogenase activity in the endoplasmic reticulum is not sufficient to compensate for its deficiency in cytosol in an RPE1 mutant background. Interestingly, it has been shown [115] that mammalian cells have two sets of enzymes catalyzing the reactions of the pentose phosphate pathway (including glucose-6-phosphate dehydrogenase) with the more active set in the cytoplasm and the less active in the endoplasmic reticulum. If this holds true for yeast then this could resolve the inconsistency. An alternative hypothesis for fixing the inconsistency is that ZWF1 independently codes for glucose-6-phosphate dehydrogenase activity in the endoplasmic reticulum (i.e., reaction G6PDH2er) as well as in the cytosol (i.e., reaction G6PDH2). In another example, FUR1 (YHR128W) and URA3 (YEL021W) form a GSL. They catalyze two reactions that serve as alternative production routes for ump (uridine mono-phosphate) a biomass precursor. Application of the GrowMatch procedure to fix this GSL suggests repressing reaction PYNP2r (pyrimidine-nucleoside phosphorylase [uracil]). According to the iMM904 model, reaction PYNP2r provides a source of uri (uridine), which subsequently is converted into ump. Suppression of PYNP2r will thus render the simultaneous deletion of FUR1 and URA3 lethal by blocking all production routes for the biomass precursor ump (see Figure 2.3). Note that since PYNP2r does not have any gene association in the model, its suppression to fix the GSL inconsistency also raises the possibility that this reaction is erroneously included in the model and should thus be removed.

Not surprisingly all the identified suppressions for correcting GSLs for the minimal medium were also valid for the YP medium. We identified only one case of a gene suppression to resolve a GSL, which was valid for only the YP medium. Genes MET12
Figure 2.3. Resolution of a GSL through reaction suppressions. (A) *FUR1* and *URA3* form a GSL since based on iMM904 there exists alternative pathways for producing the biomass precursor ump in the absence of these genes. (B) By suppressing (or removing) the reaction PYNP2r, all the production routes towards the biomass precursor ump is blocked, rendering the simultaneous deletion of *FUR1* and *URA3* lethal.

(YPL023C) and *MET13* (YGL125W), identified as a GSL in both minimal and YP media provide isozymes for reaction MTHFR3 (5,10-methylenetetrahydrofolatereductase (NADPH)), which is involved in folate metabolism. Application of the modified GrowMatch to resolve this inconsistency suggests suppression of *CYS3* (YAL012W) in both minimal and YP media. However, suppressing *CYS3* under the minimal medium will change some GGs to NGGs (i.e., it is a conditional modification) whereas, it is a global suppression in the YP medium. Interestingly, in support of *CYS3* suppression in the YP medium, it was reported before that this gene is cystein, methionine and glutathione suppressed [116].

Finally, in some cases GSL mismatches allude to incorrect GPR relationships in the model. For example, we found four GSLs (under minimal medium) involved in a GPR that implicates genes *PRS1* through 5. The first two include two genes (i.e., (*PRS1,PRS5*) and (*PRS3,PRS5*)) and the other two contain three genes (i.e.,
Figure 2.4. Change of GPR for reaction PRPPS. The GSL inconsistencies \((PRS1,PRS5), (PRS3,PRS5), (PRS1,PRS2,PRS4)\) and \((PRS3,PRS2,PRS4)\) imply that the GPR for reaction PRPPS should be changed from \((PRS1 OR PRS2 OR PRS3 OR PRS4 OR PRS5)\) to \((PRS1 AND PRS3) OR [(PRS2 OR PRS4) AND PRS5]\). \((PRS1,PRS2,PRS4)\) and \((PRS3,PRS2,PRS4)\). All these genes provide isozymes for the essential reaction PRPPS (phosphoribosylpyrophosphate synthetase) according to the iMM904 model. The inconsistency between model predictions and experimental observations can be resolved by changing the GPR relationship from \((PRS1 OR PRS2 OR PRS3 OR PRS4 OR PRS5)\) to \((PRS1 AND PRS3) OR [(PRS2 OR PRS4) AND PRS5]\) to render the simultaneous deletion of each of the aforementioned gene pairs/triple combinations lethal (see Figure 2.4). Interestingly, in a previous study a strong interaction between \(PRS1\) and \(PRS3\) as well as between \(PRS5\) and either \(PRS2\) or \(PRS4\) has been described [117]. In total, we fixed eleven GSLs in the minimal medium and five in the YP medium by converting them to SLSL consistencies by relying on literature vetted global modifications.

2.3.2.2. Resolution of ESSL inconsistencies

The ESSL inconsistencies refer to cases where the model predicts that one or both of the genes are essential even though the experimental results show that they form a SL. Based on the available data in literature, we identified 13 ESSLs containing seven \textit{in silico} essential genes in the minimal medium and nine ESSLs containing five \textit{in silico} essential genes in the YP medium. Not surprisingly, all these \textit{in silico} essential genes were part of the previously identified NGG mismatches for the single gene mutations. Consequently, a direct resolution strategy to fix the ESSLs is resolving the NGGs for the
single gene perturbations. Following this route, we found that three ESSLs in the minimal medium and one in the YP medium were automatically rectified and converted to SLSLs through the resolution of NGGs. However, the rest of them were not resolved, instead they were converted into new GSL mismatches. Except for one case, application of the methods discussed in the previous section for resolving GSL mismatches did not identify any global or literature supported correction strategies.

An interesting example of ESSL mismatches is the gene pair CHO2 (YGR157W) and OPI3 (YJR073C) where each gene is essential in vivo [14]. CHO2 catalyzes the first step in conversion of pe_SC (phosphatidylethanolamine) to pc_SC (phosphatidylcholine) in phospholipid biosynthesis pathway (reaction PETOHM_SC in the iMM904 model), whereas, OPI3 catalyzes the last two steps (reactions MFAPS_SC and PMETM_SC). We were not able to fix neither CHO2 nor OPI3 as separate NGGs by any of the methods discussed previously. However, the fact that these two genes form a SL in vivo suggests that they may act as isozymes for reactions PETOHM_SC, MFAPS_SC or PMETM_SC. Interestingly, by mining the literature we found that previous studies have already demonstrated that OPI3 can partially contribute as an isozyme in catalyzing reaction PETOHM_SC [118] implying that the GPR for this reaction should be changed to (CHO2 OR OPI3). There were no reports implicating CHO2 as an isozyme for catalyzing reactions MFAPS_SC or PMETM_SC. However, previous studies have reported that S. cerevisiae is flexible with respect to phospholipid composition and can substitute pe_SC, ptdmeeta_SC (phosphatidyl-monomethylethanolamine) or ptd2meeta_SC (phosphatidyl-dimethylethanolamine) for pc_SC to a substantial extent [118,119,120,121]. In order to capture this lack of specificity we removed pc_SC from the biomass reaction and instead added a proxy phospholipid compound with the same stoichiometric coefficient as that for pc_SC. Subsequently, we added four hypothetical reactions to the model that produce the phospholipid from any of pc_SC, pe_SC, ptdmeeta_SC or ptd2meeta_SC. Note that these modifications to the model were all global and fixed the inconsistencies for OPI3 as a NGG as well as (CHO2, OPI3) as an ESSL mismatch.
2.3.2.3. Resolution of SLG inconsistencies

The SLG inconsistencies represent a mismatch where the two genes form an *in silico* SL, however, their simultaneous deletion results in a viable strain. SLG mismatches for double gene perturbations can be viewed as NGG mismatches for single gene perturbations in the sense that they both imply that certain functionalities are missing in the model. The only difference between NGGs and SLGs is that the “no-growth” in NGGs is due to the essentiality of single genes whereas in SLGs it is the result of synthetic lethality of gene pairs. Therefore, we simply adapt the same procedure that we used for fixing NGGs to resolve SLG inconsistencies.

We found and resolved one case of such an inconsistency in both minimal and YP media. This SLG pertains to genes *PGM1* (YKL127W) and *PGM2* (YMR105C) that code for isozymes of reaction PGMT (phosphoglucomutase) involved in glycolysis/gluconeogenesis. The simultaneous deletion of *PGM1* and *PGM2* is lethal based on the iMM904 model since glycogen, a biomass precursor, cannot be produced in the absence of these two genes. Application of the GrowMatch procedure to fix this SLG, did not lead to any literature-supported correction strategy. We next mined the literature for a possible isozyme for this reaction and identified that gene *PGM3* (YMR278W) is known to catalyze the interconversion of glucose-1-phosphate to glucose-6-phosphate [122], however, it is missing in the iMM904 model. The addition of *PGM3* to the GPR for reaction PGMT (i.e., *PGM1* OR *PGM2* OR *PGM3*) renders the simultaneous deletion of *PGM1* and *PGM2* non-lethal.

2.3.2.4. Resolution of SLES inconsistencies

The SLES mismatches denote cases where the model predicts that only the simultaneous deletion of both genes is lethal whereas one of the genes (or both) is essential *in vivo*. This implies that deletion of one of these two genes cannot be compensated for by the other gene under the experimental conditions. Therefore, SLES inconsistencies are rectified by suppressing in the model the gene that is not essential *in vivo*. Essential genes participating in *in silico* SLs yield GNG inconsistencies. Therefore, the resolution of these GNGs also fixes the SLES mismatches for the double gene mutations. We identified 59 SLES mismatches for the minimal medium and 42 for the
YP medium. Fifty one of these SLEs in the minimal and two in the YP medium were fixed by using the global modifications found for GNGs. For example, gene *IPPI* (YBR011C) which codes for PPA (inorganic diphosphatase involved in oxidative phosphorylation) forms a SL pair under both minimal and YP medium with gene *IPP2* (YMR267W) that codes for the same reaction in mitochondria (reaction PPAm). However, *IPPI* has been found to be essential *in vivo* [14]. Therefore, to resolve this inconsistency, we conditionally suppress gene *IPP2* under aerobic conditions. Interestingly, by investigating the available expression data for these genes we found that the expression level of *IPPI* under aerobic conditions with glucose as the carbon source is almost 37 times higher than that for *IPP2* [123], which may explain why *IPP2* is not able to compensate for the deletion of *IPPI*.

In another example, gene *PGK1* (YCR012W), which is associated with reaction PGK (phosphoglycerate kinase) involved in glycolysis/gluconeogenesis, participates in as many as 39 *in silico* SLs under minimal medium. However, it has been found to be essential *in vivo* [14]. This implies that suppressing at least one of the 39 genes forming a synthetic lethal with *PGK1* would resolve all these SLEs. Although, we did not find any evidence confirming this resolution hypothesis for any of these genes except for *PCK1* (YKR097W), which is known to be suppressed in presence of glucose [124,125]: this gene is involved in gluconeogenesis, a process allowing yeast to synthesize glucose from non-carbohydrate precursors such as ethanol or glycerol. Notably, suppression of *PCK1* in the *iMM904* model will block production of three biomass precursors, i.e., phe-L (L-Phenylalanine), trp-L (L-Tryptophan) and tyr-L (L-Tyrosine) in the absence of *PGK1*.

Overall, the resolution of model inconsistencies for double gene deletions improved the specificity of *iMM904* model from 12.03% to 23.31% (out of a total of 133) for the minimal medium and from 6.96% to 13.04% (out of a total of 115) for the YP medium. It is worth noting that these corrections are based on only the incomplete list of SL data available in literature.

### 2.3.3. Auxotrophy inconsistencies

These mismatches refer to cases where the essentiality of single gene deletions or synthetic lethality of double gene knockouts are in agreements with *in vivo* observations,
however, the model predictions for supplementation rescue (i.e., auxotrophy) scenarios are inconsistent with experimental data. We found seven such these inconsistencies under minimal and five under YP medium, respectively, for correctly predicted essential genes, as well as four under both minimal and YP media for SLSL predictions. Notably, for all of these mismatches, the experimental results show that the single or double gene mutant strains can restore growth if additional compounds are added to the growth medium, while the model predictions imply that these genes remain essential or synthetic lethal even in the presence of these compounds. These inconsistencies can be treated in exactly the same way as the NGG or SLG mismatches were treated, since they refer to the functionalities that are missing in the model but present under experimental conditions.

As an example, it has been reported that a strain containing the \textit{FOL1} (YNL256W) deletion can grow if the medium is supplemented with folic acid [126]. Nonetheless, folic acid is not included in the list of metabolites in \textit{iMM904} model. By adding folic acid as well as exchange and transport reactions (between cytosol and extracellular environment) to the \textit{iMM904} model \textit{FOL1} remains essential even though folic acid is allowed to be taken up. GrowMatch suggested addition of any of the two reactions R00937 (5,6,7,8-tetrahydrofolate:NAD+ oxidoreductase) or R02236 (dihydrofolate:NADP+ oxidoreductase) from the KEGG database to the model so as to connect folic acid to rest of the network. Interestingly, by searching the KEGG database we found that the enzyme catalyzing these two reactions is present in yeast and the gene coding for this enzyme (\textit{DFR1}: YOR236W) is already present in the \textit{iMM904} model.

The gene pair \textit{HMG1} (YML075C) and \textit{HMG2} (YLR450W), which forms a SLSL under both minimal and YP media, is an example of auxotrophy mismatches for double gene perturbations: although \textit{in vivo} observations show that a mutant strain lacking these two genes can be rescued through addition of mev-R (mevalonate) to the growth medium [127], \textit{in silico} predictions imply that their double deletion is still lethal even in the presence of mev-R. The addition of an exchange and transport reaction between cytosol and extracellular environment for mev-R to the model resolves this auxotrophic inconsistency. Notably, the addition of this import pathway for mev-R to the model also
fixed the auxotrophy inconsistency for the essential gene \textit{ERG10} (YPL028W), which is involved in mevalonate biosynthesis.

Overall, upon including only the global modifications for which a supporting evidence was found, we could fix three auxotrophy inconsistencies for essential genes as well as one for SLs under both minimal and YP media. A summary of all the suggested modifications for the \textit{iMM904} model by using all types of inconsistencies for single and double gene perturbations is given in Figure 2.5.

### 2.4. Discussion and conclusion

We identified 120 corrections with supporting evidence to the \textit{iMM904} metabolic model of yeast by using essentiality and synthetic lethality data. Previous studies geared
towards improving the predictive ability of metabolic models have used growth phenotype inconsistencies for single gene mutation experiments in microbial systems such as *E. coli* [50,70]. Here, we go a step further by demonstrating the utility of synthetic lethality data for improving the accuracy of a multi-compartment metabolic model for a eukaryotic organism. This revealed missing or erroneously present metabolic functions in the model that could not be captured by only single gene perturbations. In addition, we found that in some cases fixing a mismatch for double gene deletions automatically fixes one or more mismatch(s) for single gene perturbations. This was illustrated for the NGG, CHO2 for which none of the mechanisms proposed to resolve the NGGs were found to be successful. Approximately, 20% of the total suggested corrections for the iMM904 model use information from both single and double gene perturbations whereas 17% of them were exclusively discerned from double gene perturbations. A far larger contribution of synthetic lethals in providing model refinement strategies is thus expected as more synthetic lethality data are becoming available.

The high number of GNG and GSL inconsistencies identified in this study runs contrary to the general perception that predictive inaccuracy of genome-scale metabolic models is primarily due to missing metabolic capabilities. It appears that the presence of not properly restricted to specific conditions functionalities in the model is the largest contributor to inconsistent predictions. Application of GrowMatch to eliminate and/or properly regulate these functionalities led to the identification of 30 growth medium-specific regulatory constraints. These growth prediction inconsistency-based constraints complement existing regulatory constraints based on gene expression data [116,128,129].

In this study we considered not only essentiality and synthetic lethality predictions but also disagreements in auxotrophy complementation. We also demonstrated that the identified growth-phenotype discrepancies are sometimes due to an incorrect or incomplete *in silico* description of the complex growth medium not the inaccuracy of the metabolic model. Overall, we significantly improved the predictive capability (i.e., specificity, sensitivity and false viability rate) of the iMM904 model for essentials and synthetic lethals by incorporating a minimum of 112 (out of 120) suggested corrections. All of these modifications are global as they do not invalidate any of the correct model predictions. The proposed corrections span a wide array of changes to the model.
including relaxation of the irreversibility constraints on existing reactions in the model, adding new reactions, compounds or genes to the model, modifying the biomass equation, changing the GPR associations and medium-specific regulatory constraints. As we found independent corroborating evidence for the proposed corrections, the vast majority (i.e., 103) of them remain relevant even for the latest update of the community yeast model (Yeast 4.0) [84]. This includes 73 model refinement strategies as well as 30 identified medium-specific regulatory constraints. Of the remaining modifications to the iMM904 model, twelve were also independently incorporated in Yeast 4.0 whereas five are different from the ones adopted in Yeast 4.0.

In addition to model refinement strategies with supporting evidence, we identified more than 60 other global modifications for which there was neither conflicting nor supporting evidence. These modifications can be treated as testable hypotheses for which experiments can be designed to prove/disprove their validity. Overall, our study demonstrates the value of bringing to bear multi-gene deletion data to further improve the predictive capability of genome-scale metabolic models. The availability of high-throughput experimental techniques [19,20,21,130] as well as efficient computational tools [1,34,35] to elucidate synthetic lethal interactions opens the door to rapidly reveal additional model deficiencies. The model refinement approaches presented in this study are versatile enough to be employed for a wider range of experimental conditions (e.g., other growth media) or synthetic lethal interactions of increasing size (e.g., triples, quadruples, etc).
Chapter 3

OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities

This chapter has been previously published in modified form in PLoS Computational Biology [131].

