METABOLIC BEHAVIOR PREDICTION UNDER GENOME-SCALE TRANSCRIPTION PERTURBATIONS

A Thesis

by

PUHUA NIU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Xiaonng Qian
Co-Chairs of Committee,	Nicholas Duffield
Committee Members,	Jiang Hu
	Anxiao Jiang
Head of Department,	Miroslav Begovic

December 2021

Major Subject: Computer Engineering

Copyright 2021 PUHUA NIU

ABSTRACT

Advances in bioengineering have enabled numerous bio-based commodities. Yet most traditional approaches do not extend beyond a single metabolic pathway or do not attempt to modify gene regulatory networks in order to buffer metabolic perturbations. This is despite access to near universal technologies allowing genome-scale engineering. To help overcome this limitation, we have developed a pipeline enabling analysis of Transcription Regulation Integrated with MEtabolic Regulation (TRIMER). TRIMER utilizes a Bayesian network (BN) inferred from transcriptomic data to model the transcription factor regulatory network. TRIMER then infers the probabilities of gene states that are of relevance to the metabolism of interest, and predicts metabolic fluxes resulting from deletion of transcription factors at the genome scale. BN-based modeling of transcription regulation can faithfully capture global dependencies in the network and allow more flexible transcriptional changes, thereby enabling one to predict condition-dependent metabolic behaviors for more general genetic engineering strategies. Additionally, we have developed a simulation framework to mimic the TF-regulated metabolic network, capable of generating both gene expression states and metabolic fluxes, thereby providing a fair evaluation platform for benchmarking models and predictions. Here, we present this computational pipeline. We demonstrate TRIMER's applicability to both simulated and experimental data and show that it outperforms current approaches on both data types.

DEDICATION

To my parents, instructors for their support and help.

ACKNOWLEDGMENTS

First, I would like to thank my parents for financially supporting my study in Texas A&M University for the past my two years.

Next, I would like to thank my advisors Prof. Xiaoning Qian and Prof. Byung-Jun Yoon. They are always patient in discussing with me for progress and gave me tremendous help for my research work. I will not be able to proceed my research smoothly without their kind guidance.

Lastly, I am thankful to Prof. Nick Duffield, Prof. Jiang Hu and Prof. Anxiao Jiang for their efforts and kind help in serving as committee members.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was mainly completed in conjunction with of Prof. Xiaoning Qian (advisor) and Prof. Byung-Jun Yoon of the Electrical and Computer Engineering Department.

The experimental validation and data collection were conducted by Drs. Maria J. Soto and Ian Blaby of the Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory.

All other conducted research work for the thesis was finished by the student independently.

Funding Sources

The materials presented in this thesis are based upon the work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract number DE-SC0012704. The work in this thesis is partially supported by the National Science Foundation under Grant CCF-1553281.

NOMENCLATURE

TRIMER	Transcription Regulation Integrated with MEtabolic Regulation
BN	Bayesian network
TF	Transcription Factor
FBA	Flux-Balance Analysis
PROM	Probabilistic Regulation Of Metabolism
GPR	Gene-Protein-Reaction
KO	KnockOuts
TRN	TF-Regulated gene Network
DAG	Directed Acyclic Graphs
MLE	Maximum Likelihood Estimate
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion
LP	Linear Programming
MOMA	Minimization Of Metabolic Adjustment
ROOM	Regulatory On/Off Minimization
FVA	Flux Variability Analysis
CMPI	Common Mathematical Programming Interface
TF	$\in \{0,\ 1\}$, transcription factor expression state
g	$\in \{0,\ 1\}$, gene expression state
G	index set of genes
$p(\vec{g}\|\vec{TF})$	$\in\ (0,\ 1)$, conditional probabilities of gene states given TF states

r	index of metabolic reaction
R	index set of metabolic reactions
$G(r)$	index set of genes that regulate reaction r
v_r	$\in [lb_r,\ ub_r]$, steady-state metabolic fluxes of reaction r
lb_r	$\in \mathbb{R}$, flux lower bound of reaction r
ub_r	$\mathbb{R}^+$, flux upper bound of reaction r
S	a real-valued matrix, stoichiometric coefficients in the metabolic reaction network
$v_{max}(r)$	$\in \mathbb{R}^+$, maximum flux of reaction r, estimated via flux variability analysis
$v^0(r)$	$\in [lb_r,\ ub_r]$, wild-type steady-state metabolic fluxes of reaction r

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND BACKGROUNDS

1.1 Introduction

There has been extensive research in *in silico* modeling and prediction of genome-scale metabolic behavior, mostly focusing on mutant strain design with metabolic reaction network modeling [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. However, living systems involve complex and stochastic processes arising from interactions between different types of biomolecules. For more accurate and robust prediction of target metabolic behavior under different conditions or contexts, not only metabolic reactions, but also the integration of genetic regulatory relationships involving transcription factors (TFs) that may regulate metabolic reactions, should be appropriately modeled. Due to the increasing computational complexity when considering multiple types of biomolecules in one computational system model, often transcription regulation has been integrated via "transcriptional regulatory constraints" with various heuristics for flux-balance analysis (FBA) of metabolic networks [14, 15, 16, 17, 18, 19, 20]. Many of these computational tools were often only validated for selected model organisms with curated data and network models. To generalize these integrated hybrid models for different organisms, the reproducibility of the results require careful validation, for example, starting from simulated ground-truth models.

Probabilistic Regulation Of Metabolism (PROM) [21] introduced probabilistic modeling of transcription regulation for better integration with condition-specific metabolism. PROM can be considered as one of the first integrated transcriptional-metabolic network models that take advantage of both existing prior knowledge and gene expression data. Specifically, conditional probabilities were inferred by microarray data analysis for annotated TF-(target gene)-reaction interactions to incorporate transcriptional regulation information in genome-scale metabolic network analysis under different conditions or contexts. IDREAM [22], an updated version of PROM, additionally allowed modeling subtle growth defects to further improve metabolic flux predictions. Recently, an algorithm called OptRAM was developed based on IDREAM for designing optimized strains

for ethanol overproduction in yeast [23].

The essential idea of PROM and its extensions is to infer the TF-gene conditional probabilities of the form Pr(gene=`ON/OFF`| TF=`ON/OFF`) so that metabolic reactions regulated by specific genes – for example, through the specific enzymes manifested as gene-protein-reaction (GPR) rules – can be modeled dependent on either genotypic or environmental changes by adjusting the reaction flux constraints in the FBA formulation for metabolic modeling. Although it is computationally desirable to simplify the TF regulatory roles by introducing TF-gene conditional probabilities estimated by local frequentist estimates based on gene expression profiles, global TF-gene dependency structures may not be well captured. The existing models are also limited in the sense that only conditional probabilities based on univariate conditions were modeled. More flexible modeling that enables predictions with more complicated condition changes, for example, multiple TF knockouts when designing mutant strains, is still lacking in the literature.

The main contribution of this work is to introduce a new flexible genome-scale simulation and analysis pipeline, *TRIMER*—Transcription Regulation Integrated with MEtabolic Regulation, for integrative systems modeling of TF-regulated metabolism. Specifically, a Bayesian network (BN) is employed in TRIMER, instead of local TF-gene conditional probabilities or transcriptional regulatory constraints, thereby aiming at effectively capturing the global transcriptional regulatory relationships that may affect metabolism. Through this BN, the influence of transcription regulation (and its changes) on metabolic behavior under different conditions will be manifested more accurately via more flexible conditional probability inference which is linked to metabolism through the prior knowledge on TF-gene-reaction interactions. In the prediction mode of TRIMER for a given model organism, expression data, and prediction tasks, a BN will be first inferred based on gene expression profiles with the prior knowledge on TF-gene-reaction interactions. Based on the inferred Bayesian network, given a condition (for example, multiple TF knockouts), we can infer the corresponding probabilities of gene states and consequently flux predictions can be performed by corresponding *in silico* metabolic models.

In addition to the modeling and analysis functionalities in TRIMER, we have also developed

a simulator that simulates the TF-regulated metabolic network, which can generate both gene expression states and metabolic fluxes from a given transcriptional-metabolic hybrid model. Such a simulator provides a fair performance evaluation platform to help better benchmarking and validating new model inference and flux prediction methods in computational systems biology.

1.2 Backgrounds

We here provide the brief review on the basics of metabolic engineering based on flux balance analysis (FBA) [24, 12], probabilistic regulatory network modeling including both simplistic conditional probability models as adopted in Probabilistic Regulation Of Metabolism (PROM) [21] as well as its extension, IDREAM [22], and more general Bayesian network (BN) modeling [25].

1.2.1 Metabolic engineering for mutant strain design

Since it has been proposed in [3, 24, 12], FBA, as a simplified network analysis model for metabolic flux analysis, has been widely adopted for steady-state flux analyses by assuming the balance of production and consumption fluxes of metabolic reaction network models. Mathematically, with the prior stoichiometry knowledge, FBA assumes that the weighted sum of reaction fluxes, denoted by the vector $\vec{v}$, based on calibrated stoichiometric coefficients S, is 0: $S\vec{v} = 0$. Such a steady-state flux analysis can be performed by assuming that the corresponding wild-type microbial species always optimizes for its growth:

$$\begin{aligned} \max_{\vec{v}} \quad & \text{biomass}(\vec{v}) \\ \text{s.t.} \quad & S\vec{v} = 0; \\ & lb_i \leq v_i \leq ub_i, \quad \forall i \in \{1, \ldots, m\}, \end{aligned}$$

where v_i, $1 \leq i \leq m$ denotes the flux value for the ith metabolic reaction of the total m reactions in the metabolic network, and S an $m \times n$ stoichiometric matrix involving all the n metabolites in the given metabolic reaction network model. The biomass production flux: $\text{biomass}(\vec{v}) = \sum_{j \in I_{\text{biom}}} c_j v_j$ is based on the annotated set of reaction indices, I_{biom}, involving the metabolite

precursors that contribute to the biomass production in FBA with the corresponding given weights c_j [6]. Each reaction flux is bounded by the corresponding lower and upper bounds lb_i and ub_i.

