# Abstract

# Acknowledgements

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Background and Motivation

## 1.2 Research Objective

A valid theoretical framework for discrimination of the two types of constraints in statistical properties of the production data

First, formulate the binning methods here because this will describe the hypothesis underlining my thesis.

Explanation of my hypothesis is a theoretical/conceptual framework as a starting point for the investigation. It is a well-defined object and based on facts.

## 1.3 Research Plan and Thesis Organization

Methods are introduced here as indicative of two fundamentally different constraints acting on the manufacturing process: technological constraints on the one hand and constraints related to material flow and production capacity on the other.

# 2 Methodology

The initial step of this master thesis work was to quantify the characteristics of two hypothetical types of constraints in industrial production: technology-driven constraints and load-driven constraints.

I am planning to achieve this with an Operations Research Model consists of two steps. First, analyzing the statistical properties of association networks over Time in an extensive data set from steel manufacturing; second, developing an abstract theoretical framework to understand better the connection between each type of constraint and the statistical patterns created by them.

## 2.1 Operations Research Model

Introduce proposed concepts in the Operations Research Model (OR model).

The OR model is a combination of Steel Manufacturing Events Analysis and Flux Balance Analysis. The art form of this model is to structure a standard data format and a shared analysis logic that allows comparing the results from manufacturing data and simulation data.

A brief introduction for Association Networks and FBA. An explanation for generating a data structure with OR-modeling in the combination of those. More detailed information to be given in the Backgroud Information Section, guiding the readers who have knowledge of FBA and Association Network concepts to the Concept Implementation Section.

Usage of linear programming and generating sets of synthetic data allow comparing the statistical characteristics of their association network with those created from the real-world data set from steel manufacturing.

### 2.1.1 Steel Manufacturing Data Analysis

#### Association Networks

Beyond a simple network graph representation of historical production data, the formation of association networks is an insightful graph-based framework combining the tool: association rules and complex networks as Merten et al. (2020) performed in their article [6]. The relevant pipeline considers sequentially revealed events of a data set. It outputs a graph that unfolds the non-random occurrence of specific events together among the complete set that took place consecutively in the production period.

Assume we have a manufacturing data set with historical order, $D$, consists of $k$ sequences and $n$ events with mass values and sequence id's included as given in Table 2.1.

| Event_ID | Mass | Sequence_ID |
|---|---|---|
| 1 | 280 | 1 |
| 2 | 250 | 1 |
| 3 | 890 | 2 |
| 4 | 850 | 2 |
| 5 | 650 | 2 |
| 6 | 745 | 2 |
| 7 | 795 | 2 |
| 8 | 150 | 3 |
| ⋮ | ⋮ | ⋮ |
| n-4 | 940 | k-1 |
| n-3 | 540 | k |
| n-2 | 520 | k |
| n-1 | 630 | k |
| n | 610 | k |

Table 2.1: Arbitrarily Created Data Set $D$.

Examining the data set, one can say that the events with mass values: $890$, $850$, $650$, $745$ or $540$, $520$, $630$, $610$ are close to each other; thus, they are produced together and likely occur in the identical sequences among the complete data. In a further step, one can label the events mentioned above with a value interval (the so-called binning size) typical for every mass value with a tiny difference to each other. Binning generation for the events allows us to investigate them in a mass-production manner. Alternative binning methods will be addressed in the following subsection.

One can hypothetically argue that the information mentioned above patterns are probably deliberate planning choices based on the related constraints acting on the manufacturing process performance. However, forming prevailing arguments is not a simple task for large and complicated data sets. Such a real-life data set may consist of more than $300,000$ events likely to have various events aggregated randomly in its large sequence groups.