3.1. Introduction

Solitary species are rarely found in natural environments as most microorganisms tend to function in concert in integrative and interactive units, (i.e., communities). Natural microbial ecosystems drive global biogeochemical cycling of energy and carbon [132] and are involved in applications ranging from production of biofuels [133,134], biodegradation and natural attenuation of pollutants [135,136,137], bacterially mediated wastewater treatment [138,139] and many other biotechnology-related processes [140,141]. The species within these ecosystems communicate through unidirectional or bidirectional exchange of biochemical cues. The interactions among the participants in a microbial community can be such that one or more population(s) benefit from the association (e.g., through cooperation), some are negatively affected, (e.g., by competing for limiting resources), or more often than not a combination of both. These inter-species
interactions and their temporal changes in response to environmental stimuli are known to significantly affect the structure and function of microbial communities and play a pivotal role in species evolution [142,143,144,145,146,147].

Recent advances in the use of high-throughput sequencing and whole-community analysis techniques such as meta-genomics and meta-transcriptomics promise to revolutionize the availability of genomic information [147,148,149]. Despite the growing availability of this high-throughput data, we still know very little about the metabolic contributions of individual microbial players within an ecological niche and the extent and directionality of metabolic interactions among them. This calls for development of efficient modeling frameworks to elucidate less understood aspects of metabolism in microbial communities. Spurred by recent advances in reconstruction and analysis of metabolic networks of individual microorganisms, a number of metabolic models of simple microbial consortia have been developed. Efforts in this direction started with the development of metabolic model for a mutualistic two-species microbial community [150]. The metabolic network of each microorganism was treated as a separate compartment in analogy to eukaryotic metabolic models [84,151]. A third compartment was also added through which the two organisms can interact by exchanging metabolites. The same approach was employed for the metabolic modeling of another syntrophic association between Clostridium butyricum and Methanosarcina mazei [152]. Lewis et al [153] have also described a workflow for large-scale metabolic modeling of interactions between various cell lines in the human brain using compartments to represent different cells. Similarly, Bordbar et al [154] developed a multi-tissue type metabolic model for analysis of whole-body systems physiology. Alternatively, others proceeded to identify and model synthetic interactions among different mutants of the same species using genome-scale metabolic models. For example, Tzamali et al [155] computationally identified potential communities of non-lethal E. coli mutants using a graph-theoretic approach and analyzed them by extending dynamic flux balance analysis model of Varma and Palsson [156]. The same researchers have recently extended their study to describe the co-growth of different E. coli mutants on various carbon sources in a batch culture [157]. Wintermute and Silver [158] identified mutualistic relationships in pairs of auxotroph E. coli mutants. Each pair was modeled using an extended form of the
minimization of metabolic adjustment (MOMA) hypothesis [159]. More recently, the concept of inducing synthetic microbial ecosystems not by genetic modifications but rather with environmental perturbations such as changing the growth medium was introduced [160].

All these studies aimed primarily at modeling communities where one or both species benefit from the association while none is negatively affected. The first study to characterize a negative interaction between two microorganisms using genome-scale metabolic models was published by Zhuang et al [161] where similar to [155,157] an extension of the dynamic flux balance analysis [162] was employed to model the competition between Rhodoferax ferrireducens and Geobacter sulfurreducens in an anoxic subsurface environment. The same procedure was also employed in a study that characterized the metabolic interactions in a co-culture of Clostridium acetobutylicum and Clostridium cellulolyticum [163]. A wide range of methods beyond flux balance analysis have been used to model microbial communities [164,165,166,167,168,169,170,171,172,173,174,175]. For example, Taffs et al [176] proposed three different approaches based on elementary mode analysis to model a microbial community containing three interacting guilds. Other studies have drawn from evolutionary game theory, nonlinear dynamics and the theory of stochastic processes to model ecological systems [169,170,173].

Despite these efforts, all existing methods for the flux balance analysis of microbial communities are based on optimization problems with a single objective function (related to individual species), which cannot always capture the multi-level nature of decision-making in microbial communities. For example, the flux balance analysis model described in [150] is applicable only to syntrophic associations, where the growth of both species is coupled through the transfer of a key metabolite. The dynamic flux balance analysis models introduced by Zhuang et al [161] and Tzamali et al [155,157] rely on solving separate FBA problems for each individual species within each time interval. In all cases these methods cannot trade off the optimization of fitness of individual species versus the fitness function of the entire community. Therefore, there is still a need to develop an efficient modeling procedure to address this issue and to analyze and
characterize microbial communities of increasing size with any combination of positive and/or negative interactions.

Here, we introduce OptCom, a comprehensive flux balance analysis framework for microbial communities, which relies on a multi-level optimization description. In contrast to earlier approaches that rely on a single objective function, OptCom’s multi-level/objective structure enables properly assessing trade-offs between individual vs. community level fitness criteria. This modeling framework is general enough to capture any type of interactions (positive, negative or combination of both) for any number of species (or guilds) involved. In addition, OptCom is able to explain in vivo observations in terms of the levels of optimality of growth for each participant of the community. We first analyze a simple and well-determined microbial community involving a syntrophic association between D. vulgaris and M. maripaludis [150] to demonstrate the ability of OptCom in recapitulating known interactions. Next, OptCom is employed to model the more complex ecological system of the phototrophic microbial mats of Octopus and Mushroom Springs of Yellowstone National Park and compare our results with those obtained using elementary mode analysis [176]. OptCom identifies the level of sub-optimal growth of one of the guilds (SYN) in this community to benefit other community members and/or the entire population. Finally, we use OptCom to elucidate the extent and direction of inter-species metabolite transfers for a model microbial community [177], identifying the proportion of metabolic resources apportioned to different community members and predicting the relative contribution of hydrogen and ethanol as electron donors in the community. Addition of a new member to this community is also examined in this study.

3.2. Methods

OptCom postulates a separate biomass maximization problem for each species as inner problems. The inner problems capture species-level fitness driving forces exemplified through the maximization of individual species’ biomass production. If preferable, alternate objective function (e.g., MOMA [159]) could be utilized in the inner stage to represent the cellular fitness criteria. Inter-species interactions are modeled with appropriate constraints in the outer problem representing the exchange of metabolites.
among different species. The inner problems are subsequently linked with the outer stage through inter-organism flow constraints and optimality criteria so as a community-level (e.g., overall community biomass) objective function is optimized. Figure 3.1A schematically illustrates the proposed concept. OptCom is solved using the solution methods previously developed for bilevel programs [1,40,50,178] (see Appendix A for details of the optimization formulation and solution). Note that since OptCom yields a (non-covex) bilinear optimization problem, all case studies presented in this paper were solved using the BARON solver [179], accessed through GAMS, to global optimality.

It is important to note that OptCom can be readily modified to account for the case when one or more organisms show a form of cooperative behavior that benefits the whole population, but comes at the expense of growing at rates slower than the maximum possible [146,180]. To quantify the deviation of community members from their optimal behavior, we introduce a metric called optimality level for each species $k$ (i.e., $c^k$). The optimality level for each one of the microorganisms is quantified using a variation of OptCom which we refer to as descriptive. Descriptive OptCom incorporates all available experimental data for the entire community (e.g., community biomass composition) as constraints in the outer problem and all data related to individual species as constraints in the respective inner problems while allowing the biomass flux of individual species to fall below (or rise above) the maxima ($v_{opt}^{k\text{biomass}}$) of the inner problems (see Figure 3.1B). We note that here the optimum biomass flux for each species ($v_{opt}^{k\text{biomass}}$) is community-specific as it is computed in the context of all microorganisms striving to grow at their maximum rate (using the formulation given in Figure 3.1A). An optimality level of less than one for a microorganism $k$ implies that it grows sub-optimally at a rate equal to $100c^k\%$ of the maximum ($v_{opt}^{k\text{biomass}}$) to optimize a community-level fitness criterion while matching experimental observations. Alternatively, an optimality level of one implies that microorganism $k$ grows exactly optimally at a rate equal to $v_{opt}^{k\text{biomass}}$ whereas a value greater than one indicates that it achieves a higher biomass production level than the community-specific maximum (i.e., super-optimality) by depleting resources from one or more other community members. It is worth noting that super-optimality is achievable for a species only at the expense of sub-optimal behavior of at
A separate biomass maximization problem is defined for each species as inner problems. These inner problems are then integrated in the outer stage through the inter-organism flow constraint to optimize a community-level objective function. (B) Structure of the Descriptive OptCom to determine the optimality level of each species ($c^k$), given a set of experimental data. The available experimental data for the entire community and the individual species are described using constraints in the outer and inner problems, respectively, whereas, sub- or super-optimal behavior of each microorganism is captured by using a constraint for the respective inner problem.
least one other member in the community. The identified combination of sub- and/or super-optimal behaviors of individual species is driven by the maximization of a community-level criterion (e.g., maximize the total community biomass).

OptCom can capture various types of interactions among members of a microbial community. Symbiotic interactions between two (or more) populations can be such that one or more species benefit from the association (i.e., positive interaction), are negatively affected (i.e., negative interactions), or combination of both. Mutualism, synergism and commensalism are examples of positive interactions, whereas parasitism and competition are examples of negative interactions. A pictorial representation of how these interactions can be captured within OptCom by appropriately restricting inter-organism metabolic flows is provided in Figure 3.2 (see Appendix A for implementation details).

3.3. Results

3.3.1. Modeling a mutualistic microbial community

We first explore the capability of OptCom to model and analyze a relatively simple and well-characterized syntrophic association between two microorganisms, namely *Desulfovibrio vulgaris* Hildenborough and *Methanococcus maripaludis*. Syntrophy is a mutualistic relationship between two microorganisms, which together degrade an otherwise indigestible organic substrate. A prominent example of syntrophic interactions is interspecies hydrogen transfer, where the hydrogen produced by one of the species has to be consumed by the other to stimulate the growth of both microorganisms [181,182,183,184]. In these communities degradation of a substrate by fermenting bacteria is energetically unfavorable as it carries out a reaction, which is endergonic under standard conditions. However, if this fermenting bacteria is coupled with a hydrogen scavenging partner such as methanogenic bacteria, the organic compound degrading reaction can proceed [185]. Methanogens use hydrogen and energy gained from the first reaction and reduce CO$_2$ to methane [183,185].

Here we focus on such a syntrophic association between *Desulfovibrio vulgaris* Hildenborough and *Methano- coccus maripaludis* S2, for which genomes-scale metabolic models as well as experimental growth data for the co-culture are available [150]. With
Figure 3.2. Pictorial illustration of the customized OptCom for various types of interactions. OptCom (top panel) can be readily customized for each type of interaction through properly adjusting the inter-organism flow constraints as demonstrated for a typical microbial community composed of two interacting members.
lactate as the sole carbon source and in the absence of a suitable electron acceptor for the sulfate reducer, *M. maripaludis* provides favorable thermodynamic conditions for the growth of *D. vulgaris* by consuming hydrogen and maintaining its partial pressure low. Stoylar *et al* [150] modeled this microbial community as a multi-compartment metabolic network and employed FBA to identify community-level fluxes by maximizing the weighted sum of the biomass fluxes of two microorganisms.

Figure 3.3. Comparison of the predicted metabolic activities during the syntrophic growth with experimental data. Experimentally determined (gray diamond) and predicted production fluxes by OptCom (black square) for (A) acetate, (B) carbon dioxide (C) methane and (D) total community biomass in the syntrophic growth of *D. vulgaris* and *M. maripaludis*. All experimental data were obtained through personal communications with authors of [150]. A separate simulation was performed for each time interval wherein lactate uptake and hydrogen evolution rates were fixed at their experimentally determined values for that interval. Error bars for experimental values indicate the bounds of 95% confidence intervals [150]. The error bars for OptCom predictions were calculated by performing the simulations on the upper and lower bounds of the 95% confidence intervals for measured lactate and hydrogen flux rates.
3.3.1.1. Comparing the OptCom predictions with experimental results

First, we examined whether our model is capable of predicting the relative abundance of species (i.e., composition) in the community by maximizing the community biomass as the outer problem objective function. Each microorganism was allowed to maximize its own biomass yield in the inner problems. Consistent with Stoylar et al [150], the lactate uptake rate was set to 48 μM/h and formate and hydrogen accumulation were set to zero, so as all formate and hydrogen produced by *D. vulgaris* is utilized by *M. maripaludis*. Lower and upper bounds on all other reactions (except for the uptake and export fluxes of the shared metabolites) were taken from [150]. The ratio of the biomass yields for *D. vulgaris* and *M. maripaludis* was predicted to be 2.28 based on our simulations. This is consistent with *in vivo* observation that *D. vulgaris* dominates in the co-culture by a ratio of at least 2:1 [150]. Throughout this and the following studies we assume that the biomass flux for each species is proportional to its biomass abundance in the community.

We next explore how well OptCom performs in predicting various metabolic activities across different stages of syntrophic growth. To this end, we applied OptCom for each time interval and compared the model predictions for acetate, methane and carbon dioxide evolution rates as well as total biomass production rates with experimental measurements [150]. A separate run was performed for each time interval where lactate uptake and hydrogen evolution rates were fixed at their experimentally determined values in that interval [150]. The results of this comparison are illustrated in Figure 3.3. We can see that OptCom predictions are generally in good agreement with experimental data especially for the acetate and methane production rates. The predicted CO₂ evolution rate, however, is lower in all time intervals (except for 62-76 hr) than the measured values. Between 62 hr and 76 hr the experimental data show that the CO₂ evolution rate is close to zero, which may indicate that all CO₂ produced by *D. vulgaris* is consumed by *M. maripaludis* [150]. In addition, OptCom predicts an increase in the biomass production of the whole community over time with increasing lactate uptake rate as expected, although, all of predicted yields are higher than experimental measurements. This inconsistency could be due to missing regulatory information, incorrect modeling of ATP utilization and maintenance energy requirements and/or the presence of futile cycles.
in the metabolic models of one or both species. It is worth noting that all predictions by Stolyar’s multi-compartment approach are also very close to the results obtained by OptCom. This is because in this syntrophic microbial community the growth of both species is coupled and uniquely dependent on the exchange of hydrogen and/or formate. This allows for a single fitness function to describe the behavior of the entire community.

3.3.1.2. The role of hydrogen and formate in interspecies electron transfer

Hydrogen and formate are primary shuttle compounds for interspecies electron transfer. There are two enzymes in *D. vulgaris* that are involved in production of hydrogen and formate namely pyruvate oxidoreductase and pyruvate-formate lyase [150,186]. While both of these enzymes convert pyruvate to acetyl-CoA, the former produces reduced ferredoxin, which is then used for hydrogen production, whereas the latter produces formate, which can be secreted to the medium. For an uptake rate of 10 μmol/hr, OptCom predicts that a total of 18.6 μmol/hr of electron transfer in the form of hydrogen and/or formate transfer are required to achieve the maximum growth for both species and community. To investigate the relative contribution of formate and hydrogen in interspecies electron transfer, we examined what portion of the total required electron transfer could be carried by hydrogen or formate while maintaining the maximum biomass yield for both species. This analysis showed that hydrogen could be used as the sole electron carrier to support the maximum growth for both microorganisms even if no formate is secreted by *D. vulgaris*. Formate, on the other hand, could only account for up to 26% (4.9 μmol/hr) of the total electron transfer to maintain the biomass productions at their maximum. In addition, OptCom results show that formate exchange rates of more than 5.5 μmol/hr (~30%) are not able to support growth for any of the two species. Using OptCom we find that *D. vulgaris* is unable to produce sufficient formate to meet the minimum electron transfer required to maintain the redox balance in the absence of hydrogen.

When hydrogen production by *D. vulgaris* is constrained to be at most 13.7 μmol/hr (i.e., the rest of 4.9 μmol/hr electron transfer is assumed to be carried out by formate if possible), OptCom predictions show that in a co-culture consisting of *D. vulgaris* and a mutant of *M. maripaludis* the growth rate of both *D. vulgaris* and *M.
maripaludis is reduced by 26%. The simulation results also show that no formate is produced by D. vulgaris in this case, which was expected, as it cannot be consumed by the M. maripaludis mutant. Despite no formate production by D. vulgaris, OptCom reveals that the flux through pyruvate formate lyase is higher compared to the community having the wild-type strains. Further investigation of the in silico flux distributions shows that the entire amount of formate produced by the pyruvate formate lyase reaction is directed towards CO₂ production. This in turn results in an increased consumption of CO₂ by the M. maripaludis mutant and consequently a lower accumulation of CO₂ in the extracellular environment compared to the community with the wild-type strains. The predictions by OptCom for the community with mutant of M. maripaludis are in agreement with experimental results by Stolyar et al [150] who established a syntrophic association between D. vulgaris and the M. maripaludis mutant MM709 lacking the two annotated formate dehydrogenase enzymes. It was observed that this co-culture is able to grow, confirming that hydrogen alone can support the syntrophic growth of both species. Nevertheless, the growth rate, biomass yield and lactate uptake rates were lower compared to the syntrophic growth between the wild-type strains [150]. Notably, OptCom predictions suggest that if the wild-type D. vulgaris in Stolyar’s experiment is replaced with a mutant lacking pyruvate-formate lyase, so as all electron equivalent is produced in the form of hydrogen, then the co-culture should be able to restore growth to that of wild-type species community as hydrogen alone can carry all required electron equivalents.