For wild-type microbial strains, a common assumption is that their steady-state flux values follow an optimal distribution that maximizes the biomass production rate. The steady-state flux distribution can be approximately solved as a linear programming (LP) problem to maximize the biomass production flux subject to the FBA stoichiometry constraints as the above formulation.

However, when modeling mutant strains, researchers found that the biomass maximization assumption for wild-type strains may not approximate the steady-state fluxes well. To achieve better agreement with experimental observations, approximation formulations of knockout metabolic fluxes undergoing a minimization of metabolic adjustment (MOMA) process [6] or by the regulatory on/off minimization (ROOM) [8] have been proposed to address the long-term post knockout metabolic flux distribution predication problem.

Existing microbial strain design formulations based on these formulations often ignore changing conditions or contexts due to interventions. They search for the knockouts to optimize the desired flux predictions by bi-level optimization formulations to make sure about the mutant survival at the same time. One of such representative methods is OptKnock [7]. However, when modeling condition/context dependency in hybrid models involving transcriptional regulations, such methods are not directly applicable.

1.2.2 PROM: A brief review

PROM aims to predict metabolic fluxes of the knockout mutants in transcription factor (TF)-regulated metabolic networks. Specifically, PROM is built upon the FBA framework. PROM first estimates the probability of "reaction-targeted" gene expression (ON/1 or OFF/0) given transcription factor (TF) expression PR$(gene = 1|TF = 0)$ based on a certain set of microarray expression data using annotated TF-gene-reaction interactions. Based on that, PROM solves the following LP

problem given transcription factor knockout (KO) perturbations:

$$\begin{aligned} \max_{\vec{v},\alpha,\beta} \quad & \text{biomass}(\vec{v}) - \kappa(\alpha+\beta) \\ \text{s.t.} \quad & S\vec{v} = 0; \\ & lb'_i - \alpha \le v_i \le ub'_i + \beta, \quad \forall i, \\ & \alpha \ge 0; \quad \beta \ge 0, \end{aligned}$$

where α and β can be considered as slack variables and lb'_i and ub'_i are perturbed flux bounds based on transcriptional regulations. In particular, Flux Variability Analysis (FVA) [26] is performed together with network-based metabolic behavior prediction [27] to get the minimum and maximum fluxes. The inferred conditional probabilities due to a specific TF KO will then be multiplied based on the transcriptional regulations on the corresponding metabolic reactions, for which the metabolic models from either the KBase [28] or `COBRA` toolbox [29] can be used.

PROM consists of multiple steps from microarray data analysis, flux bound manipulations, and FBA based on these steps with the aim to have their model prediction to better fit with the flux measurements at different conditions.

1.2.3 IDREAM

IDREAM is an improved version of PROM. They differ only in the way of conditional probability computing. While PROM uses experimentally verified interaction list directly (i.e. EcoMAC), interactions used in IDREAM are further pruned by identifying the common interactions with respect to the computationally derived ones by EGRIN [30]. EGRIN is constructed using two existing computational tools: cMonkey [31] and Inferelator [32]. Given gene expression data, corresponding genes are first grouped into clusters by cMonkey and then regulators in each cluster are further identified by Inferelator. After obtaining the common interactions, conditional probabilities corresponding to these interactions are derived by bootstrapping using Inferelator trained on subsets of gene expression data.

1.2.4 Modeling transcription regulations using Bayesian Network

A Bayesian network (BN) is a probabilistic graphical model (PGM) that can be used to represent the joint probability distribution of a set of variables $\mathbf{X} = \{X_1, \cdots, X_n\}$, whose dependencies are described by a directed acyclic graph (DAG) $\mathcal{G}$. Each node in the DAG corresponds to a variable $X_i \in \mathbf{X}$ of interest, and a directed edge $X_j \rightarrow X_k$ represents the possible causal relationship between the variables X_j and X_k. Following the topology of the DAG, the joint distribution of $\mathbf{X}$ can be written as a product of conditional probabilities:

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i|Pa(X_i)) \tag{1.1}$$

where $p(X_i|Pa(X_i))$ is the conditional distribution function of X_i given the set of variables $Pa(X_i)$, which denote the set of its parent nodes in $\mathcal{G}$. In BN, graph topology captures the complex dependencies among the variables, resulting in a compact representation of the joint probability distribution of $\mathbf{X}$ by factorizing it into a product of local probability models as in (1.1). This compact representation reduces data requirements for learning the distribution from data and also greatly enhances the computational efficiency of making probabilistic inference based on the distribution [25].

The main novelty of the hybrid models in our TRIMER is to model the TF-regulated network (TRN) using the more general Bayesian network model to better capture regulatory relationships. Unlike PROM, in which TF-regulations are represented simply by inferring the maximum likelihood estimates (MLEs) of the involved conditional probabilities $p(gene = 1|TF = 0)$, TRIMER adopts a full-fledged BN to capture the transcription regulations. Based on the available gene expression data, we can infer the BN by first inferring the structure of the network and then estimating the parameters of the local probability model (i.e., conditional probability distributions). In this manner, TRIMER can better capture both the local and global dependencies between TFs and genes, thereby better model the TF knockout effects on metabolic fluxes. Furthermore, the BN enables the incorporation of available prior knowledge regarding TF regulations, enhancing

the quality of the inferred network compared to a solely data-driven inference approach.

2. MATERIALS AND METHODS

We introduce the main components of TRIMER organized in two major modules – namely, the *transcription* regulation network module and the *metabolic regulation* module – that are integrated within a unified interacting framework (Figure 2.1). The proposed hybrid model enables condition-dependent transcriptomic and metabolic predictions for both wild-type and TF-knockout mutant strains, through general Bayesian network (BN) modeling of transcriptional regulations. We also provide the details of our TF knockout experiments from which the experimentally observed fluxes validate the *in silico* flux predictions made by TRIMER.

2.1 TRIMER: Transcription Regulation Integrated with MEtabolic Regulation

Before presenting each component, we first provide a brief overview of our proposed hybrid TF-regulated metabolic network model, *TRIMER*: TRIMER differs from the existing methods in the way of systematic prediction of effective intervention strategies when applied to the transcription regulatory network for regulation of metabolism. Specifically, TRIMER is based on a Bayesian Network (BN) for learning transcription regulation from gene expression data. Instead of utilizing simple conditional probabilities of the form Pr(gene=`ON/OFF`| TF=`ON/OFF`) as in PROM [21], the BN can be used to determine a probabilistic inference of the effect of alterations (e.g., gene deletions) of multiple TFs (or genes). While the framework presented is independent of the nature of TF engineering, we focus herein on gene deletions (i.e. knockouts (KO)). Furthermore, BN modeling enables intuitive incorporation of prior knowledge (e.g., pathways or pairwise regulatory relationship between genes) for learning the *TF-Regulated gene Network (TRN)*.

In TRIMER, a BN is trained from the gene expression data to model the joint distribution for all the relevant TFs and genes, where the resulting BN can be subsequently used to infer the steady-state conditional probabilities of the form Pr(gene(s)|TF(s))$= p(\vec{g}|\vec{TF})$ – i.e., the probability of gene states given the states of TFs of interest. For example, we can use the BN to estimate the probability that a target gene known to regulate a specific metabolic pathway is induced given

that expression of one or more TFs is abolished by gene deletion. The estimated probabilities can be used to constrain the metabolic reaction fluxes of interest, based on which the flux changes of selected metabolites resulting from the genetic alteration (e.g., TF gene deletion) can be predicted via flux balance analysis (FBA). The gene-protein-reaction (GPR) rules, which inform us of the respective metabolic pathways regulated by different genes, are used to link the transcription regulation modeled by the BN with the metabolic regulation simulated by FBA.

TRIMER, which jointly models transcription regulation and metabolic regulation via the aforedescribed hybrid approach, allows us to assess the efficacy of potential TF engineering strategies