To distinguish random co-occurrences from meaningful ones in production sequences and assess the complexity of production patterns before creating the network graphs, we extract the association rule from the set of sequences. With a similar approach as Merten et al. (2020) applied [6], an association rule measure known as "Lift" was picked and calculated for every possible pairwise subset of events that occurred in the same sequences. By having a natural threshold of Lift measure $1$. The lift can be computed by the ratio of pair items joint probability divided by the multiplication of each item's marginal probability as

$$Lift(A \leftrightarrow B) = \frac{P(A,B)}{P(A) * P(B)}, \tag{6}$$

thus, in the case of $Lift(A \leftrightarrow B) > 1$, B occurs likely if A occurs whereas $Lift(A \leftrightarrow B) < 1$, B unlikely occurs if A occurs. Indication of random and non-random co-occurrences as $0$ and $1$ in an adjacency matrix will provide the data structure to form a network as shown in Fig. 2.1.

| Events | 1 | 2 | 3 | ... | n-1 | n |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | ... | 0 | 0 |
| 2 | 1 | 0 | 0 | ... | 0 | 0 |
| 3 | 0 | 0 | 0 | ... | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| n-1 | 0 | 0 | 0 | ... | 0 | 1 |
| n | 0 | 0 | 0 | ... | 1 | 0 |

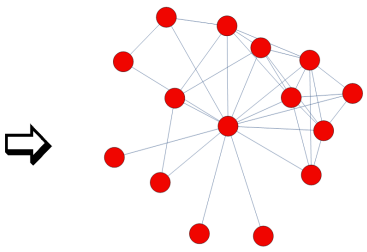


Figure 2.1: An Arbitrary Representation for Adjacency Matrix and Its Graph.

### Binning Methods

Alternative binning methods for the production events underlie the developed hypothesis of this thesis work: Non-random features of the association networks derived from these two methods explained in the introduction part.

Two distinguished network generation approach can be derived considering the alternative binning tools mentioned above: Fixed Step Size network (FSSn); it has graph nodes as binning members with equal bin sizes, and Fixed Bucket

Size network (FBSn); its nodes are binning members with an equal amount of events per bin. Forcing events to occur in the nodes with constant interval boundaries allows us to see how the aggregations take place within orders, whereas defining a typical bucket size for the network nodes results in arbitrary interval boundaries for each node; still, it allows to control their population.

*D* Labeled with FSS Bin Sizes

| Event ID | Mass | FSS Bin Size | Sequence ID |
|---|---|---|---|
| 1 | 280 | 200-299 | 1 |
| 2 | 250 | 200-299 | 1 |
| 3 | 890 | 800-899 | 2 |
| 4 | 850 | 800-899 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| n-2 | 520 | 500-599 | k |
| n-1 | 630 | 600-699 | k |
| n | 610 | 600-699 | k |

*D* Labeled with FBS Bin Sizes

| Event ID | Mass | FBS Bin Size | Sequence ID |
|---|---|---|---|
| 1 | 280 | 200-599 | 1 |
| 2 | 250 | 200-599 | 1 |
| 3 | 890 | 630-899 | 2 |
| 4 | 850 | 630-899 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| n-2 | 520 | 200-599 | k |
| n-1 | 630 | 630-899 | k |
| n | 610 | 600-629 | k |

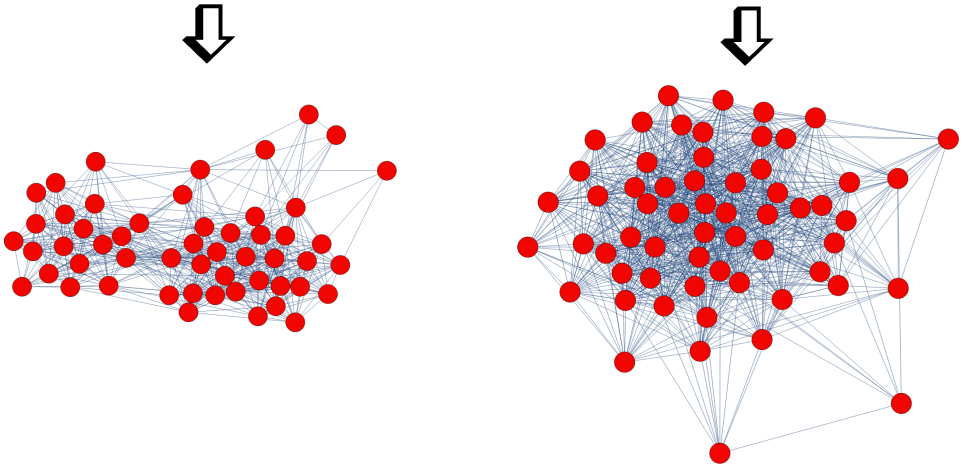


Figure 2.2: Graph Results For Two Different Network Approaches.