3.3.2. Assessing optimality levels in a phototrophic microbial community

Here we examine the applicability of OptCom for modeling a more complex microbial community containing three interacting guilds, the phototrophic microbial mats of Octopus and Mushroom Springs of Yellowstone National Park (Wyoming, USA) [187]. The inhabitants of this community include unicellular cyanobacteria related to Synechococcus spp (SYN), filamentous anoxygenic phototrophs (FAP) related to Chloroflexus and Roseiflexus spp and sulfate-reducing bacteria (SRB) as well as other prokaryotes supported by the products of the photosynthetic bacteria [176,187]. Diel (day-night) variations in metabolic activities of members of this community were
observed before [188,189,190]. During the day when the mat is oxygenated cyanobacteria appear to be the main carbon fixer, consuming CO$_2$ and producing storage products such as polyglucose as well as O$_2$ as a by-product of photosynthesis. High levels of O$_2$ relative to CO$_2$ stimulate the production of glycolate. Glycolate is then used as a carbon and energy source by other community members such as photoheterotrophic FAP. At night, the mat becomes anoxic and cyanobacteria start to ferment the stored polyglucose into small organic acids such as acetate, propionate and H$_2$. FAP can incorporate fermentation products photoheterotrophically while SRB oxidizes the fermentation products under anaerobic condition and produces sulfide [187,191,192,193]. A schematic diagram representing the interactions in this community is given in [176].

This microbial community has been previously modeled and analyzed by Taffs et al [176] using a representative microorganism for each guild: Oxygenic photoautotrophs related to *Synechococcus* spp were chosen to represent the mat’s primary carbon and nitrogen fixers. FAP from the family *Chloroflexaceae*, were selected to represent metabolically versatile photoheterotrophs that capture light energy as phosphodiester bonds but require external reducing equivalents and carbon sources other than CO$_2$. A SRB guild representative whose metabolic behavior was based on several well-studied sulfate-reducing bacteria was also included in the community model description [176]. The metabolic networks representing central carbon and energy metabolism for each guild were then constructed and three different modeling approaches based on the elementary mode analysis were employed to elucidate sustainable physiological properties of this community [176]. Here, we focus only on daylight metabolism (for which more experimental data is available) to assess the efficacy of OptCom in describing carbon and energy flows and the biomass ratio between guilds.

**3.3.2.1. Analysis of the daylight metabolism**

The relative abundance of various species in a microbial community (i.e., composition) is of significant ecological importance. The ratio of cyanobacterial (SYN) to FAP biovolumes in a Mushroom Spring mat was determined experimentally to be 1.6:1 [194]. It was assumed that biomass formation rates and biovolume of species in the community are directly related [176]. In another study the biomass ratio in the top 1 mm
of Octopus and Mushroom Spring mats was estimated to range from 1.5:1 to 3.5:1 based on the relative abundances of metagenomic reads [176]. We used OptCom to model this community postulating that each guild strives to maximize its biomass and examined if the biomass ratio of SYN/FAP can be correctly predicted. We chose as the outer problem objective function to maximize the total community biomass (i.e., SYN biomass + FAP biomass + SRB biomass). During the day O₂ competes with CO₂ for the rubisco active site, leading to production of glycolate (O₂ + ribulose − 5 − P + ATP → glycolate + triose phosphate + ADP) instead of additional reduced carbon (CO₂ + ribulose−5−P + ATP → 2 triose phosphate + ADP) [176]. The flux ratio of these two reactions (O₂/CO₂) was measured for the Octopus and Mushroom Spring microbial mats and reported to vary approximately between 0.03 and 0.07 [176,195]. We incorporated this information into our modeling framework by fixing the flux ratio of these reactions at different values between 0.03 and 0.07 (using a constraint in the inner problem of SYN). Lower and upper bounds on all reactions (except for the uptake and export fluxes of the shared metabolites) were taken from [176]. Under these conditions, the SYN/FAP biomass ratio was predicted to range from 7.94 (for O₂/CO₂ flux ratio = 0.07) to 20.26 (O₂/CO₂ flux ratio = 0.03), which are significantly higher than the experimentally determined values of 1.5 to 3.5. This suggests that the reason for the discrepancy in prediction may be that the SYN guild does not maximize its biomass. Therefore, we decided to test this hypothesis by using the descriptive mode of the OptCom procedure (see Figure 3.1B) and establish the optimality level of SYN and other members of this community. To this end, we added a constraint to the outer problem to fix the SYN/FAP biomass ratio at different values in the experimentally observed range (1.5 to 3.5). The objective function of the outer problem was assumed to be maximization of the total community biomass. We determined the optimality levels across different values of SYN/FAP biomass and O₂/CO₂ flux ratios in their experimentally determined ranges (see Figure 3.4). OptCom finds that the observed SYN/FAP biomass ratios are consistent with SYN guild growing sub-optimally at 61-82% of its community-specific maximum with lower values corresponding to higher O₂/CO₂ flux ratios (see Figure 3.4A). On the other hand, FAP guild appears to benefit from this sub-optimal behavior of SYN by growing at rates,
Figure 3.4. Optimality levels for the SYN and FAP guilds and their effect on the total community biomass. Optimality levels for (A) SYN and (B) FAP as a function of the SYN/FAP biomass ratio across different values of the O$_2$/CO$_2$ flux ratio (C) Comparison of the predicted total community biomass (1/h) for the case when SYN grows sub-optimally and when it grows optimally. Note that, to compute the total community biomass when SYN grows optimally only O$_2$/CO$_2$ flux ratio was fixed at values in the experimentally determined range (i.e., 0.03 to 0.07), whereas for all other cases, in addition to O$_2$/CO$_2$ flux ratio, SYN/FAP biomass ratio was also fixed at values measured experimentally (i.e., 1.5 to 3.5). Lower and upper dashed lines in (C) represent the maximum and minimum predicted community biomass (when SYN grows sub-optimally) across various SYN/FAP biomass ratios.

which are approximately 4.5 to 8.5 times higher than its community-specific maximum (see Figure 3.4B).

SYN grows sub-optimally in this community to benefit other community members (e.g., FAP) and optimize a community-level fitness criterion (e.g., maximize the total community biomass). We investigated the effect of sub-optimal growth of the SYN guild on the total community biomass production across different values of SYN/FAP biomass
and O$_2$/CO$_2$ flux ratios (see Figure 3.4C). As illustrated in Figure 3.4C, at higher O$_2$/CO$_2$ flux ratios, the total community biomass is higher compared to the case when SYN grows optimally. The metabolic reason for this lower growth of SYN is that fixing more carbon (manifested by 3-7 times more predicted glycolate and acetate production) to supply other guilds and increase the overall community biomass imposes extra energy demands on the SYN guild. In contrast, for low O$_2$/CO$_2$ flux ratios the maximum community biomass when SYN grows sub-optimally is lower compared with when it grows optimally (i.e., both dashed lines lie below the solid line in Figure 3.4C). A possible reason for this discrepancy is that the experimental measurements for SYN/FAP biomass ratio were performed when the O$_2$/CO$_2$ flux ratio was high. This could also be a consequence of the experimental underestimation of glycolate production due to consumption of radio-labeled photosynthate during incubation as stated in [176]. Alternatively, SYN may grow sub-optimally so that it can divert some resources towards polysaccharide production to fuel night-time maintenance energy and morning nitrogen fixation. This is another type of a cooperative behavior by SYN.

Notably, two different cases were considered by Taffs et al [176] using the elementary modes and compartmentalized approach: a selfish criterion where each guild attempts to maximize its own biomass and an altruistic criterion where the guilds strive to maximize the total community biomass. It was concluded that predictions using the first criterion are in better agreement with experimental data. OptCom, on the other hand reveals that a trade-off between these two criteria appears to be driving the metabolism in this community. While some guilds strive to maximize their own growth, others (e.g., SYN) grow sub-optimally to maximize the biomass of entire community or benefit the nighttime metabolism, or a combination of both, depending on O$_2$/CO$_2$ flux ratio and environmental conditions.

**3.3.3. Elucidating trophic and electron accepting interactions in sub-surface anaerobic environments**

In a recent study, Miller et al [177] established a model microbial community to better understand the trophic interactions in sub-surface anaerobic environments. This community was composed of three species including *Clostridium cellulolyticum*,
*Desulfovibrio vulgaris* Hildenborough, and *Geobacter sulfurreducens*. Cellobiose was provided as the sole carbon and energy source for *C. cellulolyticum* whereas the growth of *D. vulgaris* and *G. sulfurreducens* were dependent on the fermentation by-products produced by *C. cellulolyticum*. *D. vulgaris* and *G. sulfurreducens* were supplemented with sulfate and fumarate, respectively, as electron-acceptors to avoid electron acceptor competition [177]. The experimental measurements for the biomass composition of the community showed that, as expected, *C. cellulolyticum* was the dominant member in the co-culture and confirmed the presence of *D. vulgaris* and *G. sulfurreducens*. It was, however, not possible to quantify experimentally the flow of shared metabolites among the community members as their concentrations were below the detection limits. Therefore, the authors proposed an approximate model of the carbon and electron flow based on some measurements of the three species community at steady-state, pure culture chemostat experiments and data from the literature [177].

Here, we model this microbial community by making use of the corresponding bacterial metabolic models and employ OptCom to elucidate the inter-species interactions. The metabolic models of *C. cellulolyticum* (i.e., iFS431) and *G. sulfurreducens* were reconstructed by Salimi et al [163] and Mahadevan et al [196], respectively. A basic metabolic model of *D. vulgaris* containing 86 reactions was introduced by Stolyar et al [150], however, this model had only a compact representation of the central metabolism. For example, the model was not able to support growth in the presence of acetate or ethanol as the sole carbon source. Therefore, we expanded this model by adding new reactions from a first draft reconstructed model in the Model Seed [197] and the KEGG database [96] using the GrowMatch procedure [50]. The updated model of *D. vulgaris* consists of 145 reactions and is capable of supporting growth on acetate as well as ethanol.

### 3.3.3.1. Fumarate consumption by *G. sulfurreducens*

FBA simulations showed that the metabolic model for *G. sulfurreducens* [196] is not able to capture the experimental observation that the amount of fumarate consumed is higher than the amount of succinate produced. In addition, the model predicts that no malate is produced under the examined conditions. An inspection of the metabolic model
of *G. sulfurreducens* revealed that the only included uptake pathway for fumarate is through mutual dicarboxylic acid transporter (fumarate[e] + succinate[c] ↔ fumarate[c] + succinate[e]) implying that the amount of succinate produced must be equal to the amount of fumarate consumed. Interestingly, in support of the observations by Miller *et al* [177], a recent study [198] has confirmed that the fumarate consumption rate by *G. sulfurreducens* is higher than the succinate production rate and demonstrated using $^{13}\text{C}$-based metabolic flux analysis that fumarate can be used as an additional carbon source through the TCA cycle where it is converted to malate by fumarase, and oxaloacetate via malate dehydrogenase. These findings suggest that the *dcu* gene family (responsible for the uptake of dicarboxylates such as fumarate) in *G. sulfurreducens* may have a dual function, i.e., they can act both mutually (with exchange of another compound such as succinate) or independently (i.e., protonated), similarly to those in *E. coli* [199]. This was verified by performing a bi-directional BLAST analysis that revealed high sequence similarity between the *dcu* gene families in *G. sulfurreducens* and *E. coli*. It is worth noting that addition of an alternative succinate transporter to the model could also have been another way of explaining the experimental data, however this hypothesis was not supported by the BLAST analysis. Therefore, in the absence of any other experimental data, we decided to add a protonated transport reaction for fumarate to the model. In our simulations we restricted the flux of this reaction to 15.5% of the fumarate transfer by dicarboxylic acid transporter based on the metabolic flux data under electron acceptor limited conditions [198].

**3.3.3.2. Uncovering the inter-species metabolite transfers in the community**

While the relative molar abundance of each species was measured experimentally by Miller *et al* [177], the metabolite flows across community members were untraceable. We thus chose to use OptCom to gain insight into inter-species metabolite trafficking. To this end, we employed the descriptive mode of OptCom (see Figure 3.3.1B) first to establish the optimality levels of species participating in this community, by fixing the biomass composition of the community at the values obtained experimentally by adding constraints to the outer problem. The objective function of the outer problem was maximization of the total community biomass. Descriptive OptCom revealed that the
experimentally determined biomass composition in this community was consistent with optimal growth for all microorganisms (i.e., optimality level of one for all species involved). Upon verifying that biomass maximization was driving metabolism in this community, we used OptCom to make predictions about inter-organism flow rates with a basis of 1 mole/gDW.hr of cellobiose uptake by *C. cellulolyticum* so that we can directly compare our results with the estimates in Miller et al [177]. The lower bound and upper bounds on all reactions (except for the uptake and export fluxes of the shared metabolites) were taken from the publications of the respective metabolic models [150,163,196]. Because *D. vulgaris* has a much more efficient enzymatic process for hydrogen consumption than *G. sulfurreducens*, we initially allowed *G. sulfurreducens* to take up only a small portion (between 1 to 10%) of the total hydrogen produced by *C. cellulolyticum*. However, the total predicted acetate and CO$_2$ accumulation in the extracellular environment deviated significantly from the experimental observations by Miller et al [177]. Therefore, we decided to perform the remaining simulations assuming that *D. vulgaris* consumes all hydrogen produced by *C. cellulolyticum* (even though this may not be the only way of reconciling model predictions and the experimental data). OptCom found that under these conditions 1 mol/gDW.hr of cellobiose leads to 2.48 moles/gDW.hr of acetate and 3.22 moles/gDW.hr of CO$_2$ in the extracellular environment which agree well with 2.7 and 3.3 moles/gDW.hr of acetate and CO$_2$, respectively, observed in the supernatant of the bioreactor (per mole of cellobiose) by Miller et al [177]. We note, however, that the predicted level of acetate production by *C. cellulolyticum* metabolic model (1.65 mol/gDW.hr) is lower than what was estimated in Miller’s model (2.9 mol/gDW.hr). In general, however, the predicted allocation of metabolic resources to different members of the community by OptCom is in good agreements with estimations in Miller [177] (see Figure 3.5). For example, OptCom suggests that about 13% of the acetate produced by *C. cellulolyticum* is directed towards *G. sulfurreducens*, which is very close to the 15.5% value estimated in [177].

OptCom results also show that hydrogen and ethanol produced by *C. cellulolyticum* can be completely utilized by *D. vulgaris* to reduce sulfate to hydrogen sulfide. A rough estimate for the ratio of hydrogen to ethanol, which serve as electron donors for *D. vulgaris* to reduce sulfate to hydrogen sulfide. A rough estimate for the ratio of hydrogen to ethanol, which serve as electron donors for *D. vulgaris* to reduce sulfate to hydrogen sulfide.
Figure 3.5. Comparison of the predicted fluxes by OptCom with estimates in the proposed model of [177]. The total predicted acetate and CO₂ production rates by OptCom are in good agreement with experimental measurements by Miller et al [177]. Note that it was not possible to determine experimentally how much of the total acetate or CO₂ available in the supernatant of the bioreactor is produced by which microorganism (the values provided by Miller et al [177] for the acetate and CO₂ production by each species as well as all inter-organism flow rates are estimates and not experimental measurements). The values associated with the biomass of each microorganism represent fluxes (1/h) for OptCom predictions and concentrations (M) for experimental measurements [177].

_vulgaris_, is given in by Miller et al [177] (H₂/Ethanol = 20) based on the pure culture data under similar conditions. The simulations with OptCom using genome-scale metabolic models of the community members, however, indicate a much higher contribution of ethanol in inter-species electron transfer (H₂/Ethanol = 2.34). We performed a flux variability analysis to see if this ratio can change under the examined condition, while maintaining the maximum community biomass, but no changes in this ratio were possible. This suggests that under the observed experimental condition, a H₂/Ethanol ratio of 2.34 is needed to support the maximum growth for each species as well as for the community as a whole. While acetate serves as the only carbon substrate for both _G. sulfurreducens_ and _D. vulgaris_, it was not possible to determine experimentally if _D. vulgaris_ directly uses the available acetate in the medium released...
by *C. cellulolyticum* or it derives acetate from ethanol. OptCom results support the latter scenario (see Figure 3.5). This is more likely to happen because acetate is already available internally to *D. vulgaris* from the cytosolic oxidation of ethanol. OptCom also identifies that 77.6% of the converted ethanol to acetate is secreted to the medium by *D. vulgaris*, while the rest is incorporated into biomass (see Figure 3.5). This is in good agreement with the estimate by Miller *et al* [177] suggesting that *D. vulgaris* does not consume any acetate produced by *C. cellulolyticum* and that it exports 62.5% of the assimilated ethanol to the medium as acetate. Elucidation of the metabolic interactions among the members of this community was achieved by OptCom after verifying that all species appear to grow optimally based on the *in vivo* observations for the community biomass composition.