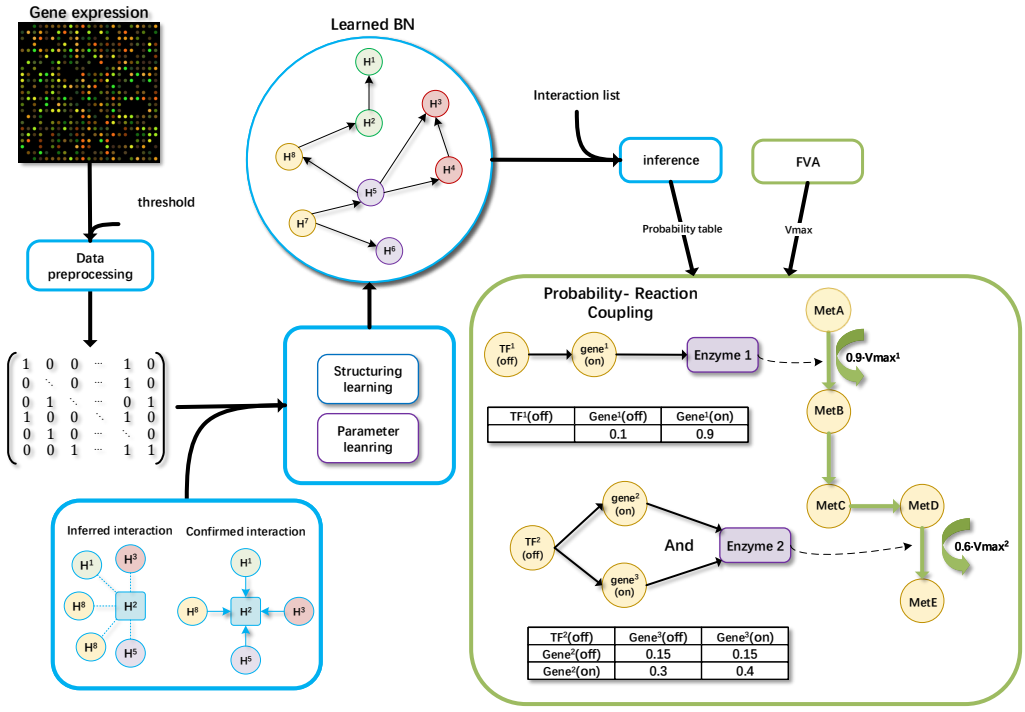


Figure 2.1: Illustrative overview of TRIMER. Gene expression data are used to infer the Bayesian network (BN) modeling the transcriptional regulations with the prior knowledge on molecular interactions. The impacts of transcription factor knockout on downstream target genes that affect metabolic pathways are inferred using the BN. The estimated probability that a given target gene being turned on modulates the constraints in the flux variability analysis (FVA), resulting in probabilistic metabolic predictions. Each module component in TRIMER has the detailed explanations in the flowchart in Figure 2.2 with the matched box boundary colors for the corresponding transcription regulation and metabolic flux prediction modules.

and identify the optimal strategy for modulating the metabolic fluxes of interest. The desirability of a given genetic alteration can be assessed *in silico* using TRIMER, which can be validated through actual TF deletion and screening experiments in the laboratory.

Figure 2.1 provides a high-level overview of TRIMER, illustrating its main workflow. As shown in this diagram, TRIMER consists of two main modules: (1) the BN module for modeling and inference of transcription regulation and (2) the metabolic flux prediction module for estimating the impact of alterations in the TRN on the metabolic outcomes. The two modules are linked to each other by the GPR rules. Furthermore, Figure 2.2 depicts the overall workflow of module components with explanations in TRIMER, including the interconnections among the computational modules that comprise TRIMER.

2.2 Transcription Regulation Inference in TRIMER

In this secction, we provide a detailed description of each components of transcription regulation network module in TRIMER.

2.2.1 Gene expression data preprocessing

The gene expression data need to be discretized for BN learning as in TRIMER, the *TF-Regulated gene Network (TRN)* concerns 'ON/OFF' states of TFs and genes in the network. In our implementation, quantile normalization is first applied to raw data. Then the threshold for a given quantile value is computed and data is binarized according to the threshold. The choice of the quantile value can be either set manually or be similarly determined as in PROM [21]. In other words, we search for the best value based on the prediction performance of the learned BN. Based on the results of our experiments, the suggested quantile value for thresholding is in the range of $[0.3, 0.4]$.

2.2.2 BN learning

The key component of TRIMER is to model the genome-scale TF-regulated gene network by a Bayesian network (BN) learned from discretized gene expression. This BN is expected to capture the interactions between regulators (TFs) and target genes. For this purpose, we have integrated

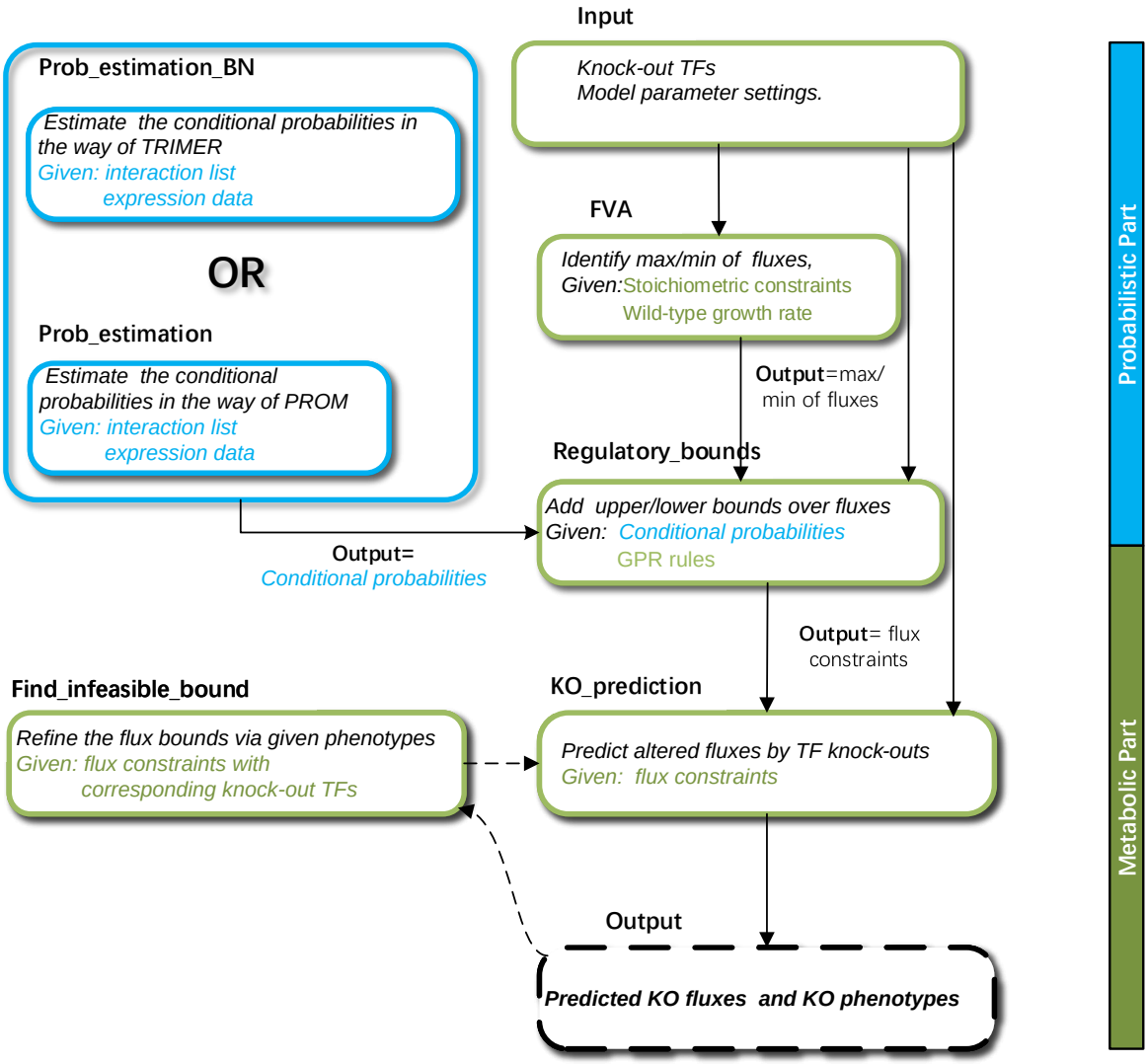


Figure 2.2: TRIMER flow-chart, where the explanations of the major computational module component for both transcription regulatory network modeling and metabolic flux prediction in TRIMER, together with their interconnections are illustrated. The blue boxes denote the module components for transcription regulation and the green ones denote the metabolic reaction network model components.

`bn-learn`, a Bioconductor package for Bayesian network modeling of biological networks [33]. A naive way to learn a BN from available observed gene states is to search over the space of all possible directed acyclic graphs (DAGs) and identify the one that optimizes a given objective function

evaluating the goodness of fit. However, the search space of BN model structures grows exponentially with the number of variables (nodes in the BN). Without restricting the BN structures, the BN learning can easily become infeasible even when considering only a dozen variables.In our experiments, we implemented two structure learning strategies, *tree-based* search for learning tree-based BN in a restricted family of Chow-Liu trees [34] and Tabu search [35], a greedy algorithm for learning *general BNs* that incorporate prior knowledge of TF-gene interactions. After finding the desired BN structure, BN model parameters are estimated by maximum likelihood estimates (MLEs).

In TRIMER, we have implemented Chow-Liu-tree-based BN learning for tree-based search. For learning general BNs, Tabu search, a modified greedy hill-climbing optimization strategy, is implemented in `bn-learn` as the search method based on a chosen score function, for example, either Bayesian information criterion (BIC) or Akaike information criterion (AIC). In our implementation, we further explore the proposed bootstrap resampling in [36] to learn a more robust structure. Specifically, we search for high-score BN structures by bootstrapping multiple expression samples from the given total samples (simulated or from expression databases). The inferred edges present in at least N% of the learned BNs are finally included in the final structure. N is a threshold value, which is determined automatically as described in [36]. Such a model averaging strategy helps to establish the significance of the edges in the final "average" structure for robustness against the potential data uncertainty and scarcity.

We further note that in our experiments, to restrict the search space of general BN structure learning, only experimentally confirmed gene-gene interactions are considered as candidate edges in BNs. For *Escherichia coli*, we have employed the interactions archived in RegulonDB [37]. When needed, interaction inference and validation methods, such as `GENIE3` [38], `TIGRESS` [39], or `Inferelator` [32], can also serve as the prior knowledge to extend the search space for structure learning.