### Network Metrics Analysis

### Modularity Check

### Different Types of Null Model

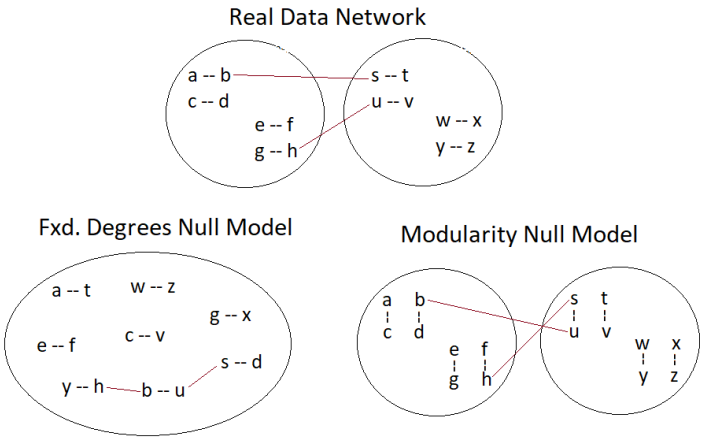



Figure 2.3: Formation of Different Null Models.

### 2.1.2 Flux Balance Analysis

The genome-scale integrated networks are necessary tools used by metabolic engineers on model generation, theoretical and computational analysis for microbial organisms. In addition, the network theory tools expand the feasible space for the following analysis techniques in the field.

Introducing stoic. matrix. Explain how the two graphs are formally obtained by manipulated the stoichiometric matrix, explain the metabolite-centric network is $S * S^T$ and binarized. In contrast, the reaction-centric network is $S^T * S$ and binarized. Introduce the general idea of FBA as an optimization scheme in a steady-state solution space. Although the networks shown in Fig. 2.4 do not contain any information about directionality or effectiveness of the reactions to the system, the set of rules take place in networks can be represented in more detail and stoichiometrically by an m-by-r matrix formulation (the so-called Stoichiometric Matrix $S$), whereas its column elements represent reactions that play a role in the chemical transformation, and its row elements represent metabolites as

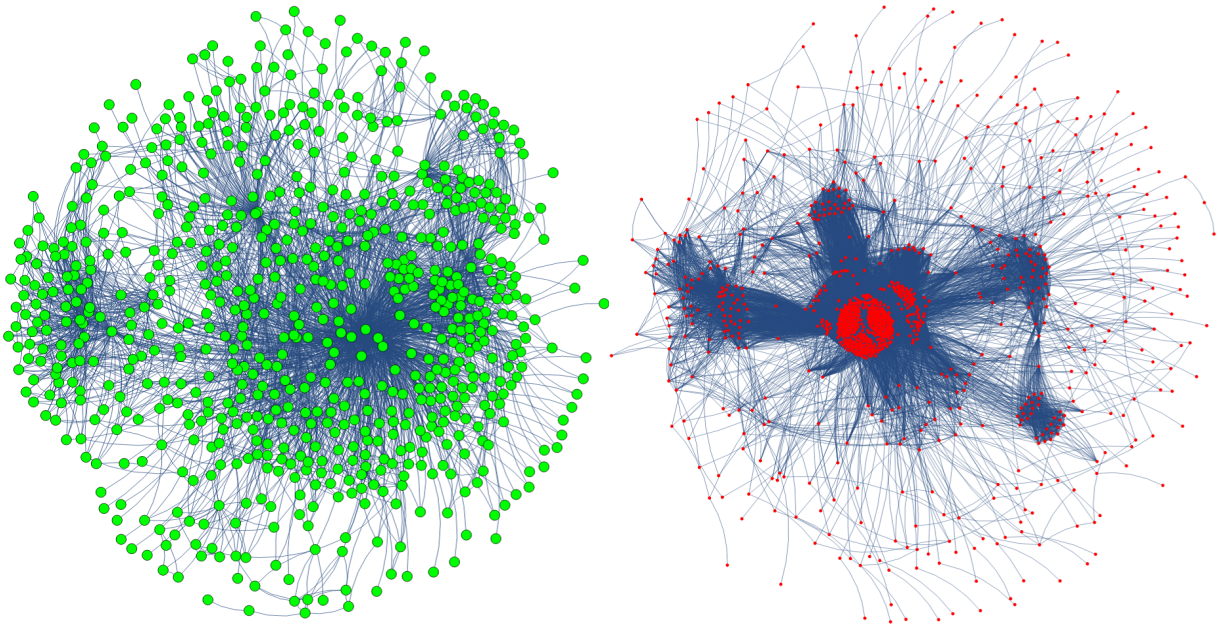
$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1r} \\ s_{21} & s_{22} & \dots & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & s_{mr} \end{bmatrix} = (s_{ij}) \in \mathbb{Z}^{mxr}, \tag{1}$$