### 3.3.3.3. Addition of a new member to the microbial community

As mentioned earlier, 2.48 moles/gDW.hr of acetate was predicted to be available in the extracellular environment (per mole of cellobiose consumed) which could be utilized by other trophic anaerobic bacteria [177]. Therefore, an acetate utilizing methanogen such as *Methanosarcina* species, which are known to be avid consumers of acetate, can be envisioned as an additional member of this community. We chose *Methanosarcina barkeri* for this analysis as its metabolic model has been reconstructed by Feist *et al* [200]. Another inner problem was added to the OptCom to account for addition of *M. barkeri* to this community. Consistent with other community members the objective function for this inner problem was to maximize the biomass flux of *M. barkeri*, whereas the objective function of the outer problem was to maximize the total community biomass. The acetate uptake rates by *G. sulfurreducens* and *D. vulgaris* were fixed at the values obtained by OptCom for the tri-culture. *D. vulgaris* and *M. barkeri* were suggested to compete in anoxic environments for hydrogen [201], however, we assumed that all H₂ produced by *C. cellulolyticum* is consumed by *D. vulgaris*, as it has been reported to have much more favorable kinetic parameters for H₂ metabolism than methanogens [202,203,204]. In addition, it was demonstrated that *Methanosarcina* species can not only consume but also produce hydrogen when growing on organic substrates such as acetate [205,206]. Therefore, we allowed *D. vulgaris* to consume the
hydrogen produced by *M. barkeri* (if any) in addition to that produced by *C. cellulolyticum*.

The biomass flux of *M. barkeri* is strongly dependent on the value of growth-associated maintenance (GAM), which was found to be a function of the proton translocation efficiency of the Ech hydrogenase reaction [200]. The range of GAM values for 0.2-2 protons translocated/2e\(^-\) that result in a growth yield consistent with *in vivo* observations was computed by Feist *et al* [200]. Here, we examined the variability in growth yields and relative abundance of *M. barkeri* in the tetra-culture community across different GAM values associated with 0.2-2 protons translocated/2e\(^-\). This analysis showed that *M. barkeri* is capable of consuming the entire 2.48 moles of acetate produced by *C. cellulolyticum* and *D. vulgaris*. Depending on the GAM value and the proton translocation efficiency, *M. barkeri* was predicted to constitute 2.5 to 10.4% of the total community biomass (assuming that the biomass fluxes are proportional directly with the abundance levels of species in the community) with the other three members growing at rates similar to the ones obtained for the tri-culture. *C. cellulolyticum* still dominates the co-culture as before with biomass fractions ranging from 69.6 to 75.7% (depending on *M. barkeri*’s biomass flux). The methane evolution rate by *M. barkeri* was predicted by OptCom to range from 2.36 to 2.45 moles/gDM.hr. It is important to note that previous studies have reported that the internal carbon and electron flow of *M. barkeri* could be altered by *D. vulgaris* in a co-culture grown on an organic substrate such as acetate, [207]: It was suggested that *D. vulgaris* strives to keep the partial pressure of hydrogen low enough to shift the catabolic redox system of methanogen so that more H\(_2\) is produced by *M. barkeri* (compared to pure cultures) and more acetate is oxidized to CO\(_2\) instead of methane [207]. Even though we allowed *D. vulgaris* to take up all hydrogen produced by *M. barkeri* (in addition to that produced by *C. cellulolyticum*), no such shift in methanogenesis was observed for the tetra-culture according to the OptCom predictions. A possible reason might be that enough hydrogen (as well as ethanol) is already available to *D. vulgaris* from *C. cellulolyticum*, obviating the need to alter methanogenesis in order to gain the reducing equivalents. This hypothesis is supported by the experimental observation that if excess H\(_2\) is added to the co-culture of *M. barkeri*.
and *D. vulgaris*, it is completely consumed by *D. vulgaris* and the acetate catabolism by *M. barkeri* is no longer affected [207].

Even though 3.22 moles/gDW.hr of CO$_2$ produced by *C. cellulolyticum* and *G. sulfurreducens* is available in the medium, OptCom predicts that it remains completely unused in the tetra-culture. This was expected as growth of *M. barkeri* on CO$_2$ relies on presence of hydrogen, which we assumed that it was consumed completely by *D. vulgaris*. In order to examine if *M. barkeri* is indeed capable of utilizing the available CO$_2$ as a carbon source (in addition to acetate), we temporarily allowed *M. barkeri* to take up the hydrogen produced by *C. cellulolyticum*. For this case, OptCom revealed that if the entire hydrogen produced by *C. cellulolyticum* is available to *M. barkeri*, it can support growth on CO$_2$ only for proton translocation efficiencies of less than one/2e$^-$. Notably, for proton translocation efficiencies of more than one, even though no CO$_2$ is assimilated by *M. barkeri*, OptCom shows that the availability of hydrogen will lead to an increase in the methane production by about 26-28%.

### 3.4. Discussion and conclusion

Here, we introduced OptCom, a comprehensive computational framework for the flux balance analysis of microbial communities using genome-scale metabolic models. We demonstrated that OptCom can be used for assessing the optimality level of growth for different members in a microbial community (i.e., Descriptive mode) and subsequently making predictions regarding metabolic trafficking (i.e., Predictive mode) given the identified optimality levels. Unlike earlier FBA-based modeling approaches that rely on a single objective function to describe the entire community [150,160] or separate FBA problems for each microorganism [155,157,161,163], OptCom integrates both species- and community-level fitness criteria into a multi-level/objective framework. This multi-level description allows for properly quantifying the trade-offs between selfish and altruistic driving forces in a microbial ecosystem. Species and community level fitness functions are quantified by maximizing the biomass formation for the respective entity. We note, however, that the physiology of microbial communities is highly context and environment dependent and a universal community-specific fitness criterion does not exist. Studies similar to those conducted for mono-cultures that examine and compare
various presumed hypotheses on cellular objective function [208,209,210,211,212,213] or algorithms that identify/test a relevant objective function using experimental flux data [214,215] are needed in the context of multi-species systems.

An important goal of studying natural and synthetic microbial communities is their targeted manipulation towards important biotechnological goals (e.g., cellulose degradation, ethanol production, etc.). This has motivated researchers to construct simple synthetic microbial ecosystems, which are amenable to genetic and engineering interventions, for biotechnology- and bioenergy-related applications. As an example, Bizukojc et al [152], have proposed a co-culture composed of Clostridium butyricum and Methanosarcina mazei to relieve the inhibition of fermentation products and increase production of 1,3-propanediol (PDO) by Clostridium butyricum. Mixed cultures have been also established for overproduction of polyhydroxyalkanoates (PHA) [216,217] and ethanol [218,219,220,221,222]. For example, Clostridium thermocellum, which is used for ethanol production, has been found to be capable of utilizing hexoses, but not pentose sugars generated from breakdown of cellulose and hemicellulose [222]. Therefore, cultivation of C. thermocellum with other thermophilic anaerobic bacteria capable of utilizing hexoses as well as pentose to produce ethanol (e.g., Clostridium thermosaccharolyticum and Thermoanaerobacter ethanolicus) has been previously examined in vivo [218,219,220,221,222]. The multi-objective and multi-level structure of the OptCom procedure, introduced here, can help assess the metabolic capabilities of such synthetic ecosystems. Taking a step further, OptCom can be readily modified to identify the minimal number of direct interventions (i.e., knock-up/down/outs) to the community leading to the elevated production of a desired compound (e.g., by considering the overproduction of desired compound as the outer problem objective function), thus extending the applicability of strain design tools such as OptKnock [40], OptStrain [178], OptReg [223] and OptForce [224]. It is worth noting that a key bottleneck to the modeling and analysis of microbial communities is the paucity of genome-scale models for all participants in a complex microbial community. Overcoming this barrier would require the development of high-throughput metabolic reconstruction tools such as the Model Seed [197] resource. Given that microbial communities change with time (e.g., day/night cycle) and also location (e.g., nutrient
gradients), approaches that would be able to capture temporal and spatial varying inter-
species metabolic interactions are needed. For example, the separate FBA problems for
each individual species in the dynamic flux balance analysis methods of Zhuang et al
[161] and Tzamali et al [155,157] can be integrated with OptCom to account for inter-
species interactions and community-level fitness driving forces within each time interval.
Chapter 4

Metabolic Interventions for Overproduction of L-serine in *Escherichia coli* Using the OptForce Procedure

Sridhar Ranganathan served as a collaborator in this project.

4.1. Introduction

L-serine is an amino used extensively in pharmaceutical and cosmetic industry [225] with an estimated annual demand of 300 tons [226]. Serine is also an important amino acid from a metabolic perspective as it participates in biosynthesis of purines and pyrimidines, serves a direct precursor for biosynthesis of several other amino acids including glycine, cysteine and tryptophane and an intermediate for biosynthesis of spingolipids and folate in many microorganisms [227,228,229]. Production of L-serine currently relies on extraction of L-serine from protein hydrolysates or from molasses, or by enzymatic conversion of glycine and methanol using *Sarcina albida*, *Hyphomicrobium methylovorum* and *Corynebacterium glutamicum* [230,231,232]. Despite the successful production of L-serine from glycerol and methanol, a low glycine yield of 50% has made these production systems less attractive [233]. On the other hand the direct conversion of glucose to L-serine is a demanding challenge because of its pivotal function as an intermediary to biomass and amino acid biosynthesis. A few recent
studies to address this challenge have tried to overproduce serine from glucose in *Corynebacterium glutamicum* [233,234,235]. For example, in one of the first efforts by Peters-Wendisch *et al* [234], overexpression of three genes *serA*, *serB* and *serC* combined with downregulation of *glyA* and *sdaA* knockout resulted in a strain accumulating up to 86 mM L-serine (0.64 mole L-serine/model glucose). In a more recent study blocking and reducing the conversion of L-serine to pyruvate and glycine, respectively, releasing the feedback inhibition by L-serine on 3-phosphoglycerate dehydrogenase (PGDH), along with the co-expression of 3-phosphoglycerate kinase (PGK) and feedback-resistant PGDH (PGDHr) resulted in a *C. glutamicum* strain producing up to (14.22±1.41) µmol serine/gCDW [235].

Despite these efforts, the overproduction of serine in well characterized microorganisms such as *E. coli* has not been addressed yet. In this work, we present results from a computational study using OptForce [224] as a metabolic engineering tool to suggest genetic engineering interventions to overproduce L-serine from glucose in *E. coli*.

### 4.2. Methods

This section was adapted from [236], where we used the exact same procedure for overproduction of fatty acids in *E. coli*.

The iAF1260 metabolic model of *Escherichia coli* [38] was used to perform the simulations with the OptForce procedure [224,237,238] for overproduction of L-serine. Metabolic flux data for 35 reactions from the glycolytic, TCA and Pentose Phosphate pathway [236] was used to define the phenotypic space of a base strain (see Figure 4.1). All simulations were performed under aerobic minimal medium with glucose as the sole carbon source. Glucose minimal conditions were simulated by restricting the glucose uptake rate to 10 mmol gDW⁻¹ h⁻¹ and the oxygen uptake rate at 20 mmol gDW⁻¹ h⁻¹. The lower bound for the remaining exchange fluxes corresponding to the metabolites present in the minimal medium was set to -1,000 and the non-growth associated ATP maintenance was fixed at 8.39 mmol gDW⁻¹ h⁻¹ [38]. In addition, the biomass flux was
Figure 4.1. Flux map used to define the phenotypic space of a base strain [236].

fixed at the maximum achievable flux subject to the experimental flux measurements (i.e., 52% of the maximum theoretical). The upper bound for all other reactions was set to 1,000 whereas the lower bound was set to zero and -1,000 for irreversible and reversible reactions, respectively. All regulatory restrictions were imported from the \textit{iAF1260} model. The phenotypic space of the wild type strain consistent with
stoichiometry/regulation, uptake rates and flux measurements was constructed by successively maximizing and minimizing each reaction flux in the network subject to the network stoichiometry and all of the constraints mentioned above.

Similarly, the flux ranges consistent with a desired over-producing target for L-serine was obtained by iteratively maximizing and minimizing each flux subject to the network stoichiometry, uptake and medium conditions, regulatory constraints and overproduction target. In this study, we imposed a minimum production yield of 100% of the theoretical maximum (2.00 moles/mole of glucose) for L-serine. The remaining parameter values were unchanged from the wild-type case.

OptForce was subsequently used to identify the minimal set of reactions/genes that must be up/down-regulated or knocked out so as to maximize the formation of L-serine. OptForce contrasts the maximal range of flux variability between the wild-type strain against the ones consistent for the overproducing phenotype designed to meet a pre-specified yield for L-serine, respectively. As outlined in earlier efforts [224], by superimposing the flux ranges one-at-a-time, we first identify the fluxes that must depart from the original ranges in the face of overproduction (MUST_U, MUST_L, MUST_X sets). One can extend this classification procedure by considering sums and differences of two fluxes (MUST^{UU}, MUST^{UL}, MUST^{LL} sets) and arrive at a collective set of flux changes that must happen in the network for overproduction. We subsequently extract the minimal subset(s) of these reactions needed to guarantee the imposed bioengineering objective (i.e., FORCE sets).

We make use of a max-min bilevel optimization problem to identify the minimal set of engineering interventions that forces the yield of the product to the target value. The OptForce procedure identifies metabolic interventions that guarantee the imposed yield even when the network fights against these interventions. Modeled as a “worst-case” optimization problem, we iteratively solve this problem by increasing the number of direct manipulations \(k\) in each step until the target yield is achieved. Binary variables are used to identify pertinent reactions from the MUST sets whose flux should be increased, decreased, or set to zero (i.e., removed) in order to maximize the minimal product formation yield. The binary variables corresponding to the reaction interventions that appear in all solutions, as well as those corresponding to trivial solutions (e.g., the
up-regulation of the transport reaction corresponding to target product) were also fixed at one and zero, respectively to reduce run time. A biomass flux of at least 10% of theoretical maximum was enforced in all OptForce simulations, along with other constraints mentioned earlier. OptForce first identifies the interventions that have the largest contribution towards meeting the overproduction target thus providing a way to prioritize the implementation of genetic interventions. The use of integer cuts allows for the identification of alternate optimal solutions that can serve as alternate genetic intervention choices.

4.3. Results

4.3.1. Targeted pathway

Metabolic pathways that lead to the synthesis of L-serine in *Escherichia coli* follow a three-step conversion branching out from glycolysis. First, 3-phosphoglycerate formed as an intermediary from glycolysis is converted into 3-phosphohydroxypyruvate by the enzyme phosphoglycerate dehydrogenase (EC# 1.1.1.95). Secondly, the enzyme phosphoserine transaminase (EC# 2.6.1.52) catalyzes the conversion of 3-phosphohydroxypyruvate into L-phosphoserine using L-glutamate as the amino acid donor. Finally, phosphoserine phosphatase (EC# 3.1.3.3) catalyzes the conversion in L-serine (see Figure 4.2). In addition to being a direct participant in the formation of biomass, serine acts as a precursor for the synthesis of various amino acids in the metabolism.

![Figure 4.2. Metabolic pathways showing the three step conversion of 3-phosphoglycerate to L-serine.](image-url)
4.3.2. Identified FORCE sets for overproduction of L-serine

Results for the FORCE set of reactions to overproduce L-serine in *E. coli* is summarized in Figure 4.3. As the first set of interventions, OptForce suggested up-regulating any one of the three fluxes that directly lead to the synthesis of L-serine by at least 12 times of the maximum achievable flux in the wild-type. These reactions include phosphoglycerate dehydrogenase (PGCD), phosphoserine transaminase (PSERT) and phosphoserine phosphatase (PSP_L).

Metabolite 3-phosphoglycerate a precursor for production of L-serine is normally produced through forward glycolysis from glucose. Interestingly, interventions identified by OptForce suggest the use of glycolytic reactions that converge to 3-phosphoglycerate from two opposing forward and backward directions. This bi-directional mode is accomplished by knocking out glucose-6-phosphate isomerase (PGI) that blocks an ATP-consuming reaction in upstream glycolysis converting glucose-6-phosphate into fructose-6-phosphate. Instead of using upstream pathways in glycolysis, OptForce suggests using 2-dehydro-3-deoxy-phosphogluconate aldolase (EDA) from the pentose phosphate pathway to produce 3-phosphoglycerate and pyruvate. In order to maximize the flux of this reaction, OptForce suggests knockout of any one of reaction 1. phosphogluconate dehydrogenase (GND) or 2. transketolase (TKT1) or 3. transaldolase (TALA) that drive carbon away from 3-phosphoglycerate towards production of fructose-6-phosphate (see Figure 4.3). These interventions lead to not only saving one mole of ATP from glycolysis but also forming one mole of pyruvate as a by-product. Phosphoenolpyruvate synthase (PPS) next converts pyruvate using the available ATP into phosphoenolpyruvate. Subsequently, these knockouts force the reversal of enolase (ENO) and phosphoglycerate mutase (PGM) so that phosphoenolpyruvate is converted back into 3-phosphoglycerate. This further increases the availability of 3-phosphoglycerate, which serves as a precursor for serine biosynthesis.