2.2.3 Gene state inference

Once we have learned a BN, we can infer all the relevant conditional probabilities Pr(gene(s)|TF(s))= $p(\vec{g}|\vec{TF})$ that regulate the genome-scale metabolic network (iAF1260 for *E. coli* for example) so that for TF-knockout mutants, the conditional probabilities Pr(gene(s)|TF(s)) can model the effect of TF knockouts over the regulated target genes and therefore the corresponding metabolic fluxes at the genome scale. To do that in TRIMER without incurring high computational cost to exhaust all potential Pr(gene(s)|TF(s)) for metabolism regulation, we only focus on the TF-target interaction list to determine which genes can be affected (annotated as target genes) when one or multiple TFs are knocked out. Generally speaking, due to potential I-equivalent classes when learning BNs from data, determining the exact causal relationships from the learned BN structure is difficult. We rely on the annotated TF-gene interaction list (in RegulonDB for example). The Kolmogorov-Smirnov test [40] is performed to select significantly coupled TFtarget pairs in the interaction list. Then the filtered list is further pruned by removing the pairs that are d-separated in the learned BN. In cases when multiple TFs are knocked out at the same time, the list of the affected genes is the union of the target gene lists corresponding to each knockout TF in TRIMER. In addition, we only care about the probabilities that will affect metabolic reactions so that only the target genes that are associated with the metabolic reactions as described by the gene-protein-reaction (GPR) rules will be considered. Given this pruned interaction list, TRIMER infers corresponding conditional probabilities by BN inference algorithms. In TRIMER, exact inference is performed by the integrated package `gRain` [41] and approximate inference in `bn-learn` [33] can also be directly utilized for computational efficiency.

2.3 Metabolic Flux Prediction in TRIMER

The workflow of connecting the BN inference and metabolic flux prediction modules is illustrated in the flowchart shown in Figure 2.2. In the following text, we detail the corresponding TRIMER implementations for the constituting module components for the metabolic flux prediction module.

2.3.1 Construct transcriptional constraints over flux variables

The first module component for metabolic flux prediction is to integrate transcriptional changes into metabolic network modeling. Metabolism regulation in TRIMER is achieved by integrating the inferred conditional probabilities under different conditions from the BN to construct constraints for the corresponding metabolic reaction fluxes according to the GPR rules. From the BN learning and inference module, we derive a list of conditional probabilities associated with the corresponding metabolic reactions in the metabolic network model. Similar as in PROM, these probabilities together with the fluxes bounds estimated via flux variability analysis (FVA) [26] are used to constrain the reaction flux bounds through GPR rules. FVA helps determine alternative optimal solutions for the constraint-based linear programming formulation of FBA by screening the corresponding polygon boundaries of the feasible solution space, which identifies the minimum and maximum possible fluxes through a reaction in the metabolic model. To integrate transcription regulation into the metabolic models, GPR rules are represented as Boolean expressions associated with corresponding reactions to describe the nonlinear relationships between genes and reactions. In TRIMER, we have implemented a general platform as in `TIGER` [10] to convert the conditional probability values into linear constraints over flux variables and integrate them with the metabolic model in `COBRA` [29] for flux prediction.

We have adopted two ways to derive the updated reaction flux constraints according to the two ways of inferring conditional probabilities based on the learned BN.

The first way is the same as the one adopted in PROM. Suppose there are M genes denoted as $G = \{g_1, \ldots, g_m, \ldots, g_M\}$ that are regulated by the corresponding TF(s). Then via the provided GPR rules in the `COBRA` model, we can find the corresponding affected reactions denoted as $R = \{r_1, \ldots, r_n, \ldots, r_N\}$. For each $r_n \in R$, we can find a subset of regulating genes in G, denoted as $G(r_n)$, based on the corresponding GPR rules. With the corresponding TF knockout mutants, the

reaction flux bounds are then adjusted in the following way:

$$\begin{aligned} ub_{r_n} &= \min_{g \in G(r_n)} \{p(g=1|TF=0)\} \times v_{max}(r_n); \\ lb_{r_n} &= \min_{g \in G(r_n)} \{p(g=1|TF=0)\} \times (-v_{max}(r_n)), \end{aligned} \tag{2.1}$$

where $v_{max}(r)$ is estimated by FVA for reaction r. An example is given in the Appendix A to illustrate this operation.

In TRIMER, we have also implemented a more general way for integrating both the probabilities and the GPR rules into the flux constraints, so we can obtain the joint probabilities of the states of multiple genes regulating the same reaction, instead of simply combining the conditional probabilities for individual genes in the heuristic manner as in the previous approach. The reaction flux bounds can be set by directly multiplying the maximum flux with the sum of all probabilities with the corresponding gene states that affect the corresponding reaction according to the GPR rules:

$$\begin{aligned} ub_{r_n} &= \sum_{Bool(\pi)=1} p(G(r_n)=\pi|TF=0) \times v_{max}(r_n); \\ lb_{r_n} &= \sum_{Bool(\pi)=1} p(G(r_n)=\pi|TF=0) \times (-v_{max}(r_n)), \end{aligned} \tag{2.2}$$

where $Bool(\pi)=1$ denotes that the corresponding GPR rules between the genes and the reaction are satisfied with the state profile π representing the corresponding states of genes. Note that the above flux constraints are directly derived based on the conditional joint probabilities of all the regulating genes for a given reaction r_n. One illustrative example is given in the Appendix A.

Finally, we note that the above equations can be extended to experiments that involve multiple TF knockouts, enabled by flexible BN-based transcription regulation modeling.In the remaining content, we use TRIMER-C to denote the TRIMER implementations including the flux constraints computed in the first way and TRIMER-B for the second way.

2.3.2 Data structure for metabolic reaction network

TRIMER adopts a data structure organized in a similar way as that in the `TIGER` package [10] to represent the TF-regulated metabolic reaction network. In this data structure, constraints, lower/upper bounds, variable types of the reaction flux variables provided in the model files from the `COBRA` toolbox, together with the corresponding information for additional variables are represented and stored in a unified framework. As shown in the data structure representation in Figure 2.3, fields `obj`, `varnames`, `vartype`, `lb`, and `ub` correspond to the coefficient vector used in the objective function of the corresponding metabolic network model formulations, such as FBA or ROOM [8]; descriptive names of involved variables; variable types; and lower/upper bounds. Fields `A`, `b`, `ctype` store all the information about the constraints over variables, including the specific parameter setups in the corresponding metabolic model under given conditions. Stoichiometry constraints $S\vec{v} = 0$ for flux variables $\vec{v}$ and all the other additional linear constraints over the decision variables in the data structure specified by users are collected into the matrix A and vector b and represented as a single expression $A\vec{var}\ op\ b$, where $\vec{var}$ denotes all the variables included in TRIMER and op is an operator vector constituting $\{'>','<','='\}$ stored in the field `ctype`. In TRIMER, build-in functions are implemented to provide a standardized way to build the aforementioned data structure. One example can be found in the Appendix A.

2.3.3 Metabolic flux prediction

In TRIMER, we have implemented two variations of the FBA formulations for metabolic flux prediction, in addition to the standard FBA formulation with biomass as the objective function as described earlier. When predicting corresponding reaction fluxes of knockout mutants for all these formulations, let $\vec{v}$, $\vec{v}^0$, ub, $lb \in \mathbb{R}^m$ and I, denote the flux variables, wild-type optimal flux vector (the fluxes obtained by performing the standard FBA with the initial flux bounds given by the `COBRA` toolbox), flux upper and lower bounds for all the m reactions, as well as the set of reactions affected by the corresponding TF knockout(s). For each affected reaction, the reaction flux bounds are modified as described previously. With that, the optimization formulation for

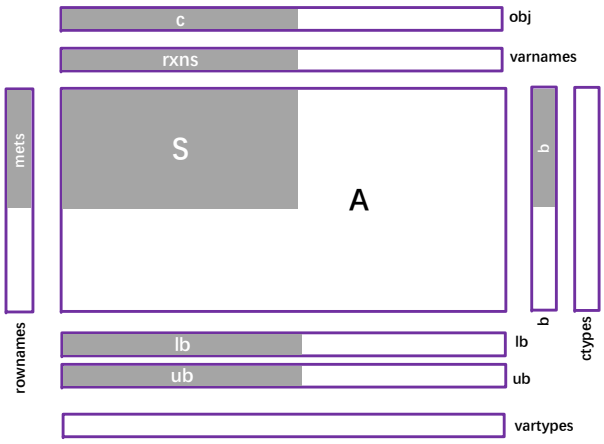


Figure 2.3: Metabolic models represented as Matlab data structures: Boxes indicate size and orientation of the fields. Black text denotes the corresponding field names. Gray areas contain data from the metabolic model, with white text indicating the relevant field names.

mutants with the biomass objective and slack variables allowing violating flux bound constraints, denoted as **sFBA**, is as follows:

$$
\begin{aligned}
\max_{\vec{v},\vec{\alpha},\vec{\beta}} \quad & \text{biomass}(\vec{v}) - \vec{\kappa}^{\mathrm{T}}(\vec{\alpha}+\vec{\beta}) \\
\text{s.t.} \quad & S\vec{v} = 0; \\
& lb_i - \alpha_i \le v_i \le ub_i + \beta_i, \quad \forall i \in \{1,\ldots,m\}; \\
& \kappa_i \begin{cases} = \dfrac{biomass(\vec{v}^0)}{max(|v_{max}(i)|, v_{thresh})}, & \forall i \in I; \\ = 0, \quad \text{otherwise}, \end{cases}
\end{aligned}
$$

where $\vec{\alpha}$ and $\vec{\beta}$ can be considered as slack variables and $\vec{\kappa}_i$ is a coefficient vector controlling which reactions are allowed to exceed the upper/lower bounds and the penalty for exceeding the bounds.