Fig. 2.4 shows two differently constructed networks showing interactions between metabolites, intermediate or end products and metabolic reactions for a particular metabolism: Homo Sapien. In Fig. 2.4a the graph nodes stand for the metabolites, graph edges are the reactions. In contrast, in Fig. 2.4b the roles are reversed so that the graph edges represent the metabolites, and the graph nodes represent the reactions.

Studying biological metabolic systems, generated models to achieve cellular objectives like cell growth or ATP production brings the necessity of various tools to analyze reconstructed genome-scale networks. [1, 2]. One of the commonly used tools is Flux Balance Analysis (FBA). It is a constraint-based modelling approach to simulate microbial metabolisms and can be applied to biochemical-reaction networks containing the chemical transformations and flux exchanges in that particular network [3, 4].

while one can express the fluxes in a one-dimensional array (the so-called Flux Vector $V$) as

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_r \end{bmatrix} = (v_i) \in \mathbb{R}. \tag{2}$$



(a) Metabolic-centric Network

(b) Reaction-centric Network

Figure 2.4: Network Representations for Homo Sapiens Metabolic Model

$V$ contains flux exchange values for the corresponding reactions in the system and gives information about the flux distribution; hence, those can be both positive and negative real numbers. Definition of a mass balance ($S.V = 0$) constraint in the FBA enables us to analyze the metabolic network operations in a steady-state [3, 4].

$$S.V = \begin{bmatrix} s_{11}v_1 + s_{12}v_2 + \cdots + s_{1r}v_r \\ s_{21}v_1 + s_{22}v_2 + \cdots + s_{2r}v_r \\ \vdots \\ s_{m1}v_1 + s_{m2}v_2 + \cdots + s_{mr}v_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3}$$

The higher amount of metabolite consideration in the set of rules, $S$, in other words, the larger matrix size by its rows amount means the more complex organization structure taken into account while preserving the steady-state in the whole system.

More than one steady-state solution might be present since it is impossible to identify all constraints in a cellular system [3]. Therefore, one can formulate an optimization approach to identify reaction network steady-states that maximize the biomass [3, 4] or control the production of specific metabolites [5] within a defined objective function under the consideration of the system constraints. According to Price et al. (2004), there are three primary purposes to generate objective functions: to discover allowable characteristic properties

in the genome-scale network reconstruction; to mimic probable physiological functions like biomass or ATP production to be able to determine likely physiological states; and lastly, to design a genetic variant or sub-type to obtain a desired particular product [4].

The objective function can be thought as a production plan that gives an idea about the diversity of products that the relevant system can produce, and one can express its coefficients in a one-dimensional array as

$$O = \begin{bmatrix} o_1 & o_2 & \dots & o_r \end{bmatrix} = (o_i) \in \mathbb{R}. \tag{4}$$

As given in Eq.(5), the Objective Function, $Z$, rules the maximized output based on its non-zero coefficients, which are the decisive ones for the flux elements of $V$ to be considered.

$$Z = O.V = (o_1 v_1 + o_2 v_2 + \cdots + o_r v_r) \in \mathbb{R}_{\geq 0}. \tag{5}$$

Stoichiometry and mass-balance are the constraints introduced so far in Eq.(1) and Eq.(3). In addition, upper and lower bounds are introduced for particular fluxes in $V$ during the optimization process. The bounds are used in the reactions for uptake and secretion of any organic metabolite. In the uptake reactions, nutrients are transported to the inside of the metabolic network. In the secretion reactions, products are exported to the outside of the network. The rest of the fluxes in $V$ are used in the exchange reactions, namely the intermediate reactions in the network. The constraints are decisive on the reactions for uptake and secretion, whereas no limitation is considered in the exchange reactions. Quantification of imported nutrients and exported outputs (the so-called Resources and Wastes) by constraining them with upper and lower bounds to fulfil a single objective function goal might play a significant role in the optimization process.