Glycine hydroxymethyl-transferase (GHMT2r) utilizes L-serine towards the formation of glycine, methionine and other amino acids that can be synthesized using alternate biochemical pathways. Not surprisingly, OptForce suggests knocking out this reaction to avoid the drain of L-serine towards production of these compounds.
Figure 4.3. Metabolic interventions suggested by OptForce.

OptForce suggests these metabolic interventions in a growth medium where there exist at most one mole of oxygen/gDM.hr and at least two moles of NH$_4$/gDW.hr per each mole of glucose uptake. The collective implementation of the aforementioned interventions under these conditions guarantees achieving a serine production yield of at least 1.65 moles/gDW.hr per each mole of glucose (i.e., ~83% of the maximum theoretical yield) according to the simulations.

4.4. Discussion and conclusion

In this study, we described computationally derived metabolic engineering strategies for overproduction of L-serine in *E. coli*. The suggested modifications include not only straightforward up-regulations of genes in serine biosynthesis pathway and knockouts or downregulation of pathways consuming serine towards production of other metabolites but also many non-intuitive modifications distant from the target product. Some of the suggested modifications in this study such as upregulation of *serA, serB*
and/or serC and downregulation of glyA have been already implemented experimentally by other researchers in C. glutamicum [233,234,235] and proven to be effective in serine overproduction. It should be noted that OptForce suggests knocking out the gene glyA to avoid consumption of serine towards glycine production, however this gene has been found to be essential under the aerobic minimal medium. Nonetheless, OptForce suggest deletion of its corresponding reaction, GHMT2r, because this reaction is not essential under the minimal medium based on the iAF1260 model of E. coli, i.e., this is a model/experiment inconsistency. To resolve this problem, one can instead downregulate glyA, as has been done in previous studies. Further simulations revealed that if the flux of GHMT2r is downregulated by at least fivefold compared to that in the wild-type strain, one can still achieve a yield of up to 80% of the theoretical maximum (i.e., 1.6 moles/gDW.hr per each mole of glucose). The rest of modification suggested by OptForce are non-intuitive and cause a drastic redirection of the metabolic flux in glycolysis and PP pathway to save one mole of ATP and enable bidirectional convergence of metabolic flux towards 3-phosphoglycerate thereby increasing the availability of this precursor for serine biosynthesis.
Chapter 5

Optimization-driven Identification of Genetic Perturbations Accelerating the Convergence of Model Parameters in Ensemble Modeling of Metabolic Networks

5.1. Introduction

Mathematical models have been used extensively to elucidate the functional and organizational principles of the metabolic systems and to predict their response to genetic and environmental perturbations. These models are useful in many biotechnological applications by identifying targeted modifications (e.g., modulation of enzyme expression levels) leading to the optimization of a desired behavior of the system (e.g., overproduction of a biochemical). Last decade has witnessed a large number of studies using the constrained-based modeling of cellular metabolism relying on only stoichiometric description of the underlying system [156,239,240,241,242]. A key advantage of this modeling formalism lies in the minimal amount of biological knowledge and data it requires to make quantitative predictions about the phenotypic behavior of the system. Despite the many successes of the constrained-based modeling
and flux balance analysis (FBA) [156] in the analysis of metabolic systems and biotechnological and biomedical applications [1,29,40,131,153,154,178,224,243,244,245,246,247,248,249,250,251,252], this simplicity comes at the expense of a number of limitations. For example, FBA cannot capture metabolite concentrations information and nonlinearities due to kinetic/regulatory effects. In addition, all biological processes are inherently dynamic and one of the primary challenges in the post-genome era is to understand the temporal behavior of living systems. Given that FBA-based approaches assume steady state for the metabolic network, they fail to capture the transient behavior of the metabolic fluxes and metabolite concentrations. Dynamic flux balance analysis (dFBA), proposed by Mahadevan el al [162] partially addresses the latter limitation by solving a separate FBA problem along a sequence of time intervals.

A thorough assessment of the dynamic features of a metabolic network requires employing more sophisticated and often nonlinear mathematical models. Kinetic models rely on the identification of model parameters that best explain metabolite concentrations and enzymatic reaction rates in the network. Mass balances are expressed as nonlinear ordinary differential equations representing the rate of change in concentration of each participating metabolite in the network as a function of kinetic equations. Kinetic models for the central metabolism of yeast (i.e., glycolysis) and E. coli (i.e., glycolysis and pentose phosphate pathway) were proposed by Rizzi et al [253] and Chassagnole et al [254], respectively and were subsequently improved/extended by others [255,256]. Other examples include kinetic models of TCA cycle in Dictyostelium discoideum [257] and glycolysis in Trypanosoma brucei [258]. In contrast to stoichiometric-based modeling, which is amenable to descriptions up to genome-scale level and beyond [131,154,160], expanding kinetic models of metabolism leads to significant challenges. These challenges are caused by the difficulty in devising portable kinetic expressions for all biochemical transformations, lack of experimental information on concentrations and fluxes at a scale needed to support unambiguous parameter value identification, and more importantly poor understanding of how genetic and/or environmental perturbations may affect parameter value or even model structure. In some cases it might be even impossible to determine the kinetic rate laws and parameters experimentally. In addition, these
experiments are often performed under *in vitro* conditions, which may not truly represent the physiological conditions inside the cell [254]. Moreover, the range of validity of kinetic parameters and even the functional form of the kinetic rate laws may vary under different physiological conditions or due to evolutionary processes [259].

Existing methods are aware of these difficulties and tried to come up with ways of resolving or bypassing them. Many studies have incorporated *in vivo* measurements, such as metabolite concentrations and reaction rates in response to targeted environmental perturbations, in the estimation of kinetic parameters [253,254,260,261]. However, they still rely on kinetic rate equations derived from *in vitro* experiments. Other studies adopted kinetic representations drawn from generalized approximate rate equations [262] to side-step the need for exact determination of the kinetic rate laws. These include power-laws (or S-systems or generalized mass action) [263,264,265,266], log-lin [267,268,269], and lin-log kinetics [270,271], saturating and cooperative [272], convenience rate laws [273] and more [274,275,276,277]. Alternatively, other researchers have explored a combination of lumped (e.g., Michaelis-Menten) and approximate rate equations [278]. Using these approximate kinetic equations may lead to some loss of accuracy, however, they are guaranteed to be valid over narrower ranges of conditions. Conforming to the same general kinetic form for all reactions in the network simplifies model construction and parameterization. In addition, kinetic parameters can be obtained using more established experimental techniques such as metabolic control analysis (MCA) [279,280,281,282,283].

More recently the integration of stoichiometric and kinetic modeling has been pursued [284,285,286,287,288,289]. For example, in the method proposed by Smallbone et al [284] the computed fluxes using FBA are allowed to change dynamically according to lin-log kinetics. In another effort a step-wise procedure called mass action stoichiometric simulation (MASS) was proposed to create larger scale kinetic models of metabolism by combining mass action rate law and stoichiometric modeling while using omic (metabolomic, fluxomic, and proteomic) data [285].

Faced with the difficulty in pinpointing unique parameter values for kinetic models applicable across a wide range of conditions, methods have been developed that aimed at identifying a range of kinetic parameter values [259,290,291,292]. For example, Famili et
al [290] proposed an approach which integrates metabolomic data and constrained-based modeling to construct a kinetic solutions space called the $k$-cone. In another effort Steuer et al [291] proposed a method for the quantitative analysis of dynamic capabilities of a metabolic network (e.g., the possibility of bifurcation or sustained oscillation) without requiring an explicit knowledge of functional form of the kinetic rate laws. This is achieved by constructing an ensemble of local linear models encompassing the space of all kinetic models such that any element of each model is either accessible experimentally (e.g., by incorporating flux or concentration data) or amenable to clear biochemical interpretation. A limitation associated with determination of the feasible space of kinetic parameters is that this space can be extremely large and thus difficult to faithfully sample. In addition, random sampling of this space can lead to the emergence of additional issues such as increased computational time or physiologically irrelevant model predictions. Therefore, approaches that efficiently make use of the available experimental data to reduce the feasible parameter space are highly desirable. Ensemble modeling (EM) of metabolic networks proposed by Tran et al [292] was developed towards achieving this goal through successively reducing the parameter space by using phenotypic data such as experimental flux measurements. Even though these phenotypic data are measured at steady state they are terminal points of complex dynamic processes. In addition, such these data are often available as part of conventional metabolic engineering efforts and thus the EM procedure does not require dedicated experiments for kinetic model construction, such as measuring dynamic metabolite concentrations [293]. Ensemble modeling of metabolic networks has been successfully applied to guide the overproduction of L-lysine [294] and aromatic products [295] in E. coli, and to decipher the effects of perturbations to the hepatic metabolic network on fatty acid oxidation and glucose uptake [296].

The EM procedure relies on constructing an initial ensemble of models all of which converge to the same (reference) steady state flux distribution (see Figures 5.1A and 5.1A’) followed by rounds of model screening by simulating genetic perturbations resulting in new steady state flux distributions compared to the reference strain. After a number of screening steps, a minimal set of physiologically relevant model parameterizations are extracted which are then used in a predictive fashion. A limitation
associated with the current form of the EM procedure is that the enzyme(s) that should be perturbed and the type of perturbation (i.e., knock out/down/up) are chosen based on biological insight and intuition. This may not be the most efficient approach as the nonlinearities in the model propagate throughout the metabolic network making *a priori* predictions about the strength of the effect of perturbations on steady state fluxes uncertain. The objective of this study is to systematically identify which genetic/enzymes to perturb and what type of perturbation to choose in order to maximally reduce the number of retained models in the ensemble after each round of model screening. To this end, we devised an optimization-based algorithm that successively identifies single, double, triple and higher order combinations of genetic perturbations resulting in the maximum divergence of the predicted fluxes over the ensemble parameterizations thus allowing for maximal elimination of models from the ensemble and faster convergence to the correct ones. We applied this algorithm to a model of the central metabolism in *E. coli* spanning reactions in glycolysis, pentose phosphate (PP) pathway and tricarboxylic acid (TCA) cycle. This study revealed that knockouts (as opposed to overexpressions) are enzyme perturbations having the highest “scattering” effect on the predicted fluxes and thus maximally screening models out of the ensemble. Surprisingly, we find that enzyme perturbations distant from the target fluxes are in some cases the ones leading to the elimination of most models from the ensemble.
Figure 5.1. Comparison of variance and departure from steady state as descriptors for the most informative enzyme perturbations. In (A)-(C) reactions are projected onto their first two principal components (PC1 and PC2), whereas (A’)-(C’) show a corresponding sample output from the EM procedure. Figures (A)-(C) provide only a qualitative representation of the model predictions. In (B) and (C) only the models whose predicted fluxes are located inside the dashed circle, and in (B’) and (C’) only those lying inside the dashed lines are accepted and the rest are rejected. (A) At reference steady state all models reach the same steady state. (B) Perturbations leading to the maximum variance between the model predictions maximally reduce the number of retained models in the ensemble. (C) Maximum departure from steady state is not always a good descriptor for the most informative enzyme perturbations as the predicted fluxes can significantly depart from the reference steady state but may be very close to each other and to the experimental flux measurements for the perturbed strain thereby causing most models to pass the screen.
5.2. Methods

5.2.1. Ensemble modeling of metabolic networks

The ensemble modeling of metabolic networks is based on decomposition of enzymatic reactions in the network to a series of elementary reaction steps collectively governing the overall behavior of that reaction [292]. For example, if reaction $X_i \rightarrow X_{i+1}$ is catalyzed by enzyme $E$, it can be broken down to its corresponding elementary reactions as follows:

$$X_i + E \xleftarrow{v_1} X_i E \xrightarrow{v_2} X_{i+1} E \xleftarrow{v_3} X_{i+1} + E$$

The rate law for each individual elementary reaction follows the mass action kinetics. For example:

$$v_i = k_i [X_i] [E] \quad (5.1)$$

Here $[X_i]$ and $[E]$ denote the concentration of metabolite $X_i$ and free enzyme $E$. The right-hand side of equation (5.1) can be expressed in dimensionless form through dividing the concentration of each metabolite by its concentration at the reference steady state (i.e., $[X_{i,ss,ref}]$) and that of free enzymes and enzyme complexes by the total enzyme concentration at the reference state (i.e., $[E_{total,ref}]$). For example, $v_i$ in equation (5.1) can be expressed with dimensionless variables as follows

$$v_i = \left( k_i X_{i,ss,ref}^{E_{total,ref}} \right) \frac{[X_i]}{X_{i,ss,ref}} \frac{[E]}{E_{total,ref}} = \tilde{K}_i \tilde{X}_i \tilde{e} \quad (5.2)$$

where, $\tilde{X}_i$ and $\tilde{e}$ are normalized metabolite concentration and enzyme fraction, respectively. The advantage of working with these two dimensionless quantities is that it obviates the need to know the metabolite and enzyme concentrations in the network. A similar equation can be written for each elementary reaction associated with to overall reaction $j$. Note that due to the conservation of total amount of enzyme in the system, sum of the enzyme fractions corresponding to free enzymes and enzyme complexes of each particular enzyme $r$ is set to one:

$$\sum_{i' \in \text{enzyme } r} \tilde{e}_{i'} = 1 \quad (5.3)$$
In addition, the rate of each overall reaction $j$ in the network is the difference between the forward and backward rates of each elementary step $m$ as following:

$$V_j = v_{j,2m-1} - v_{j,2m}$$  \hspace{1cm} (5.4)

Subscripts $2m-1$ and $2m$ denote forward and backward elementary reactions of step $m$, respectively. The EM procedure also incorporates thermodynamic information in the construction of the ensemble of models by linking the reaction reversibilities for each elementary step $m$ ($R_m$) and free Gibbs energy change of each overall reaction $j$ in the network ($\Delta G_j$) as follows [292]:

$$\frac{\Delta G_j}{RT} \leq \text{sign}(V_j) \sum_m \ln R_m \leq \frac{\Delta G_j}{RT}$$  \hspace{1cm} (5.5)

where, $R$ is the universal gas constant and $T$ is temperature. Reaction reversibility for each elementary step $m$ is defined as following:

$$R_m = \frac{\min(v_{j,2m-1}, v_{j,2m})}{\max(v_{j,2m-1}, v_{j,2m})}$$  \hspace{1cm} (5.6)

or alternatively,

$$R_m = \left(\frac{v_{j,2m}}{v_{j,2m-1}}\right)^{\text{sign}(V_j)}$$  \hspace{1cm} (5.7)

Reversibility ranges between zero and one, where zero signifies the complete reaction irreversibility (in the forward or backward direction) and one indicates that the reaction is at equilibrium (i.e., rates of forward and backward reactions are equal). By combining equations (5.4) and (5.7) one can express the rate of forward and backward elementary reactions as a function of reversibility:

$$v_{j,2m-1} = \frac{V_j}{1 - R_m^{\text{sign}(V_j)}}$$  \hspace{1cm} (5.8)

$$v_{j,2m} = \frac{V_j R_m^{\text{sign}(V_j)}}{1 - R_m^{\text{sign}(V_j)}}$$  \hspace{1cm} (5.9)

Given a set of experimental flux measurements at steady state for the wild-type system the EM procedure constructs an ensemble of $N$ models such that they all 'anchor' to the same reference steady state [292], i.e., they can replicate the wild-type (reference)
flux data correctly. We note that at the reference steady state $[X_i] = X_i^{ss,ref}$ and thus $\tilde{X} = 1$. Therefore, equation (5.2) is reduced to:

$$v_1^{ss,ref} = \tilde{K}_1 \tilde{e}^{ss,ref}$$  \hspace{1cm} (5.10)

Given that the number of variables ($v_1^{ss,ref}, \tilde{K}_1$ and $\tilde{e}^{ss,ref}$) is higher than the number of equations (equations 5.3 and 5.10) the system is underdetermined and thus $\tilde{K}_1 = f(v_1^{ss,ref}, \tilde{e}^{ss,ref})$. Since the rate of elementary reactions can be expressed in terms of reversibilities $\tilde{K}_1 = f(R_m, \tilde{e}^{ss,ref})$. The stepwise EM procedure for constructing the kinetic models can now be summarized as follows:

**Step 1- Assign enzyme fractions:** Randomly sample enzyme fractions at the reference steady state subject to conservation of enzymes as stated in equation (5.3).

**Step 2- Assign reversibilities:** Randomly sample reaction reversibilities for each elementary step subject to thermodynamic constraints in equation (5.5).

**Step 3- Compute the rate of elementary reactions:** Having assigned reversibilities and the reference steady state fluxes, compute the rate of forward and backward elementary reactions using equations (5.8) and (5.9).

**Step 4- Compute the kinetic parameters:** Having enzyme fractions and the rate of elementary reactions, compute the kinetic parameters using equation (5.10).