We have also implemented ROOM, which is believed to better model mutant strains [8]. In the ROOM formulation, the objective is to minimize the number of reactions with significant changes

from the wild-type fluxes $\vec{v}^0$. TRIMER solves the following optimization problem:

$$\begin{aligned}
\min_{y} \quad & \sum_i y_i \\
\text{s.t.} \quad & S\vec{v} = 0; \\
& lb_i \leq v_i \leq ub_i, \quad \forall i \in I; \\
& v_i - (ub_i - w_i)y_i \leq w_i, \quad \forall i; \\
& v_i - (lb_i - w_i)y_i \geq w_i, \quad \forall i; \\
& w_i = v_i^0 + \delta|v_i^0| + \epsilon, \quad \forall i,
\end{aligned}$$

where δ and ϵ are two hyperparameters used in the original ROOM formulation to define the allowed flux changes from the wild-type fluxes $\vec{v}^0$.

The corresponding lower and upper reaction flux bounds in these metabolic network models are modified based on the inferred conditional probabilities given transcriptional changes as described in the previous subsections.

Following `TIGER` [10], TRIMER builds a customized Matlab CMPI (Common Mathematical Programming Interface) for metabolic flux prediction based on the data structure detailed above. This CMPI defines a consistent structure for mathematical programming solvers, including `CPLEX` and `GLPK`.

2.4 TRIMER as a simulator

With all these components, TRIMER can also serve as a simulator to generate both gene expression and metabolic flux data. Specifically, given model organisms or specific pathways of interest, metabolic network model can be first extracted from the existing models in the `COBRA` toolbox. Based on GPR rules as well as available knowledge on gene-gene interactions, we can simulate a BN (of different size if needed by growing from a set of metabolism-regulating genes to the whole genome for example) by randomly sampled genes and edges between the selected gene pairs. Given a simulated BN structure, BN model parameters characterizing the corresponding condi-

tional probability tables can also be simulated. With the simulated BN connecting to the metabolic model through GPR rules, we can first sample the gene expression data based on the BN. At the same time, with the simulated conditional probabilities under different conditions, for example with TF knockouts, metabolic flux predictions can be computed by constructing the corresponding transcriptional constraints as described previously. When serving as a simulator, TRIMER directly simulates a BN instead of learning the BN based on the regulatory prior knowledge and gene expression data as shown in Figure 2.2. Through this simulation procedure, TRIMER can serve as a fair benchmark platform for validation and comparison of different transcriptional-metabolic prediction methods as we have showcased in the latter simulation experiments.

2.5 Datasets and Software Packages

TRIMER integrates several existing packages. For the BN learning and inference module, `bn-learn` [33] and `gRain` [41] are adopted for Bayesian network learning and inference respectively. For the metabolic flux prediction module, TRIMER supports `CPLEX` and `GLPK` as solvers for the three aforementioned FBA formulations. In addition, TRIMER is also compatible with the CMPI module in that `TIGER` package [10] to interface with the corresponding FBA solvers.

2.5.1 Metabolic model

In general, TRIMER can take any metabolic model in the `COBRA` format based on the organism under study.We focuse on the analyses with *E. coli* and *yeast* by TRIMER in the thesis. We have used the iAF1260 and iMM904 model for *E. coli* and *y*east from the `COBRA` toolbox [29] throughout all the current experiments as the lab experimental data are collected from *E. coli* wild-type strains and knockout mutants.

2.5.2 Microarray datasets

To infer the TF regulation network and determine the 'ON/OFF' gene states, quantile normalization is performed over the archived microarray data in EcoMAC [42, 43] and the data in [43] from `https://sourceforge.net/projects/gemini-data/`.

2.5.3 TF-gene interaction annotations

For *E. coli*, we have used the interaction set in EcoMAC [42]. These data comprise all archived interactions in RegulonDB v8.1 [37] that were experimentally validated to support the existence of regulatory interactions.For *yeast*, we used the interactions in YEASTRACT [44, 43] database.Interactions with genes not included in the microarry datasets are pruned out resulting in 3,704 and 31075 regulatory interactions in total for *E. coli* and *yeast* respectively. Serving as prior knowledge, those interaction pairs helped to learn the BN from microarray data and derive the TF-target list for metabolism regulation as detailed previously

2.5.4 GPR rules

In `COBRA` [29] model including iAF1260 and iMM904, GPR rules are provided for most of the metabolic reactions. TRIMER takes these GPR rules from `COBRA` model directly.

2.6 Experimental Data Collection

2.6.1 *E. coli* mutants and validation

Strains deleted for genes encoding transcription factors used in this study were obtained from the Keio collection; an *E. coli* mutant library [45]. All comparisons were made to BW25113, the parent strain of the collection. Mutants were validated with internal gene-specific primers by colony PCR.

2.6.2 Kovac's Assay for indole quantification

The amount of indole produced by each mutant of interest was quantified by Kovac's assay as described in [46]. Briefly, total indole concentrations were determined by growing strains at 37°C overnight in LB or M9 minimal media, data for growth in each media is provided in Appendix C, and normalized to an OD_{600} of 0.3 the following morning. 60 μl of Kovacs reagent (comprised of 150 ml isoamyl alcohol (IAA), 50 ml concentrated hydrochloric acid (HCl) and 10 g of para-dimethylaminobenzaldehyde (DMAB)) was added per 200 μl of normalized culture and incubated for 2 minutes. 10 μl were subsequently removed and added to 200 μl of an HCl-IAA solution, and

the absorbance measured at 540 nm. Indole concentrations were then calculated using an indole standard curve prepared in the same manner as described above.

3. RESULTS

In this chapter, we present the experimental results based on both simulated and experimental data to demonstrate the effectiveness and flexibility of TRIMER for metabolic flux prediction under different conditions.

3.1 Simulation of *E. coli* Transcription Regulatory Network

In this set of experiments, we first validate that the BN learning in TRIMER can capture the regulatory relationships, which thereafter leads to reliable metabolic flux predictions with TF knockouts, based on a simulated BN model as the ground truth.

We first describe the procedure to simulate the BN for *E. coli* TRN based on the TRIMER simulator, given the corresponding metabolic network model. We start with a simulated small-scale BN and then demonstrate the scalability and flexibility of TRIMER with multiple TF knockout results in a large-scale BN in this section. With the simulated ground-truth BN, the metabolic network model is adopted to simulate reaction fluxes with the constraints based on the simulated conditional probabilities.

3.1.1 Simulating integrated transcription and metabolic regulations with a small-scale BN

For the experiments with the small-scale BN, we simulated the interactions between key genes related to indole production. Besides 12 transcription factors studied in the aforementioned TF knockout experiments using the Keio library, 32 corresponding target genes were also taken as the backbone nodes for the small-scale BN. To be specific, we selected these target genes by computing Pearson correlation coefficients (PCC) between the 12 indole-related TFs and the remaining genes in the gene expression data from EcoMAC [42]. The resulting 32 target genes were selected as each of these had PCC> 0.65 with at least one of these 12 TFs.

We took these 12 TFs and 32 selected target genes as the backbone nodes (44 in total) to simulate the BN as the transcription regulatory network. When simulating edges in this small-scale BN, directed regulatory interactions of these nodes were initialized with the following restrictions:

1) only the nodes corresponding to the TFs can serve as parent nodes in the simulated regulatory interactions; 2) the maximum number of edges between one TF parent node and all the other TF nodes was restricted to be half of the total number of TFs, and the maximum number of edges between one TF parent node and all the other target gene nodes was restricted to be half of the total number of target genes (This restriction prevents from simulating a very dense BN); 3) the edges were randomly generated between every valid pair of nodes with the corresponding values of conditional probability table (CPD) for each node being initialized randomly according to the uniform distribution Unif$(0, 1)$. The gene expression data were first generated by sampling based on the simulated BN. Ten sample sets of 2000 binary gene expression profiles were drawn via the forward sampling procedure on the simulated BN. For each sample set of 2000 generated samples, five subsets of 100, 200, 400, 800, and 1600 samples were randomly selected to construct the corresponding training sets for performance evaluation. In this way, 50 datasets in total with sizes ranging from 100 to 1600 were obtained.

On the other hand, regulating targets for each TF were found by the dependency between pairs of nodes, which can be obtained from the simulated BN structure. Based on these interactions, we can infer the probabilities of the corresponding gene states for different TF knockouts as well as the wild-type from the simulated BN. Based on the inferred probabilities and the gene-reaction relationships, the flux constraints in FBA can be adjusted to predict corresponding reaction fluxes of TF knockout mutants and the wild-type. For both the simulating ground-truth BN and the inferred BNs based on the simulated expression data, the corresponding metabolic fluxes can be simulated based on this procedure for performance evaluation. Note that all of our simulation experiments are based on the *E. coli* iAF1260 metabolic network model for FBA.

3.1.2 Small-scale BN structure inference based on simulated gene expression data

Given the simulated gene expression data, the first task was to learn the BN structure that best fits the simulated expression data for performance evaluation of discovering the regulatory interaction between TFs and target genes. In our experiments, we used score-based structure learning methods for this task, where the quality of the learned BN structure was measured by the Bayesian

Table 3.1: False negatives/positives for learned BN structures by Tabu search

Training dataset size	100	200	400	800	1600
False positive (avg± std)	39.2± 3.7	23.6± 6.0	17.0± 3.2	11.8±2.9	10.4±2.5
False negative (avg± std)	55.8± 3.8	36.6± 5.4	24.4± 2.2	15.6± 2.5	13.8± 2.3

[1] The total number of edges in the ground-truth simulated Bayesian network is 137.