The above-explained optimization process is a linear programming problem since the mass balance (Eq.(3)), the Objective Function (Eq.(5)), and linear equations formulate the upper & lower bounds for fluxes. The linear optimization result maximizes the structured Objective Function in the form of a flux distribution [3, 4]. Since each term in Eq.(5) is a produced biomass expression for the fluxes, the summation of those terms will give the overall growth of the system for a single network state.

### Fluxes for Uptake and Secrete Reactions

Let

$$V^* = (v_1^*, v_2^*, \dots, v_x^*) = (a_i \leq v_i^* \leq b_i) \in V \tag{7}$$

is a set of fluxes picked from $V$ to be limited with the bounds: $a_i$ and $b_i$ which are used in the reactions for uptake and secretion as previously introduced.

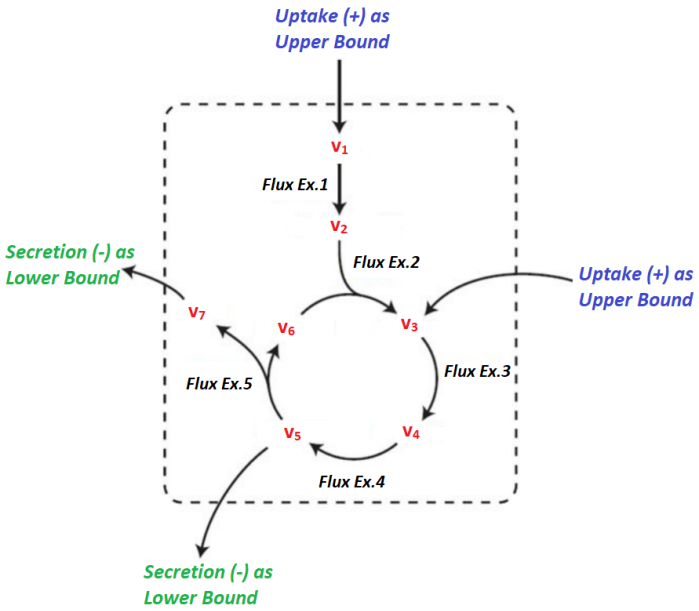



Figure 2.5: A Simplified Reaction-centric Network Sketch Shows The Reactions for Exchange, Uptake and Secretion.

### Upper and Lower Bounds

### Objective Functions

## 2.2 Applications and Results of the Developed Framework

### 2.2.1 Integration of Concepts

What happens here is to bring the simulated data into a format that is compatible with my analysis of the real-life events data.