Once the kinetic parameters are determined and the ensemble is populated by different sets of models, one can simulate the dynamic behavior of the system by writing the mass balance for each metabolite, free enzyme and enzyme complex as follows:

$$\frac{dC_i}{dt} = \sum_j \sum_m S_{i,j,m} v_{j,m}$$  \hspace{1cm} (5.11)

where, $C_i$ denotes the normalized concentration of each metabolite $i$ or the enzyme fraction for a free enzyme or enzyme complex $i$. $S_{i,j,m}$ is the stoichiometric coefficient of $i$ in (forward or backward) elementary reaction of step $m$ corresponding to overall reaction $j$ in the network. Note that the rates of elementary reactions in this equation are
substituted by their corresponding kinetic expressions from equation (5.2). Solving this nonlinear system of ODEs along with equation (5.3) would enable one to simulate the dynamic behavior of the wild type (or reference) system as well as the steady state normalized metabolite concentrations and fluxes. It is worth noting that the EM procedure is amenable to incorporating lumped kinetic rate laws, such as Michaelis-Menten, Hill equation, and allosteric enzyme kinetics, for any reaction if available [297].

The central task of the EM procedure involves screening of the models in the ensemble by using enzyme perturbations (knock-outs/ups/downs). Upon each perturbation the models in the ensemble reach a new steady state different from the wild-type strain thereby providing a basis to screen out the models whose predictions do not agree with the flux measurements for the perturbed strain. After sufficient rounds of perturbations one can finally arrive at a minimal set of physiologically relevant kinetic models, which can then be used as a reliable basis for other types of analyses (e.g., computational strain design). Enzyme perturbations can be modeled by modifying equation (5.3) as follows:

$$\sum_{i' \in \text{enzyme } r} \hat{e}_{i'} = L_r$$  \hspace{1cm} (5.12)

where, $L_r$ can take a value of zero, less than one, one or greater than one for each enzyme $r$ representing knockout, downregulation, no perturbation and overexpression, respectively.

### 5.2.2. Identifying the most informative enzyme perturbations

EM procedure in its current form does not use any rationale to choose which enzymes to perturb and which type of perturbation (i.e., knock-out/down/up) to perform. This may significantly delay convergence to a final set of physiologically acceptable kinetic models. Therefore, we aimed to pro-actively identify genetic/enzyme perturbations that maximally reduce the number of retained models in the ensemble after each round of model screening (i.e., the most informative perturbations). As noted earlier, variance is a good descriptor of such these perturbations whereby the most informative enzyme perturbations are those causing the maximum degree of separation among the predicted fluxes by models in the ensemble thereby leading to the elimination of a larger
number of models at each screening step and accelerating the convergence. We
developed an optimization-based procedure for the targeted identification of higher order
combinations of such these perturbations whose mathematical description requires the
definition of the following sets:

\[ N = \{n\} = \text{Set of the models in the ensemble} \]
\[ J = \{j\} = \text{Set of the reactions in the model} \]
\[ M = \{m\} = \text{Set of the elementary steps for each reaction} \]
\[ I = \{i\} = \text{Set of all metabolites, free enzymes and enzyme complexes} \]
\[ R = \{r\} = \text{Set of enzymes} \]

The enzyme perturbations are enforced using three binary variables encoding which
enzyme(s) should be knocked out, overexpressed, or downregulated:

\[
y_X^r = \begin{cases} 
1 & \text{if enzyme } r \text{ is knocked out} \\
0 & \text{otherwise} 
\end{cases}
\]
\[
y_U^r = \begin{cases} 
1 & \text{if enzyme } r \text{ is upregulated} \\
0 & \text{otherwise} 
\end{cases}
\]
\[
y_D^r = \begin{cases} 
1 & \text{if enzyme } r \text{ is downregulated} \\
0 & \text{otherwise} 
\end{cases}
\]

The following mixed-integer nonlinear programming (MINLP) formulation is
constructed for the targeted identification of the most informative enzyme perturbations:

Maximize \[ z = \left( \frac{\sum_{j \in \text{basis rxns}} Var_j}{N_{\text{basis rxns}}} \right) \]

s.t.

\[ Var_j = \sum_{n} \left( V_j^n - \sum_{n} V_j^n / N_{\text{Ensemble}} \right) \left( N_{\text{Ensemble}} - 1 \right) \quad \forall j \]  
(5.13)
\[
\sum_{j} \sum_{m} S_{i,j,m} v_{j,m}^{n} = 0 \quad \forall i,n \tag{5.14}
\]

\[
v_{j,m}^{n} = \bar{K}_{j,m}^{n} \left( \sum_{i \in \text{metabolites}} \tilde{X}_{i}^{n} \right) \left( \sum_{i \in \text{free enzyme & enzyme complexes}} \tilde{e}_{i}^{n} \right) \quad \forall j,m,n \tag{5.15}
\]

\[
V_{j}^{n} = v_{j,2m-1}^{n} - v_{j,2m}^{n} \quad \forall j,m,n \tag{5.16}
\]

\[
\sum_{i' \in \text{enzyme } r} \tilde{e}_{i'}^{n} = 1 - y_{r}^{X} + (L_{r}^{D} - 1)y_{r}^{D} + (L_{r}^{U} - 1)y_{r}^{U} \quad \forall r,n \tag{5.17}
\]

\[
y_{r}^{X} + y_{r}^{U} + y_{r}^{D} \leq 1 \quad \forall r \tag{5.18}
\]

\[
\sum_{r} (y_{r}^{X} + y_{r}^{U} + y_{r}^{D}) \leq (\text{Allowable # of perturbations}) \tag{5.19}
\]

where, \( Var_{j}, N_{\text{ensemble}}, \) and \( N_{\text{basis rxns}}, \) represent the variance of reaction \( j, \) number of models in the ensemble and number of reactions used as the basis to compute the variance. Superscript \( n \) for parameters and variables represents that they are related to model \( n \) in the ensemble. The objective function of this problem maximizes the average variance over a set of pre-specified reactions. Constraint (5.13) computes the prediction variance for each reaction over all models in the ensemble. Constraint (5.14) represents the steady state mass balance for each metabolite, free enzyme and enzyme complex. Constraints (5.15) is the generalized form of equation (5.2), and constraint (5.16) is a recast of constraint (5.4) for each model in the ensemble. Constraint (5.17) forces an enzyme to be knocked out, overexpressed or downregulated by \( L_{r}^{U} \) and \( L_{r}^{D} \) folds, respectively, whereas constraint (5.18) restricts each enzyme to undergo a maximum of only one type of perturbation. Finally, constraint (5.19) imposes an upper limit on the total number of interventions. We note that this optimization problem is nonlinear and non-convex due to the presence of bilinear terms in the rate equations (constraint 5.15). Moreover, incorporation of lumped kinetic equations such as Michaelis-Menten in the EM framework introduces additional nonlinearities into the optimization problem.
Given the large number of binary variables, nonlinear terms and constraints in this formulation, regular nonlinear optimization solvers such as CONOPT and BARON [179] cannot be used to solve this optimization problem and using customized decomposition algorithms to reduce the solution time is inevitable. To this end, we relied on a customized outer-approximation algorithm [298] previously used for the design of synthetic genetic circuits using a kinetic description of basic parts characteristics and interactions [299]. The basic idea here is to create a sequence of converging lower bounds and upper bounds for the original problem to bracket the optimal solution. To comply with the standard form of the optimization problem presented in [298] and [299] we recast the problem of finding the most informative enzyme perturbations as a minimization problem through multiplying the objective function by -1 (i.e., Minimize $z = - [\text{Average prediction variance}]$). The step-wise algorithm to solve this optimization problem is as follows:

**Step 1- Initialize:** Initialize the iteration counter $\text{iter} = 0$. Set lower bound (LB) to $-\infty$ and upper bound (UB) to $\infty$. Generate an initial feasible integer solution $Y_{r,\text{iter}} = \begin{bmatrix} y_{r,\text{iter}}^X, y_{r,\text{iter}}^U, y_{r,\text{iter}}^D \end{bmatrix}$ \(\forall r\). This can be simply done by solving the following optimization problem:

Minimize $z = 0$

s.t.

\[
\begin{align*}
|y_r^X + y_r^U + y_r^D| & \leq 1 \quad \forall r \\
\sum_r y_r^X + y_r^U + y_r^D & \leq \text{(Allowed # of perturbations)} \\
y_r^X, y_r^U, y_r^D & \in \{0,1\} \quad \forall r
\end{align*}
\]

**Step 2- Update upper bound:** Given the integer solution, solve the original optimization problem (subject to constraints 5.13-5.17) to compute the value of objective function $z^*,\text{iter} = - \text{(Average variance)}$. Set the UB = min ($z^*,\text{iter}$, UB).
Step 3- Update lower bound: In this step a lower bound for the objective function is computed by solving a relaxed form of the optimization problem called the master problem using the outer approximation algorithm [298], where a number of supporting hyper planes are constructed at the point of interest. Following the customized outer-approximation procedure used in [299], the supporting hyper planes are constructed at the current integer solution and the master problem can thus be formulated as following:

Minimize $\mu$

s.t.

$$\mu \geq z^k + \sum_r (Y_i - Y_r^k) \left( \frac{\partial z}{\partial Y_r} \right)_{Y_i = y^i} \quad k = 1,2,\ldots,\text{iter}$$

$$y^X_r + y^U_r + y^D_r \leq 1 \quad \forall r$$

$$\sum_r y^X_r + y^U_r + y^D_r \leq \text{(Allowed # of perturbations)}$$

$$y^X_r, y^U_r, y^D_r \in \{0,1\} \quad \forall r$$

The partial derivatives are computed using the finite difference [299]. Solving the master problem not only results in a new integer solution ($Y_i^{*,\text{master}}$) but also provided a lower bound on the objective function of the optimization problem. Set $\text{LB} = \max(\text{LB}, \mu^{*,\text{master}})$, where $\mu^{*,\text{master}}$ is the optimal objective function of the master problem.

Step 4- Test of convergence: If $\text{LB} > \text{UB}$ (cross-over) or $|\text{UB} - \text{LB}|$ is less than a pre-specified threshold then optimal solution has been found and STOP. Otherwise, increase the iteration counter ($\text{iter} = \text{iter} + 1$) and set $Y_r^{\text{iter}} = Y_r^{*,\text{master}} \quad \forall r$ and go back to Step 2.

Given that the optimization problem is not convex due to the presence of bilinear terms and other nonlinear terms when using a lumped kinetic equation instead of elementary reactions, global optimality is not guaranteed. Therefore, the procedure stated above is solved from multiple initial guesses for the binary variables and the local optimum providing the best objective function is selected as the final solution. In addition, note that by using this solution procedure one can take advantage of the fact that
the constraints of this optimization problem are decomposable for each model in the ensemble and are coupled only through the objective function and binary variables. Hence, once the integer variables are fixed, the problem is reduced to solving $N_{\text{ensemble}}$ independent system of nonlinear equations to compute the steady state reaction rates (Step 2), where $N_{\text{ensemble}}$ is the total number of models in the ensemble. This enables one to take advantage of parallels computing facilities of software packages like MATLAB to significantly reduce the solution times. Furthermore, one can also instead solve the system of ordinary differential equations (ODEs) to obtain steady state reaction rates if it is deemed to be more efficient computationally than solving the system of nonlinear equations.

5.3. Results

5.3.1. Criterion selection for describing the most informative enzyme perturbations

Two different criteria were explored to describe the most informative perturbations (i.e., the ones leading to the maximal elimination of models from the ensemble after each round of perturbations). These include (i) average prediction variance over the models in the ensemble and (ii) average departure from the wild-type steady state upon perturbation, on the basis of a specific set of reactions in the model whose fluxes can be measured experimentally. In the former, the most informative enzyme perturbations are the ones that cause the models in the ensemble to generate the most divergent flux distributions, whereas in the latter they are the ones causing the new flux distribution upon perturbation to maximally depart from the wild-type flux distribution. We postulate that variance is a better criterion to serve this purpose as the maximum degree of separation among the model predictions will lead to the rejection of a higher number of models whose predicted fluxes depart from experimental data for the perturbed strain. This is shown pictorially in Figure 5.1B where fluxes are projected onto their first two principal components for representation purposes (see also Figure 5.1B’). Conversely, if departure from steady state is used as the basis, the predicted fluxes by the models in the ensemble may significantly deviate from the wild type (reference) flux distribution but
stay very close to each other and to experimental flux measurements for the perturbed strain thereby providing poor screening as the majority of the models are accepted (see Figures 5.2B and 5.2B’). For example, consider the case where the reaction used as the screening basis is located at the end of a linear pathway, whereby removal of a reaction (i.e., gene/enzyme knockout) can completely block the metabolic flux through this pathway including the terminal reaction. Even though this perturbation causes a significant deviation from the wild-type flux distribution, the majority of models in the ensemble would be able to correctly predict a zero flux through that pathway as well as towards the terminal reaction(s) used as the screening basis irrespective of their kinetic parameterization. Therefore, in the studies described in here maximum ensemble model variance prediction is used as the criterion for the selection of perturbations.

5.3.2. The metabolic model

To demonstrate the applicability of the proposed algorithm we constructed a metabolic model representing the central metabolism of E. coli as shown in Figure 5.2. This model consists of 34 reactions and 32 metabolites and encompasses glycolysis, TCA cycle and pentose phosphate pathway. This metabolic model allows for glucose uptake using the phosphotransferase system (pts) and contains two export reactions, one for succinate (in TCA cycle) and the other for ribose-5-phosphate (in PP pathway). In addition, several known inhibitory effects [254,292] are included in this model to account for their effect on the final flux distribution in the network. A similar structure was previously used to model succinate production in E. coli using the EM procedure [297]. The exhaustive list of reactions and metabolites in this model as well as regulatory interactions are given in Appendix B. A feasible flux distribution (i.e., satisfying steady state mass balance) using 100 moles/sec of glucose uptake as the basis was obtained and chosen to represent the flux distribution for the wild-type system (see Figure 5.2).
Figure 5.2. Metabolic model used for simulations and the corresponding flux distribution for the wild-type strain. This model spans glycolysis, pentose phosphate pathway and tricarboxylic acid (TCA) cycle.
5.3.3. Perturbation results

Consistent with the EM procedure, the kinetics governing the behavior of the metabolic model was obtained by decomposing all reactions to elementary reaction steps. The reaction representing glucose uptake by the phosphotransferase system was decomposed into 20 elementary steps as detailed in [292]. An ensemble of 100 models was created using the EM procedure for this analysis. All these models can accurately predict the wild-type flux distribution. All internal enzymes were considered for perturbation, where two types of perturbations are examined in this study including knockouts and twofold increase in the enzyme levels. The impact of the perturbations on ensemble model predictions was evaluated for three target reactions (i.e., glucose uptake, succinate, and ribose-5-phosphate export reactions), which served as proxies for experimental measurements. We considered a total of seven cases including each reaction by itself, each possible pair, and all three of these reactions as the basis to compute the average variance of the predicted fluxes over all models in the ensemble. This allows for exploring the role of which reactions are measured on the selection of optimal perturbation to perform in the network. In addition, we examined using the optimization procedure proposed in the Methods section up to three enzyme perturbations to identify the ones resulting in the maximum degree of separation among the predicted fluxes. The results of this analysis are summarized in Table 5.1 and Figure 5.3-5.9.

5.3.3.1. Single enzyme perturbations

We determined the optimal enzyme perturbations in this case using an exhaustive enumeration procedure and the optimization-based method to provide a check in the calculations. As shown in Table 5.1 and Figure 5.3, when pts is used as the basis to compute the variance, knocking out gnd (coding for 6-phosphogluconate dehydrogenase) in pentose phosphate pathway results in the highest divergence in the model predictions. Removal of gnd blocks the pentose phosphate pathway leading to redirection of the entire metabolic flux towards the glycolysis. Given that the maximum amount of the metabolic flux that can be carried out by glycolysis depends on the
Table 5.1. Summary of the optimal enzyme perturbations when different sets of reactions are used as the basis to compute the variance. The optimal perturbations are also shown on the metabolic map in Figures 5.2 to 5.8.