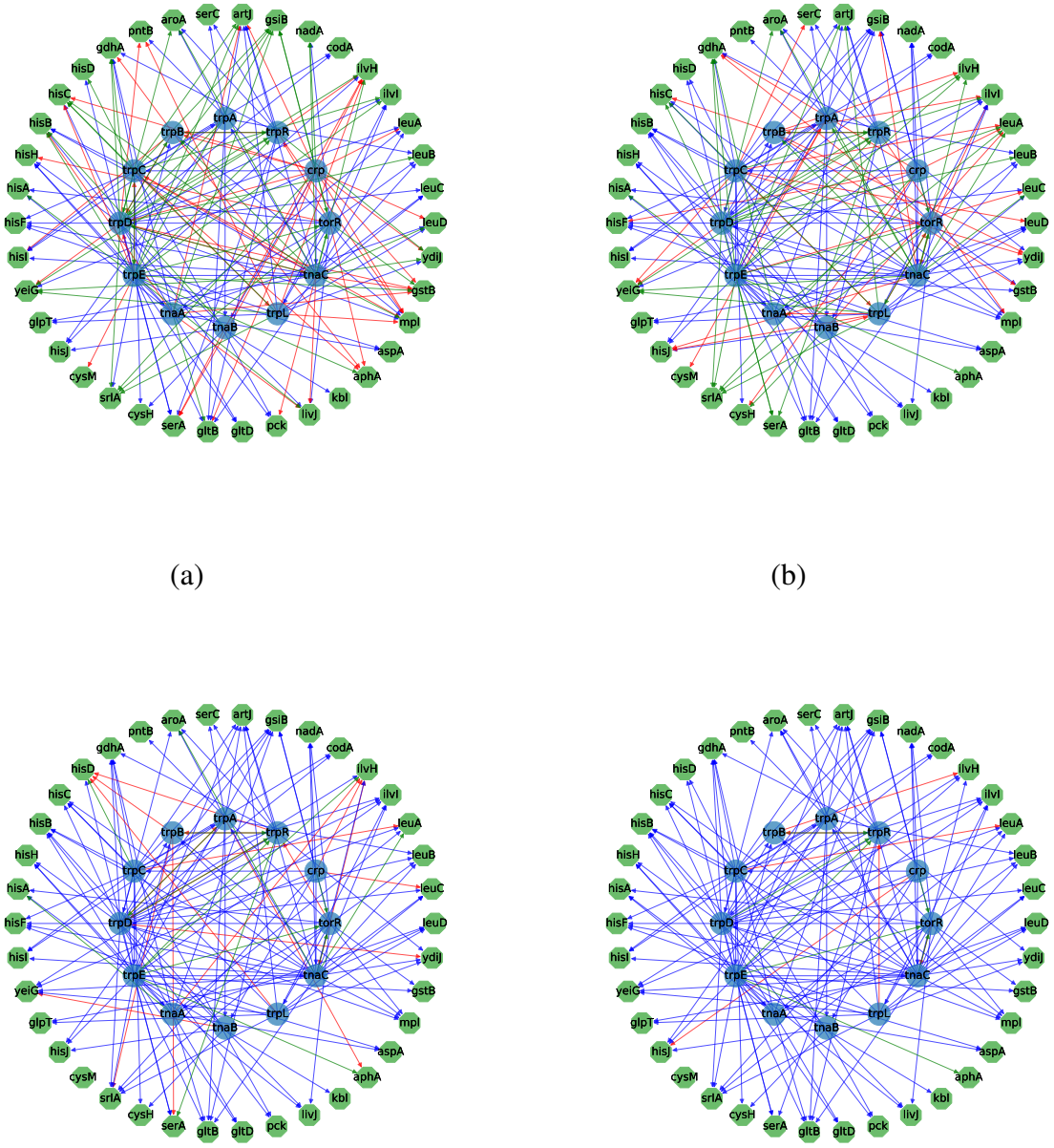
Table 3.2: False negatives/positives for learned tree-based BN structures

Training dataset size	100	200	400	800	1600
False positive (avg± std)	21.8± 3.4	16.6± 3.1	13.5± 2.1	10.1±1.0	9.4±1.2
False negative (avg± std)	107.8± 3.4	102.6± 3.1	99.5± 2.1	96.1± 1.0	95.4.± 1.2

[1] The total number of edges in the ground-truth simulated Bayesian network is 137.

Information Criterion (BIC) score. We tested two BN structures: Chow-Liu tree search algorithm for identifying the global optimal tree-based BN structure and Tabu search algorithm for more general BN structure learning. Tabu search only finds the local optimal structure. In order to guarantee the quality of the predicted solutions in our experiments, the Tabu length was set to be 100 where the best structural changes in every 100 iterations were iteratively updated as a reference for future search.

Once the BN structure was inferred based on the expression data,conditional probability tables for the BN were fit to the expression data by maximum likelihood estimates (MLEs). Finally, the regulated genes and associated conditional probabilities given TF states can be computed via examining dependency from the learned BN structure and performing the exact/approximate inference algorithm over the fitted BN. It should be noted that the original PROM estimates the conditional probabilities of gene states given TF states by MLE (relative frequencies) directly based on the expression data, while the authors stated that they only adjusted FBA constraints by investigating only the "experimentally verified" TF-gene pairs. In our experiments, the underlying dependency between pairs of nodes in the simulated ground-truth BN was considered as the actual TF-gene pairs for PROM, to some extent in favor of PROM since it did not learn the regulation network structure.



(a) (b)

(c) (d)

Figure 3.1: Examples of learned BNs from (a) 100, (b) 200, (c) 800, and (d) 1600 simulated expression profiles. The blue circled nodes represent 12 TFs while the green nodes are the corresponding target genes. The blue edges denote the accurately learned edges, the red edges are false positives where the regulatory edges were falsely added by BN structure learning, and the green edges are false negatives that BN learning was not able to identify.

3.1.3 Evaluation of flux prediction using TRIMER based on the small-scale inferred network

We further compare the flux prediction results by TRIMER-C with Chow-Liu tree (tree-TRIMER) and general BN structure (BN-TRIMER) to the results by PROM, based on simulated gene expression data. Note that we focused on applying the flux constraints based on (2.1) (TRIMER-C) for fair performance comparison with PROM. We computed the PCC between the simulated biomass and indole fluxes based on the simulated ground-truth network model and the predicted biomass and indole fluxes based on the inferred networks of both wild-type and the mutant strains deleted for TFs in the regulation network. For 10 simulated datasets of the same number of gene expression profiles, the average PCC and its standard deviation (std) were computed. Figures 3.2 and 3.3 summarize the performance comparison of TRIMER and PROM for biomass and indole flux prediction respectively, with the inferred BNs based on different numbers of simulated gene expression data.

As shown in Figure 3.2, from simulated expression data, BN-TRIMER consistently gives the closest biomass flux prediction to the simulated fluxes based on the ground-truth model. It is clear that with more expression data, the predicted fluxes can get better and vary less with different simulated expression data. With small training expression data, PROM's flux prediction can have quite weak correlation while with increasing number of expression profiles, the prediction can be improved. For tree-TRIMER, as the model class deviates from the ground-truth model, the prediction performance saturates when the number of training expression profiles is 400. On the other hand, with small training sets, tree-TRIMER still performs better than PROM.

Comparison for the indole flux predictions are also provided in Figure 3.3. Note that the ground-truth BN models were simulated based on the core subnetwork centering around indole-related reactions, identified by correlation analysis using EcoMAC gene expression data. We observe that both versions of TRIMER have better indole flux prediction performance, especially with small training data, compared to the results in Figure 3.2. The tree-TRIMER shows much better performance, which suggests that good prior knowledge on what to model for the TF reg-

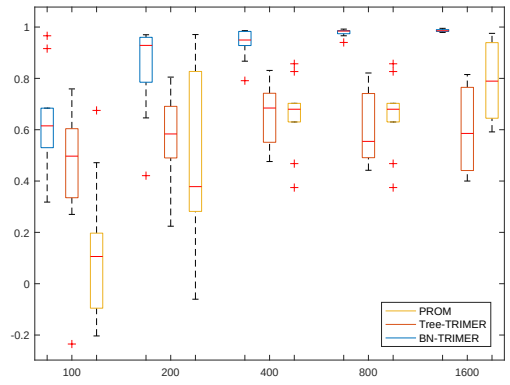


Figure 3.2: Biomass flux prediction comparison between TRIMER and PROM in the small-scale BN.

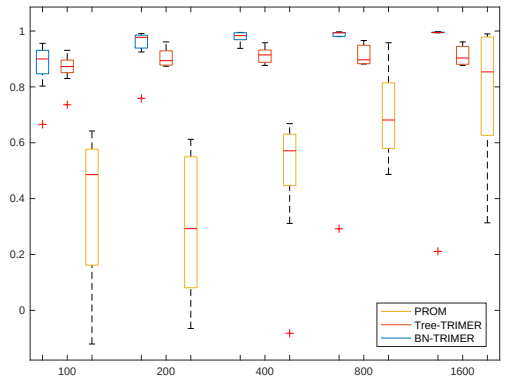


Figure 3.3: Indole flux prediction comparison between TRIMER and PROM in the small-scale BN.

ulation network may significantly enhance flux predictions. On the other hand, unlike TRIMER, PROM only models local dependency instead of global dependency, and its indole and biomass flux prediction performances are similar.

3.1.4 Simulating integrated transcription and metabolic regulations for a large-scale BN

We further simulate a large-scale BN with multiple TF knockouts to demonstrate the scalability and flexibility of the BN learning and metabolic flux prediction modules in TRIMER. To simulate a large-scale BN, we used all the genes included in the interaction list from EcoMAC and randomly selected 40% valid pairs of the interaction list as edges, resulting in a large-scale BN with 1591 genes and 1503 edges in this set of experiments.