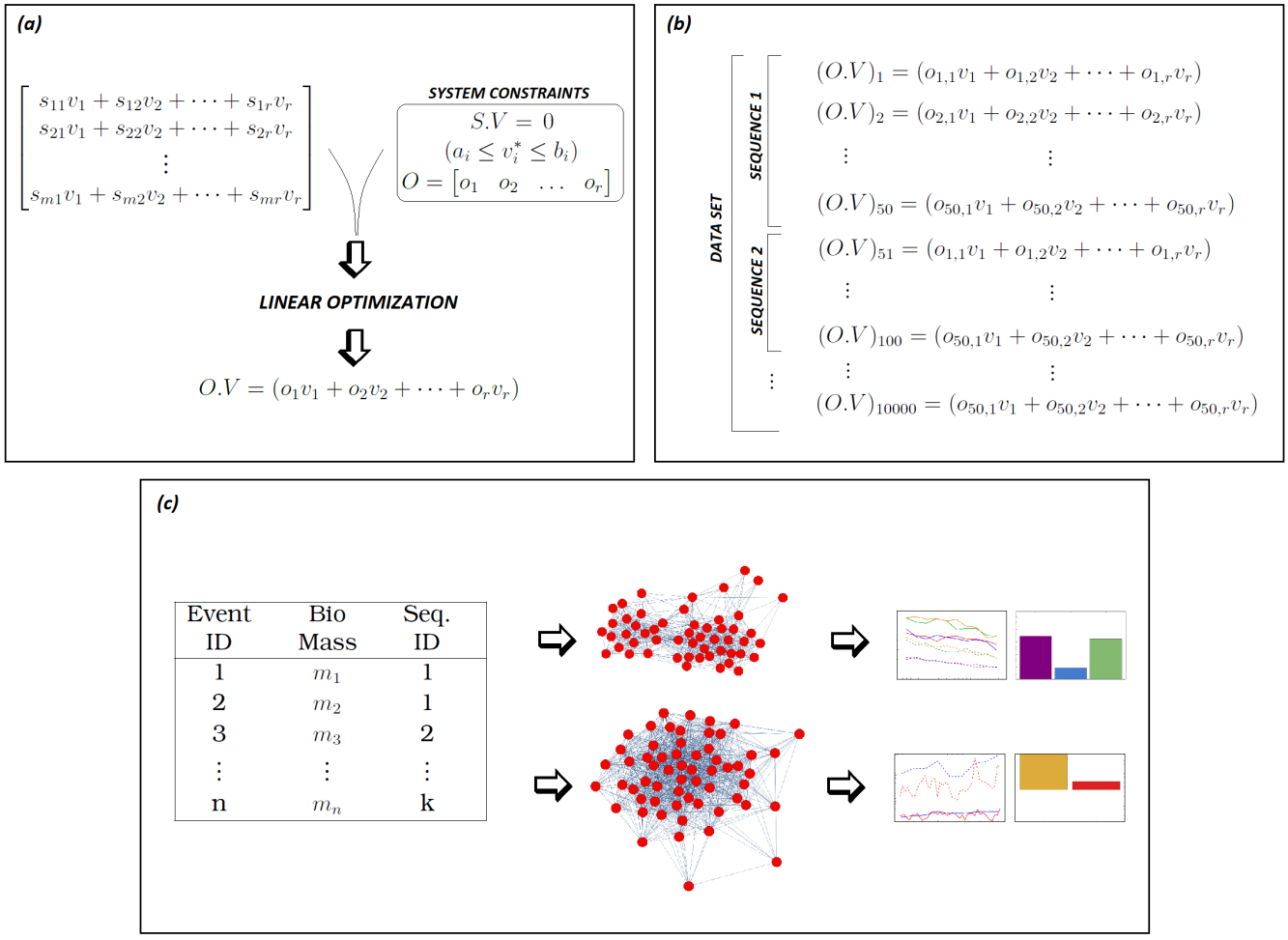



Figure 2.6: Complete Framework Sketch.

### 2.2.2 Simulation Results

Briefly explain *in silico analyses* attempts /numerical experiments from the generated data.

# 3 Conclusion And Outlook

## 3.1 Thesis Contribution

## 3.2 Outlook

# 4 Bibliography

[1] B. Kim, W. J. Kim, D. I. Kim, and S. Y. Lee, "Applications of genome-scale metabolic network model in metabolic engineering," *Journal of Industrial Microbiology and Biotechnology*, vol. 42, pp. 339–348, 03 2015.

[2] T. Hao, D. Wu, L. Zhao, Q. Wang, E. Wang, and J. Sun, "The genome-scale integrated networks in microorganisms," *Frontiers in Microbiology*, vol. 9, p. 296, 2018.

[3] K. J. Kauffman, P. Prakash, and J. S. Edwards, "Advances in flux balance analysis," *Current Opinion in Biotechnology*, vol. 14, no. 5, pp. 491–496, 2003.

[4] N. D. Price, J. L. Reed, and B. . Palsson, "Genome-scale models of microbial cells: evaluating the consequences of constraints," *Nature Reviews Microbiology*, vol. 2, no. 11, pp. 886–897, 2004.

[5] A. Varma, B. W. Boesch, and B. O. Palsson, "Biochemical production capabilities of escherichia coli," *Biotechnology and Bioengineering*, vol. 42, no. 1, pp. 59–73, 1993.

[6] D. Merten, M.-T. Hütt, and Y. Uygun, "A network analysis of decision strategies of human experts in steel production," *submitted to IISE Transactions*, 2020.

# 5 Supplements