<table>
<thead>
<tr>
<th>Reactions used to compute the variance</th>
<th># of perturbations</th>
<th>Variance</th>
<th>Optimal perturbation</th>
<th>Variance</th>
<th>Optimal perturbation</th>
<th>Variance</th>
<th>Optimal perturbation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Optimal perturbation</strong></td>
<td><strong>Variance</strong></td>
<td><strong>Optimal perturbation</strong></td>
<td><strong>Variance</strong></td>
<td><strong>Optimal perturbation</strong></td>
<td><strong>Variance</strong></td>
<td></td>
</tr>
<tr>
<td>Pts</td>
<td>Knockout gnd</td>
<td>540.74</td>
<td>Overexpress gap</td>
<td>566.92</td>
<td>Knockout sdh &amp; pyk</td>
<td>683.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>knockout gnd</td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX_succ</td>
<td>Knockout sdh</td>
<td>466.56</td>
<td>Knockout pyk &amp; sdh</td>
<td>560.55</td>
<td>Knockout sdh &amp; pyk</td>
<td>630.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX_r5p</td>
<td>Knockout pfk</td>
<td>55.90</td>
<td>Knockout pfk &amp; pdh</td>
<td>181.58</td>
<td>Knockout pfk &amp; pdh &amp; rpe</td>
<td>312.01</td>
<td></td>
</tr>
<tr>
<td>Pts &amp; EX_succ</td>
<td>Knockout sdh</td>
<td>457.36</td>
<td>Knockout pyk &amp; sdh</td>
<td>557.92</td>
<td>Knockout sdh &amp; pyk</td>
<td>657.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pts &amp; EX_r5p</td>
<td>Knockout gnd</td>
<td>270.55</td>
<td>Knockout pyk &amp; sdh</td>
<td>294.87</td>
<td>Knockout sdh &amp; pyk</td>
<td>356.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX_succ &amp; EX_r5p</td>
<td>Knockout sdh</td>
<td>245.11</td>
<td>Knockout pyk &amp; sdh</td>
<td>297.50</td>
<td>Knockout sdh &amp; pyk</td>
<td>330.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pts &amp; EX_succ &amp; EX_r5p</td>
<td>Knockout sdh</td>
<td>312.80</td>
<td>Knockout pyk &amp; sdh</td>
<td>383.43</td>
<td>Knockout sdh &amp; pyk</td>
<td>448.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overexpress pfk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

values of kinetic parameters of its reactions, in the absence metabolic flow in the PP pathway the metabolic network has to accordingly adjust the glucose uptake rate, which in turn causes the highest variation in the predicted flux through reaction pts.

When succinate export reaction (EX_succ) is used to compute the variance, sdh (succinate dehydrogenase) knockout appears as the most informative perturbation (Table 5.1 and Figure 5.4). Knocking out sdh prevents succinate from being converted to fumarate thereby redirecting the entire metabolic flux towards the oxidative half of the TCA cycle and ultimately towards EX_succ. Given that different models in the ensemble
Figure 5.3. Optimal enzyme perturbations when pts is used as the basis to compute the variance.
Figure 5.4. Optimal enzyme perturbations when EX_succ is used as the basis to compute the variance.
Figure 5.5. Optimal enzyme perturbations when EX_r5p is used as the basis to compute the variance.
Figure 5.6. Optimal enzyme perturbations when pts and EX_succ are used as the basis to compute the variance.
Figure 5.7. Optimal enzyme perturbations when pts and EX_r5p are used as the basis to compute the variance.
Figure 5.8. Optimal enzyme perturbations when EX_succe and EX_r5p are used as the basis to compute the variance.
Figure 5.9. Optimal enzyme perturbations when pts, EX_succ and EX_r5p are used as the basis to compute the variance.
have different parameter values for succinate secretion, this causes a significant divergence in the ensemble on the prediction of the models for succinate export rate. With ribose-5-phosphate export reaction (EX_r5p) as the basis to compute the variance, knockout of pfk (phosphofructokinase) is identified as the perturbation resulting in the maximum degree of separation for predictions of the models in the ensemble (see Table 5.1 and Figure 5.5). Removal of pfk blocks the metabolic flow through the upper glycolysis and redirects it towards the PP pathway thus increasing the flux towards ribose-5-phosphate export reaction and leading to a large variance among the predictions of the models in the ensemble. This is because the efficiency of ribose-5-phosphate secretion for models in the ensemble is governed by different model parameterizations.

Interestingly, we observed that sdh knockout appears to be the dominant enzyme perturbation when considering any possible pair combination of pts, EX_succ and EX_r5p which includes EX_succ, as well all three of these reactions as the basis to compute the prediction variance (see Table 5.1 and Figures 5.5, 5.6, 5.8 and 5.9). This implies that sdh knockout not only significantly affects the metabolic flux in TCA cycle, but also in glycolysis and/or PP pathways, which in turn requires the system to accordingly adjust the fluxes through pts for glucose uptake and EX_r5p for ribose-5-phosphate export. When considering both pts and EX_r5p as the target reactions, gnd knockout appears again as the optimal perturbation (see Figure 5.7) implying that the divergence of predicted rates for pts has a higher impact on the average total variance compared to combinations thereof.

**5.3.3.2. Double enzyme perturbations**

Allowing for double enzyme perturbations increases the variance by 5 to 225% compared to single perturbations depending on the reaction(s) used as the basis to compute the variance (see Table 5.1). The identified best enzyme perturbations for pts as the basis include the gnd knockout, which also appeared in the single enzyme perturbations along with two fold overexpression of gap (glyceraldehyde-3-phosphate dehydrogenase) that further increase the flux in glycolysis (see Table 5.1 and Figure 5.4). This in turn necessitates additional glucose uptake by pts leading to divergent flux predictions by the models depending on the value of kinetic parameters. For ribose-5-
phosphate export reaction (EX_r5p) as the basis to compute the variance, the pfk knockout (that also appeared in single enzyme perturbations) is selected along with pdh (pyruvate dehydrogenase) to further restrict the metabolic flux in lower glycolysis (see Table 5.1 and Figure 5.5). We note that in the absence of pdh the metabolic flux needs to be directed to the TCA cycle through ppc (phosphoenolpyruvate carboxylase) and given that according to the value of kinetic parameters for the models in the ensemble this reaction cannot carry as much flux as that of pdh, the metabolic flux in the lower glycolysis has to decrease. Therefore, less flux is directed to the lower glycolysis through tkt2 (transketolase) and instead it is further diverted towards ribose-5-phosphate export reaction leading to divergent predictions by the models in the ensemble.

Interestingly, for the remaining of the cases a combination of sdh knockout (which also appeared in the single enzyme perturbations) and pyk (pyruvate kinase) knockout were identified as the most informative perturbations (see Table 5.1 and Figures 5.4 and 5.6-5.9). Removal of pyk will require more flux to be carried out by ppc towards the TCA cycle thus changing the flux distribution in lower glycolysis and affecting the glucose uptake through pts, and ribose-5-phosphate export rates.

5.3.3.3. Triple enzyme perturbations

By moving up to triple enzyme perturbations the variance increases between about 11 to 72% compared to double perturbations (see Table 5.1). When ribose-5-phosphate export is used as the target reaction to compute the variance rpe (ribulose-5-phosphate 3-epimeras) removal in PP pathway is appended to pfk and pdh knockouts (see Figure 5.5). Knockout of rpe will block reactions tkt1 and tkt2 as well as talAB (transaldolase) and further divert the metabolic flux towards ribose-5-phosphate export reaction leading to divergent model predictions. For the rest of the cases overexpression of pfk is combined with sdh and pyk knockouts (see Table 5.1 and Figures 5.4 and 5.6-5.8). This leads to an increase in the metabolic flux towards TCA cycle through ppc and affects the demand for glucose uptake as well as the flux distribution in PP pathway and thus ribose-5-phosphate export rate.
5.4. Discussion and conclusion

In this study we proposed an optimization-based algorithm for identifying the most informative enzyme perturbations (knockout-outs/downs/ups) for the ensemble modeling of metabolic networks. The most informative perturbations are defined as those that maximally reduce the number of retained models in the ensemble after each round of model screening. We postulated in this study that the average prediction variance of steady state fluxes over all models in the ensemble is an appropriate descriptor of the most informative perturbations. We considered various sets of reactions with measurable fluxes (i.e., uptake and export) and considered up to three simultaneous perturbations. Interestingly, we observed that knockouts serve as dominant type of enzyme perturbations leading to the maximum variance in the model predictions, whereas overexpressions appear mostly for triple (or higher order) perturbations. In addition, we observed that the identified perturbations are not always located in a close distance to the reactions used as the basis to compute the variances.

The customized decomposition algorithm used in this study enabled us to circumvent the computational intractability of the problem by decomposing it into independent optimization problems for each model, which can be solved in parallel to significantly reduce the computational time. It should be noted however that due to the nonlinear and non-convex nature of the underlying problem global optimality for double and triple perturbations is not guaranteed, implying that better solutions are possible if starting from a different set of initial guesses. Extension of the decomposition algorithm used in this study to identify global optimum would be needed to address this problem. Overall, the methods developed in this study can greatly improve the efficiency of the EM procedure by enabling faster convergence to a final set of physiologically relevant kinetic models while minimizing the required experimental effort.
References


Appendix A. Optimization formulation and solution procedure for OptCom

This appendix was previously published in modified form in PLoS Computational Biology as part of the supplementary material for [131].

A.1. Mathematical description of OptCom

Mathematical description of OptCom for the flux balance analysis of microbial communities requires definition of the following sets:

\[ K = \{k \mid k = 1, 2, \ldots, K\} = \text{Set of microorganisms (or guilds) present in the community} \]

\[ J^k = \{j \mid j = 1, 2, \ldots, N^k\} = \text{Set of reactions in microorganism (or guild) } k \]

\[ I^k = \{i \mid i = 1, 2, \ldots, M^k\} = \text{Set of metabolites in microorganism (or guild) } k \]

\[ I^k_{\text{uptake}} \text{ and } I^k_{\text{export}} = \text{Set of metabolites that are exported or taken up by microorganism (or guild) } k \text{ and are shared (exchanged) with other community members} \]

\[ I^k_{\text{shared}} = \text{Set of the shared metabolites that are present in the extra-cellular environment and are shared (exchanged) among the community members. Note that for each metabolite in this set, there exists at least one corresponding metabolite in } I^k_{\text{uptake}} \text{ or } I^k_{\text{export}} \text{ for some } k \in K. \]

We also, define the following variables and parameters using these sets:

**Variables:**

\[ \nu^k_j = \text{Flux of reaction } j \text{ in microorganism (or guild) } k. \]

\[ e^k_i = \text{Export flux of metabolite } i \in I^k_{\text{export}} \text{ by microorganism (or guild) } k \left( e^k_i \in \{\nu^k_j\} \right). \]

\[ u^k_i = \text{Uptake flux of metabolite } i \in I^k_{\text{uptake}} \text{ by microorganism (or guild) } k \left( u^k_i \in \{\nu^k_j\} \right). \]

\[ e^c = \text{Export flux of the shared metabolite } i \in I^k_{\text{shared}} \text{ by the community.} \]

\[ u^c_i = \text{Uptake flux of the shared metabolite } i \in I^k_{\text{shared}} \text{ by the community.} \]
**Parameters:**

- $S_{ij}^k$: Stoichiometric coefficient of metabolite $i \in I^k$ in reaction $j \in J^k$.
- $LB_j^k$: Lower bound on reaction $j \in J^k$ in microorganism (or guild) $k$.
- $UB_j^k$: Upper bound on reaction $j \in J^k$ in microorganism (or guild) $k$.

OptCom is a multi-level and multi-objective optimization problem, which can be represented as following:

**Maximize / Minimize** $z = \text{Community-level objective}$  
subject to

\[
\begin{align*}
\text{Maximize}_j^k & v_j^k \\
\text{subject to} & \\
\sum_{j \in J^k} S_{ij}^k v_j^k &= 0 \quad \forall \ i \in I^k \quad (A.1) \\
LB_j^k & \leq v_j^k \leq UB_j^k \quad \forall \ j \in J^k \quad (A.2) \\
u_i^k &= uval_i^k \quad \forall \ i \in I_{\text{uptake}} \quad (A.3) \\
e_i^k &= eval_i^k \quad \forall \ i \in I_{\text{export}} \quad (A.4) \\
\sum_k uval_i^k + e_i^k &= \sum_k eval_i^k + u_i^k \quad \forall \ i \in I_{\text{shared}} \quad (A.5) \\
uval_i^k, eval_i^k, e_i^k, u_i^k & \geq 0 \quad \forall \ i \in I_{\text{shared}}, \ k \in K
\end{align*}
\]

Constraint (A.1) in the inner problems represents the steady-state mass balance for each microorganism (or guild) $k$. Constraint (A.2) imposes a lower and upper bound for each flux. Constraints (A.3) and (A.4) fix the uptake or export flux of a shared metabolite $i$ at the values $uval_i^k$ and $eval_i^k$, respectively, which are imposed by the outer problem. This means that $uval_i^k$ and $eval_i^k$ serve as variables for the outer problem, but act as parameters for the inner problems. Constraint (A.5) in the outer problem establishes a mass balance for each shared metabolite $i \in I_{\text{shared}}$ present in the extra-cellular environment (shared metabolite pool), where the terms $\sum_k uval_i^k$ and $\sum_k eval_i^k$ represent the total uptake and
Figure A.1- Graphical representation of the variables and equations defined in the OptCom formulation. This representation is given for a sample community composed of three microorganisms (guilds). In this figure $i$ is a typical metabolite that is exchanged (shared) among community members. The community-level objective function is assumed to be maximization of the total community biomass.

The multi-level structure of the OptCom implies that the inner problems are integrated in the outer stage so as a community-level objective function is primarily optimized (e.g., maximization of the total community biomass), while the fluxes are further constrained (by the inner problems) to maximize an individual-level criterion (i.e., export of the shared metabolite $i$ by community members, respectively. This constraint is the key equation modeling the interactions and communications among participants of the community (through exchanging or sharing metabolites). Any other microorganism or community related conditions can be readily incorporated into the OptCom formulation through addition of appropriate constraints to the inner or outer problems, respectively. A pictorial representation of this formulation is given in Figure A.1.
growth). Solution methods of the bilevel problems from [1,50] are used to convert the multi-level program of OptCom to a bilinear optimization problem by adding the constraints of the dual of each inner problem and setting the objective functions for the primal and dual problems equal to one another:

\[
\text{Maximize} \quad z = \text{Community-level objective (e.g., total community biomass)} \quad \text{[OptCom]}
\]

subject to

\[
\sum_{j \in J^k} S_{ij}^k v_j^k = 0 \quad \forall \quad i \in I^k, \quad k \in K \quad (A.1)
\]

\[
LB_j^k \leq v_j^k \leq UB_j^k \quad \forall \quad j \in J^k, \quad k \in K \quad (A.2)
\]

\[
u_i^k = \text{eval}_i^k \quad \forall \quad i \in I_{\text{uptake}}^k, \quad k \in K \quad (A.3)
\]

\[
e_i^k = \text{eval}_i^k \quad \forall \quad i \in I_{\text{export}}^k, \quad k \in K \quad (A.4)
\]

\[
\sum_{i \in I^k} S_{ij}^k \lambda_i^k + \mu_j^k - \eta_j^k = 0 \quad \forall \quad j \in J^k - \{\{v_{\text{biomass}}^k\} \cup \{u_i^k \mid i \in I_{\text{uptake}}^k\} \cup \{e_i^k \mid i \in I_{\text{export}}^k\}\}, \quad k \in K \quad (A.6)
\]

\[
\sum_{i \in I^k} S_{i,\text{biomass}}^k \lambda_i^k + \mu_{\text{biomass}}^k - \eta_{\text{biomass}}^k = 1 \quad (A.8)
\]

\[
v_{\text{biomass}}^k = \left( \sum_{j \in J^k} UB_j^k \mu_j^k - \sum_{j \in J^k} LB_j^k \eta_j^k + \sum_{i \in I_{\text{uptake}}^k} \text{eval}_i^k \alpha_j^k + \sum_{i \in I_{\text{export}}^k} \text{eval}_i^k \beta_j^k \right) \quad \forall \quad k \in K \quad (A.9)
\]

\[
\sum_k \text{eval}_i^k + e_i^k = \sum_k \text{eval}_i^k + u_i^c \quad \forall \quad i \in I_{\text{shared}} \quad (A.5)
\]

\[
\text{eval}_i^k, \text{eval}_i^k, e_i^k, u_i^c \geq 0 \quad \forall \quad i \in I_{\text{shared}}, \quad k \in K
\]

\[
\mu_j^k, \eta_j^k \geq 0 \quad \forall \quad j \in J^k, \quad k \in K
\]

\[
\lambda_i^k, \alpha_j^k, \beta_j^k \in \mathbb{R} \quad \forall \quad i \in I^k, \quad j \in \{u_i^k \mid i \in I_{\text{uptake}}^k\} \cup \{e_i^k \mid i \in I_{\text{export}}^k\}, \quad k \in K
\]
Here, $\lambda^i_j$, $\alpha^i_j$, $\beta^i_j$ (not restricted in sign), $\mu^i_j$ and $\eta^i_j$ (non-negative) are the dual variables associated with the steady-state mass balance (Constraint A.1), uptake and export constraints (constraints 3 and 4) and right- and left-hand side inequalities in constraint (A.2), respectively. Constraints (A.7) and (A.8) represent the dual constraints corresponding to uptake and export reactions for shared metabolites and the biomass reaction, respectively. Equation (A.6) is the dual constraint corresponding to all other reactions. Constraint (A.9) states that the objective functions of the primal and dual problems should be equal to achieve optimality (for inner problems). The bilinear formulation for OptCom is in general non-convex. It can be solved to optimality using the global optimization solver BARON [300]. Please note that problems with a few thousand bilinear terms in the constraints are typically solvable by BARON to global optimality. Given that the community members usually share only a limited number of metabolites the number of bilinear terms remain far below this limit for most cases. All examples presented in the manuscript were solved using this solver to achieve the global optimum. If an alternate nonlinear objective function for the inner problem such as minimization of the metabolic adjustment (MOMA) [159] deemed to be a better surrogate for cellular fitness, then instead of writing the dual constraints, one needs to write the Karush-Kuhn-Tucker (KKT) conditions for the inner problems (provided that they are convex).