One sample set of 2000 expression profiles was drawn via the forward sampling procedure as described for the small-scale BN. As done in the previous experiments, randomly selected interactions as edges in this simulated BN were taken as ground-truth interaction list and corresponding conditional probabilities associated with them were simulated. We used TRIMER-B to simulate the fluxes when we knocked out two TFs at the same time in this set of experiments.

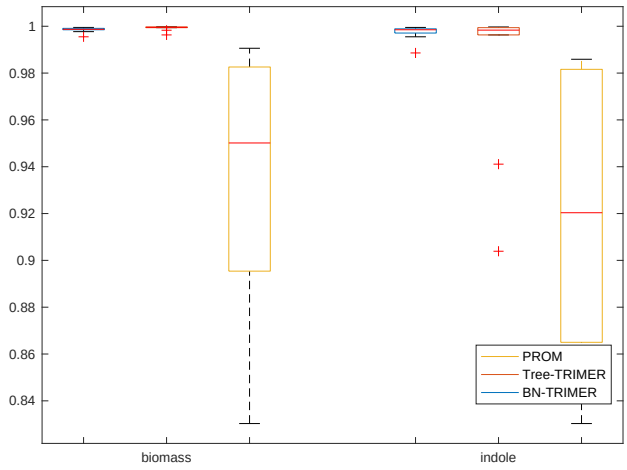


Figure 3.4: Flux prediction comparison between TRIMER and PROM for double TF knockouts in the simulated large-scale BN.

3.1.5 Evaluation of flux prediction using TRIMER based on the large-scale inferred network

We compared the flux prediction results by TRIMER-B with Chow-Liu tree (tree-TRIMER) and general BN structure (BN-TRIMER) with the results by PROM, based on PCC between the simulated and predicted fluxes for both biomass and indole production. Base on the sampled expression data from the simulated ground-truth BN model, a general-structure BN was inferred using our TRIMER package, resulting in 1377 edges, denoted as BN-TRIMER. When we restricted the BN to a Chow-Liu tree, the inferred BN had 1590 edges, denoted as tree-TRIMER. We used the simulated ground-truth interaction list for both TRIMER and PROM to construct the corresponding transcriptional constraints for flux predictions. To demonstrate the capability of TRIMER modeling mutant strains with multiple knockout TFs at the same time, ten random sets of 50 TF pairs were selected according to the EcoMAC interaction list. Figure 3.4 shows the corresponding bar plots, from which it is clear that based on the simulated expression data, BN-TRIMER consistently gave the best biomass and indole flux prediction with respect to the simulated fluxes based on the ground-truth model.

3.2 Experimental validation of metabolic flux predictions made by TRIMER

To further demonstrate the utility of TRIMER in *in silico* metabolic flux prediction for TF knockout mutants, we compared the prediction performance of TRIMER with PROM, IDREAM [22], and TR-FBA [19] for both biomass and indole flux prediction for *E. coli* TF-knockout mutants. We inferred the corresponding models based on the archived microarray gene expression data and the experimentally verified TF-gene interactions in EcoMAC [42].

For PROM, we used all 3704 interactions in EcoMAC. A key parameter for PROM is the binarization threshold value. In our experiments, it was determined by searching for the value when PROM achieved the best performance from 0.01 to 0.9 with the step-size of 0.01 based on the normalized microarray expression values.

IDREAM is an improved version of PROM as mentioned in the previous chapter, whose performance relies heavily on the inferred interactions by EGRIN [30]. As neither IDREAM [22] nor EGRIN [30] provided the inferred models for *E. coli*, we tried to derive the corresponding models using the data in EcoMAC .In our implementation for IDREAM, we tried to use the source code provided in the original IDREAM paper [22]. However, as only the part of the code using Inferelator was provided, for *E. coli* in our experiments, we had to take the available cMonkey package at `https://github.com/baliga-lab/cmonkey2` to derive the clusters in *E. coli* required for EGRIN. We used the default settings in cMonkey to first derive 250 clusters of genes in EcoMAC. Inferelator with default settings in IDREAM is performed to identify the regulators for each cluster. In our experiments, we identified 14175 interactions with 391 interactions overlapping with EcoMAC interactions. Note that in the reported results for yeast in the original IDREAM paper, 371 common interactions were identified. Finally, bootstrapping for conditional probability computation by Inferelator was done by training 200 models on randomly selected subsets of the EcoMAC gene expression data, where each subset accounts for 10% of the full expression data from EcoMAC.

In TRFBA, a constant parameter C is used to convert the expression levels to the upper bounds of the metabolic reactions. In our experiments, it was set to be the optimal value when TRFBA

achieved the best performance. For TRFBA, we also used all the interactions in EcoMAC directly.

In TRIMER, the binarized expression data are taken to infer the corresponding Bayesian network via Tabu search for modeling the TF regulation network using the general BN inference module of TRIMER. For fair comparison, we used the same binarization threshold value as PROM. With the search space restricted to 3704 EcoMAC-archived interactions, Tabu search was ran for one time based on all the expression data as we observed no significant change between learned BNs with or without bootstrapping. A BN with 1409 edges is learned from the EcoMAC expression data. Based on the inferred BN, the conditional probabilities of corresponding gene states when given TF knockouts were computed. Taking these inferred probabilities, the metabolic network flux prediction module with different implementations in TRIMER were adopted to predict biomass and indole fluxes for the corresponding TF knockout mutants.

3.2.1 Run-time

We ran our experiments on a PC with Intel Xeon 6248R processor. It should be noted that the BN structure learning part of TRIMER can be completed within 10 minutes with the search space limited to the interaction list. To predict corresponding biomass or indole fluxes for each TF-knockout mutant, it took TRIMER 7.32 seconds on average on the same PC. By comparison, the construction of the IDREAM model is time-consuming: Running cMonkey one time takes typically 5 hours; Running Inferelator for 200 times takes around 4 days. These computational challenges indeed limits the application of IDREAM to large-scale network modeling, especially considering knockout mutants.

3.2.2 Biomass prediction

We first compared the *in silico* flux predictions by TRIMER, PROM, IDREAM and TRFBA with the experimentally measured biomass productions in [1]. To compare the two ways of estimating conditional probabilities, we implemented both TRIMER models: we refer the TRIMER model with the conditional probabilities computed in the first or second way as TRIMER-C or TRIMER-B. For the biomass objective, we took `Ec_biomass_iAF1260_core_59p81M` in iAF1260

Table 3.3: Predicted biomass flux comparison for the knockout experiments in [1]. The unit of fluxes is mmol/gDCW/hr.

TF KO	Actual	TRIMER-C			TRIMER-B			PROM			MET	TRFBA	IDREAM
		FBA	sFBA	ROOM	FBA	sFBA	ROOM	FBA	sFBA	ROOM	sFBA	FBA	sFBA
WT +O2	0.710	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.708	0.563	0.708
arcA+O2	0.686	0.123	0.631	0.122	0.378	0.610	0.356	0.197	0.272	0.053	0.708	0.563	0.708
fnr +O2	0.635	0.391	0.538	0.388	0.381	0.547	0.381	0.399	0.526	0.356	0.708	0.563	0.708
arcA fnr +O2	0.648	0.127	0.395	0.055	0.315	0.619	0.298	0.197	0.272	0.015	0.708	0.563	0.708
appY +O2	0.636	0.708	0.708	0.671	0.708	0.708	0.671	0.708	0.708	0.671	0.708	0.563	0.708
oxyR +O2	0.637	0.708	0.708	0.671	0.708	0.708	0.671	0.708	0.708	0.671	0.708	0.563	0.708
soxS +O2	0.724	0.653	0.707	0.652	0.650	0.707	0.650	0.649	0.707	0.649	0.708	0.563	0.708
WT -O2	0.485	0.481	0.481	0.481	0.481	0.481	0.481	0.481	0.481	0.481	0.481	0.407	0.481
arcA -O2	0.377	0.023	0.023	0.022	0.071	0.071	0.062	0.037	0.037	0.034	0.481	0.355	0.481
fnr -O2	0.410	0.266	0.366	0.266	0.259	0.371	0.259	0.139	0.271	0.139	0.481	0.353	0.481
arcA fnr -O2	0.301	0.024	0.024	0.019	0.160	0.160	0.154	0.037	0.037	0.023	0.481	0.356	0.481
appY -O2	0.476	0.481	0.481	0.456	0.481	0.481	0.456	0.481	0.481	0.456	0.481	0.354	0.481
oxyR -O2	0.481	0.481	0.481	0.456	0.481	0.481	0.456	0.481	0.481	0.456	0.481	0.357	0.481
soxS -O2	0.465	0.443	0.481	0.443	0.442	0.481	0.442	0.441	0.479	0.441	0.481	0.355	0.481
PCC	-	0.538	0.870	0.517	0.700	0.906	0.702	0.619	0.693	0.484	0.918	0.927	0.918
rPCC	-	0.684	0.851	0.679	0.723	0.841	0.732	0.725	0.770	0.653	0.282	0.425	0.282

[1] The unit of fluxes is mmol/gDCW/hr

[2] In the FBA formulations, substrate (glucose) and oxygen uptake rates for aerobic conditions are set to be 8.5 and 14.6 mmol/gDCW/hr, respectively. They are set to 20.8 and 0 mmol/gDCW/hr for anaerobic conditions.

[3] The optimization is by the CPLEX solver.

as done in PROM. Three FBA formulations, standard FBA, sFBA, and ROOM for TRIMER-C, TRIMER-B, and PROM were implemented, where PROM-sFBA is the original PROM model. In the original TRFBA implementation, the metabolic network flux prediction formulations are based on FBA. In our experiments, the parameters δ and ϵ in the ROOM formulation were set to be 0.05 and 0.001. Binarization threshold for PROM, TRIMER and IDREAM was set to 0.33, which is used in the original PROM paper. The C value in TRFBA is set to 2.30 by searching from 0 to 3 with the step-size of 0.05.