**A.2. Determining the optimality levels using Descriptive OptCom**

In OptCom each species is assumed to maximize its own growth (i.e., in the inner problems). However, it happens quite often in microbial communities that individual microorganisms sacrifice in order to benefit other community member and improve an altruistic performance criterion. Therefore, individual species may grow at sub-optimal levels (e.g., 90% of the maximal growth rate) to benefit the entire population. This behavior can be identified and captured by OptCom through a modification of the optimization structure presented above. The idea is to quantify the level of optimality of growth for each species in the community consistent with the set of available experimental data. To this end, we introduce a new metric called optimality level for each species $k$ in the community ($c^k$) that captures the deviation of individual species growth
from their optimal behavior. Optimality levels can be determined upon incorporating all available experimental data related to the whole community (e.g., the biomass composition of the community) as constraints in the outer problem and any data related to individual microorganisms as constraints in the respective inner problems. This new mode of OptCom is called Descriptive OptCom and can be represented as following:

\[
\begin{align*}
\text{Maximize/Minimize} & \quad z = (\text{Community-level objective}) \\
\text{subject to} & \quad \begin{aligned}
 & \text{Maximize } \sum_{j \in J^k} v^k_{j, \text{biomass}} \\
 & \text{subject to} \\
 & \sum_{j \in J^i} s^k_{i,j} v^k_j = 0 \quad \forall \ i \in I^k \\
 & L^k_j \leq v^k_j \leq U^k_j \quad \forall \ j \in J^k \\
 & u^k_i = u_{\text{val}}^k \quad \forall \ i \in I_{\text{uptake}}^k \\
 & e^k_i = e_{\text{val}}^k \quad \forall \ i \in I_{\text{export}}^k \\
 & v^k_{\text{biomass}} \leq c^k \cdot \text{vopt}^k_{\text{biomass}} \\
\end{aligned}
\end{align*}
\]

Constraints describing experimental data for microorganism \( k \) (A.11)

\[
\begin{align*}
\sum_k u_{\text{val}}^k_i + e^k_i &= \sum_k e_{\text{val}}^k_i + u^k_i \quad \forall \ i \in I_{\text{shared}} \\
\text{Constraints describing experimental data for the whole community} & \quad \text{(A.12)} \\
u_{\text{val}}^k_i, e_{\text{val}}^k_i, e^k_i, u^k_i, e^k \geq 0 \quad \forall \ i \in I_{\text{shared}}, \ k \in K
\end{align*}
\]

Constraint (A.10) allows the biomass flux for each microorganism (or guild) to be lower \((c^k < 1)\), equal \((c^k = 1)\), or higher \((c^k > 1)\) than its optimum \((\text{vopt}^k_{\text{biomass}}))\). Note that \(\text{vopt}^k_{\text{biomass}}\) for each species is community-specific and is computed in the context of all microorganisms strive to maximize their own growth (using the [OptCom] formulation described in the previous section). The optimality level for each microorganism is a variable and is determined by solving the Descriptive OptCom. An optimality level of less than one for a microorganism \( k \) implies that it grows sub-optimally at a rate equal to \(100c^k\%\) of the maximum \((\text{vopt}^k_{\text{biomass}}))\) to optimize a community-level fitness criterion.
while matching experimental observations. Alternatively, an optimality level of one implies that microorganism $k$ grows exactly optimally at a rate equal to $\nu_{\text{opt}}^{k_{\text{biomass}}}$ whereas a value greater than one indicates that it achieves a higher biomass production level than the community-specific maximum (i.e., super-optimality) by depleting resources from one or more other community members. It is worth noting that super-optimality for a species is achievable only at the cost of at least one other community member growing sub-optimally. Note that the presence of inner problems in Descriptive OptCom allows each species to still optimize its fitness (i.e., maximize its own biomass) as much as possible considering the available experimental data. However, once the optimality levels are determined (i.e., when the RHS of equation 10 is a constant) the inner problems can be replaced with their respective constraints and the problem is simplified to a single-level optimization problem, which can be used to provide further predictions about the community.
Appendix B. Details of the metabolic model used in the case study of Chapter 5

Table B.1. List of reactions and inhibition information

<table>
<thead>
<tr>
<th>Reaction abbreviation</th>
<th>Reaction equation</th>
<th>inhibitor*</th>
</tr>
</thead>
<tbody>
<tr>
<td>pgI</td>
<td>G6P $\leftrightarrow$ F6P</td>
<td>6PGC</td>
</tr>
<tr>
<td>pfk</td>
<td>F6P + ATP $\leftrightarrow$ FBP + ADP</td>
<td>PEP</td>
</tr>
<tr>
<td>fba</td>
<td>FBP $\leftrightarrow$ DHAP + G3P</td>
<td>-</td>
</tr>
<tr>
<td>tpi</td>
<td>DHAP $\leftrightarrow$ G3P</td>
<td>-</td>
</tr>
<tr>
<td>gap</td>
<td>G3P + NAD $\leftrightarrow$ BPG + NADH</td>
<td>-</td>
</tr>
<tr>
<td>pgk</td>
<td>BPG + ADP $\leftrightarrow$ 3PG + ATP</td>
<td>-</td>
</tr>
<tr>
<td>gpm</td>
<td>3PG $\leftrightarrow$ 2PG</td>
<td>-</td>
</tr>
<tr>
<td>eno</td>
<td>2PG $\leftrightarrow$ PEP</td>
<td>-</td>
</tr>
<tr>
<td>pyk</td>
<td>PEP + ADP $\leftrightarrow$ Pyruvate + ATP</td>
<td>ATP</td>
</tr>
<tr>
<td>pps</td>
<td>Pyruvate + ATP $\leftrightarrow$ PEP + ADP</td>
<td>-</td>
</tr>
<tr>
<td>pdh</td>
<td>Pyruvate + NAD $\leftrightarrow$ AcCoA + NADH</td>
<td>-</td>
</tr>
<tr>
<td>ppc</td>
<td>PEP $\leftrightarrow$ Oxaloacetate</td>
<td>-</td>
</tr>
<tr>
<td>gltA</td>
<td>AcCoA + Oxaloacetate $\leftrightarrow$ Citrate</td>
<td>-</td>
</tr>
<tr>
<td>acn</td>
<td>Citrate $\leftrightarrow$ Isocitrate</td>
<td>-</td>
</tr>
<tr>
<td>icd</td>
<td>Isocitrate + NAD $\leftrightarrow$ 2-Ketoglutarate + NADPH</td>
<td>-</td>
</tr>
<tr>
<td>sucAB</td>
<td>2-Ketoglutarate + NAD $\leftrightarrow$ SuccinylCoA + NADH</td>
<td>-</td>
</tr>
<tr>
<td>sucCD</td>
<td>SuccinylCoA + ADP $\leftrightarrow$ Succinate + ATP</td>
<td>-</td>
</tr>
<tr>
<td>sdh</td>
<td>Succinate + NAD $\leftrightarrow$ Fumarate + NADH</td>
<td>-</td>
</tr>
<tr>
<td>fum</td>
<td>Fumarate $\leftrightarrow$ Malate</td>
<td>-</td>
</tr>
<tr>
<td>mdh</td>
<td>Malate + NAD $\leftrightarrow$ Oxaloacetate + NADH</td>
<td>-</td>
</tr>
<tr>
<td>zwf</td>
<td>G6P + NADP $\leftrightarrow$ 6PGL + NADPH</td>
<td>NADPH</td>
</tr>
<tr>
<td>pgl</td>
<td>6PGL $\leftrightarrow$ 6PGC</td>
<td>-</td>
</tr>
<tr>
<td>gnd</td>
<td>6PGC + NADP $\leftrightarrow$ Ru5P + NADPH</td>
<td>NADPH</td>
</tr>
<tr>
<td>rpe</td>
<td>Ru5P $\leftrightarrow$ X5P</td>
<td>-</td>
</tr>
<tr>
<td>rpiAB</td>
<td>Ru5P $\leftrightarrow$ R5P</td>
<td>-</td>
</tr>
<tr>
<td>tkt1</td>
<td>X5P + R5P $\leftrightarrow$ G3P + S7P</td>
<td>-</td>
</tr>
<tr>
<td>talAB</td>
<td>G3P + S7P $\leftrightarrow$ F6P + E4P</td>
<td>-</td>
</tr>
<tr>
<td>tkt2</td>
<td>X5P + E4P $\leftrightarrow$ F6P + G3P</td>
<td>-</td>
</tr>
<tr>
<td>ATP_rec</td>
<td>ATP $\leftrightarrow$ ADP</td>
<td>-</td>
</tr>
<tr>
<td>NAD_rec</td>
<td>NADH $\leftrightarrow$ NAD</td>
<td>-</td>
</tr>
<tr>
<td>NADP_rec</td>
<td>NADPH $\leftrightarrow$ NADP</td>
<td>-</td>
</tr>
<tr>
<td>pts</td>
<td>PEP $\leftrightarrow$ G6P + Pyruvate</td>
<td>-</td>
</tr>
<tr>
<td>EX_succ</td>
<td>Succinate --$\to$</td>
<td>-</td>
</tr>
<tr>
<td>EX_r5p</td>
<td>R5P --$\to$</td>
<td>-</td>
</tr>
</tbody>
</table>

* All considered as mixed inhibition
<table>
<thead>
<tr>
<th>Reaction abbreviation</th>
<th>Expanded reaction name</th>
</tr>
</thead>
<tbody>
<tr>
<td>pgi</td>
<td>Phosphoglucose isomerase</td>
</tr>
<tr>
<td>pfk</td>
<td>6-phosphofructokinase</td>
</tr>
<tr>
<td>fba</td>
<td>Fructose bisphosphate aldolase</td>
</tr>
<tr>
<td>tpi</td>
<td>Fructose phosphate isomerase</td>
</tr>
<tr>
<td>gap</td>
<td>Glyceraldehyde 3-phosphate dehydrogenase</td>
</tr>
<tr>
<td>pgk</td>
<td>3-phosphoglycerate kinase</td>
</tr>
<tr>
<td>pgm</td>
<td>Phosphoglycerate mutase</td>
</tr>
<tr>
<td>eno</td>
<td>Enolase</td>
</tr>
<tr>
<td>pyk</td>
<td>Pyruvate kinase</td>
</tr>
<tr>
<td>pps</td>
<td>Phosphoenolpyruvate synthase</td>
</tr>
<tr>
<td>pdh</td>
<td>Pyruvate dehydrogenase</td>
</tr>
<tr>
<td>ppc</td>
<td>Phosphoenolpyruvate carboxylase</td>
</tr>
<tr>
<td>gltA</td>
<td>Citrate synthase</td>
</tr>
<tr>
<td>acn</td>
<td>Aconitase</td>
</tr>
<tr>
<td>icd</td>
<td>Isocitrate dehydrogenase</td>
</tr>
<tr>
<td>sucAB</td>
<td>2-oxoglutarate dehydrogenase complex</td>
</tr>
<tr>
<td>succCD</td>
<td>Succinyl-CoA synthetase</td>
</tr>
<tr>
<td>sdh</td>
<td>Succinate dehydrogenase</td>
</tr>
<tr>
<td>fum</td>
<td>Fumarase</td>
</tr>
<tr>
<td>mdh</td>
<td>Malate dehydrogenase</td>
</tr>
<tr>
<td>zwf</td>
<td>Glucose-6-phosphate-1-dehydrogenase</td>
</tr>
<tr>
<td>pgl</td>
<td>6-phosphogluconolactonase</td>
</tr>
<tr>
<td>gnd</td>
<td>6-phosphogluconate dehydrogenase</td>
</tr>
<tr>
<td>rpe</td>
<td>Ribulose phosphate 3-epimerase</td>
</tr>
<tr>
<td>rpiAB</td>
<td>Ribose-5-phosphate isomerase</td>
</tr>
<tr>
<td>tkt1</td>
<td>Transketolase 1</td>
</tr>
<tr>
<td>talAB</td>
<td>Transaldolase</td>
</tr>
<tr>
<td>tkt2</td>
<td>Transketolase 2</td>
</tr>
<tr>
<td>pts</td>
<td>Glucose uptake (through pts)</td>
</tr>
<tr>
<td>EX_suc</td>
<td>Succinate export</td>
</tr>
<tr>
<td>EX_r5p</td>
<td>Ribose-5-phosphate export</td>
</tr>
<tr>
<td>Metabolite abbreviation</td>
<td>Expanded metabolite name</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>G6P</td>
<td>Glucose-6-phosphate</td>
</tr>
<tr>
<td>F6P</td>
<td>Fructose-6-phosphate</td>
</tr>
<tr>
<td>FBP</td>
<td>Fructose-1,6-bisphosphate</td>
</tr>
<tr>
<td>DHAP</td>
<td>Dihydroxyacetone phosphate</td>
</tr>
<tr>
<td>G3P</td>
<td>Glyceraldehyde-3-phosphate</td>
</tr>
<tr>
<td>BPG</td>
<td>1,3-diphosphoglycerate</td>
</tr>
<tr>
<td>3PG</td>
<td>3-phosphoglycerate</td>
</tr>
<tr>
<td>2PG</td>
<td>2-phosphoglycerate</td>
</tr>
<tr>
<td>PEP</td>
<td>Phosphoenolpyruvate</td>
</tr>
<tr>
<td>Pyruvate</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>AcCoA</td>
<td>Acetyl-CoA</td>
</tr>
<tr>
<td>Citrate</td>
<td>Citrate</td>
</tr>
<tr>
<td>Isocitrate</td>
<td>Isocitrate</td>
</tr>
<tr>
<td>2-Ketoglutarate</td>
<td>2-Ketoglutarate</td>
</tr>
<tr>
<td>SuccinylCoA</td>
<td>Succinyl-CoA</td>
</tr>
<tr>
<td>Succinate</td>
<td>Succinate</td>
</tr>
<tr>
<td>Fumarate</td>
<td>Fumarate</td>
</tr>
<tr>
<td>Malate</td>
<td>Malate</td>
</tr>
<tr>
<td>Oxaloacetate</td>
<td>Oxaloacetate</td>
</tr>
<tr>
<td>6PGL</td>
<td>6-phosphogluconolactone</td>
</tr>
<tr>
<td>6PGC</td>
<td>6-phospho gluconate</td>
</tr>
<tr>
<td>Ru5P</td>
<td>Ribulose-5-phosphate</td>
</tr>
<tr>
<td>X5P</td>
<td>Xylose-5-phosphate</td>
</tr>
<tr>
<td>R5P</td>
<td>Ribose-5-phosphate</td>
</tr>
<tr>
<td>S7P</td>
<td>Sedoheptulose-7-phosphate</td>
</tr>
<tr>
<td>E4P</td>
<td>Erythrose-4-phosphate</td>
</tr>
<tr>
<td>ATP</td>
<td>ATP</td>
</tr>
<tr>
<td>ADP</td>
<td>ADP</td>
</tr>
<tr>
<td>NAD</td>
<td>NAD</td>
</tr>
<tr>
<td>NADH</td>
<td>NADH</td>
</tr>
<tr>
<td>NADP</td>
<td>NADP</td>
</tr>
<tr>
<td>NADPH</td>
<td>NADPH</td>
</tr>
</tbody>
</table>
VITA

Ali Reza Zomorrodi

Education

<table>
<thead>
<tr>
<th>Institute</th>
<th>Field of study</th>
<th>Degree</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn State University</td>
<td>Chemical Engineering</td>
<td>PhD</td>
<td>2006-2012</td>
</tr>
<tr>
<td>Penn State University</td>
<td>Computational Science</td>
<td>PhD minor</td>
<td>2010-2012</td>
</tr>
<tr>
<td>Penn State University</td>
<td>Industrial Engineering</td>
<td>MS</td>
<td>2010-2011</td>
</tr>
<tr>
<td>AmirKabir University of Technology (Tehran Polytechnic), IRAN</td>
<td>Chemical Engineering-Biotechnology</td>
<td>MS</td>
<td>2002-2005</td>
</tr>
<tr>
<td>Sharif University of Technology, IRAN</td>
<td>Chemical Engineering</td>
<td>BS</td>
<td>1997-2002</td>
</tr>
</tbody>
</table>

Honors and Awards
- 2012 Genomic Sciences Meeting Student Travel Grant, Department of Energy (DOE), Bethesda, MD
- Larry Duda Award for Outstanding Graduate Student Performance in Chemical Engineering: Best Paper Award, Penn State, September 2011.
- 3rd place award for poster presentation in 2011 Graduate Exhibition, Penn State, March 2011
- 2010-2011 Network Science Exploration Research Grant, Penn State, April 2010
- 2009 Genomics: GTL Conference Student Travel Grant, Department of Energy (DOE), Bethesda, MD
- Graduate Fellowship, Department of Chemical Engineering, Penn State, Fall 2006

Publications


*Joint first authors  #Joint second authors


*Joint first authors

(Note: Publication name was changed from A. Zomorrodi to A.R. Zomorrodi since 2010)