Table 3.3 provides the comparison of the experimental and predicted fluxes by TRIMER-C, TRIMER-B, PROM, IDREAM and TRFBA for different TF knockout mutants. It should be pointed out that we used a fixed uptake rate for glucose and oxygen. In this way, simply using the metabolic model has no predictive capability for TF knockout mutants without integrating the change to the reaction regulations due to knockouts and the flux prediction will be the same as that of the wild-type. To illustrate how these hybrid models considering regulations improve over the simple metabolic network model, we have also included the results by simply running FBA,

denoted as MET in the table. This makes it more straightforward to compare how these integrated regulatory-metabolic model improve the predictive capability of metabolic-only model denoted as MET. Pearson correlation coefficients (PCC) were computed for performance comparison based on flux predictions for both wild-type and knockout strains. As experimental fluxes were measured under two growth conditions, we also computed PCC between experimental flux ratio and predicted flux ratio, where the flux ratio was obtained by dividing fluxes by the wild-type experimental flux in the corresponding growth condition. The PCC computed with the flux ratios is denoted as rPCC, which can better illustrate how knockout fluxes deviate from wild-type fluxes.

As shown in Table 3.3, TRIMER-B consistently achieved the highest PCC with the experimentally measured fluxes when compared to PROM and TRIMER-C under three FBA formuations. With sFBA, both TRIMER-C and TRIMER-B performed better than PROM. This shows the superiority of TRIMER over PROM. We can ascribe the overall superiority to the effective modeling of the global dependency in TF regulations through the BN learning and inference in TRIMER, in contrast to using simple conditional probability estimates adopted in PROM.

It is notable running FBA only without integrating transcription regulations always output wild-type flux predictions, which gave a high PCC, which did not provide meaningful evaluation. However, when normalized by the corresponding growth conditions, it led to a much lower rPCC. This explains why TRFBA achieved the highest PCC but the corresponding rPCC is only 0.425 since most of its knockout flux predictions were the same as its wild-type ones, which is clearly not desirable. By contrast, PCC and rPCC for TRIMER and PROM are consistent and TRIMER-sFBA achieved the highest rPCC.

Compared to the other methods, IDREAM basically predicted wild-type fluxes as MET for knockout mutants when we derive interactions based on EcoMAC data using EGRIN. We found that different numbers of interactions associated with araA, fnr, appY, osyR, and soxS (three, seven, zero, two and one, respectively) among 391 identified common interactions by EGRIN. Actually, with only these 13 interactions, no reaction in the metabolic model iAF1260 was significantly affected by TF knockouts, leading to the reported results.

Table 3.4: Predicted indole flux comparison for our TF knockout (KO) experiments in M9 minimal media. The unit of fluxes is mmol/gDCW/hr.

TF KO	Actual	TRIMER-C			TRIMER-B			PROM			TRFBA	IDREAM
		FBA	sFBA	ROOM	FBA	sFBA	ROOM	FBA	sFBA	ROOM	FBA	sFBA
fnr	0.0427	0.0231	0.0293	0.0427	0.0216	0.0301	0.0427	0.0224	0.0295	0.0427	0.0100	0.0397
soxS	0.0387	0.0366	0.0397	0.0386	0.0364	0.0397	0.0374	0.0365	0.0397	0.0367	0.0100	0.0397
crp	0.0397	0.0197	0.0197	0.0383	0.0193	0.0200	0.0367	0	0	0.0367	0.0100	0.0397
lysR	0.0400	0.0372	0.0372	0.0392	0.0370	0.0370	0.0380	0.0370	0.0370	0.0370	0.0100	0.0397
fucR	0.0390	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
malI	0.0403	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
phoB	0.0390	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
cpxR	0.0393	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
creB	0.0383	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
trpB	0	0	0	0	0	0	0	0	0	0	0	0.0397
trpD	0	0	0	0	0	0	0	0	0	0	0	0.0397
trpE	0	0	0	0	0	0	0	0	0	0	0	0.0397
paaX	0.0393	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
trpA	0	0	0	0	0	0	0	0	0	0	0	0.0397
tnaA	0.0380	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
trpL	0.0393	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
tnaC	0.0397	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
tnaB	0.0400	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
dhaR	0.0403	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0397	0.0397	0.0377	0.0100	0.0397
PCC	-	0.9270	0.9448	0.9988(7)	0.9203	0.9478	0.9988(8)	0.8305	0.8481	0.9987	0.9983	0

[1] The unit of fluxes is mmol/gDCW/hr
[2] In the FBA formulations, substrate (glucose) and oxygen uptake rates are set to be 9.5 mmol/gDCW/hr and 13.0 mmol/gDCW/hr, respectively.
[3] The optimization is by the CPLEX solver.

3.2.3 Indole flux prediction

We further validated the predicted fluxes by TRIMER with our experimentally-generated data from TF-knockout experiments for indole production as described previously. As we generated all these experimental data under the same growth condition, we only report PCC for performance comparison in this set of experiments. We used the same parameters as in the previous experiment for all the models and took `TRPS3` in iAF1260 for indole flux prediction. Table 3.4 provides the comparison of the experimental and predicted fluxes by TRIMER-C, TRIMER-B, PROM, IDREAM and TRFBA for different TF knockout mutants grown in M9 minimal media and the overall PCCs between experimental and predicted fluxes. In this set of experiments,IDREAM had the same issue as for biomass prediction.By contrast TRIMER-B achieved consistently better correlation with the experimental results with all three formulations compared to TRIMER-C and PROM. It should be also noted that TRIMER with the ROOM formulation has achieved the high-

Table 3.5: Predicted biomass flux comparison by correlation analysis for the knockout experiments in [2].

	TRIMER-C			TRIMER-B			PROM			TRFBA	IDREAM
	FBA	sFBA	ROOM	FBA	sFBA	ROOM	FBA	sFBA	ROOM	FBA	sFBA
PCC	0.479	0.499	0.482	0.492	0.503	0.497	0.327	0.350	0.340	0.161	0.388
rPCC	0.479	0.499	0.482	0.492	0.503	0.496	0.326	0.349	0.339	0.163	0.387

[1] In the FBA formulations, substrate (glucose) and oxygen uptake rates are set to be 10.0 and 3.3 mmol/gDCW/hr, respectively.
[2] The optimization is by the CPLEX solver.

est correlation values, which were significantly better than the other FBA formulations for both TRIMER and PROM. TRFBA also achieved similar PCC as TRIMER with the ROOM formulation.

Based on these results, the construction of EGRIN-derived regulatory network in IDREAM may need to be carefully tuned for different organisms, data and analytic tasks (We again note that the presented results in the original IDREAM paper were for yeast). Based on our experience, fine-tuning the IDREAM model can be time-consuming compared to PROM and TRIMER.

In summary, the predictions by TRIMER and PROM are more reliable and TRIMER performs the best for capturing the varying patterns of knockout fluxes. The experimental results also suggest that the existing hybrid models, in particular IDREAM, require careful tuning to be applicable to different organisms, available data, as well as prediction tasks. To further illustrate this and demonstrate the flexibility of our TRIMER pipeline, we performed additional experiments for metabolic flux prediction for *yeast* with different knockouts, based on the reported experiments in the original IDREAM paper [22].

3.3 Performance comparison for yeast metabolic flux predictions

To further investigate the performance of TRIMER compared to PROM, IDREAM, and TRFBA, we tested all the methods based on the experiments for yeast. We have taken the yeast metabolic model iMM904 for metabolic flux prediction. YEASTRACT interactions with genes not included in the metabolic model iMM904 is pruned out resulting in 31075 interactions in total. As done in the previous experiments, these interactions were used for BN structure learning. As gene ex-

pression data were not provided in [22], we had to use the gene expression data in [43] from `https://sourceforge.net/projects/gemini-data/` in these competing methods accordingly. As done in the previous experiments, the search space of BN structure learning for TRIMER on yeast is restricted to 31075 interactions from YEASTRACT [43]. A BN with 1809 edges were learned. It should be pointed out that the structure learning process took about 40 minutes while the search space for yeast is much larger than that in our *E.coli* experiments. For IDREAM, we directly used the provided EGRIN-derived model in [22] directly.

We compared the *in silico* flux predictions by TRIMER-C, TRIMER-B, PROM, IDREAM, and TRFBA with the experimental biomass measurements for 119 TF knockouts provided in [2]. For the biomass objective function, we took `Biomass_SC5_notrace` in iMM904 as done in [43]. The parameters δ and ϵ in the ROOM formulation were set to be 0.05 and 0.001. For PROM, IDREAM, and TRIMER, we still used the binarization threshold of 0.33 as suggested in [43]. The optimal value for C parameter in TRFBA is 2.5 by searching from 0 to 3 with the step-size 0.05.

It should be noted that the performance of PROM, IDREAM, and TRIMER mainly depends on two aspects: selection of interaction pairs and computing of corresponding conditional probabilities in the respective regulatory models. IDREAM extended PROM to incorporate a refined interaction list by taking advantage of EGRIN while one of the main contributions in TRIMER that makes it different from PROM and IDREAM is in modeling conditional probabilities with BN. Here we compared the flux prediction performance based on the EGRIN-derived interaction list as originally reported in [22], which includes 307 interactions. Hence, the experimental results provided in Table 3.5 demonstrates how conditional probability modeling with BN proposed in TRIMER may further improve flux predictions of TF knockouts. From Table 3.5, we can observe that TRFBA performed the worst in this set of experiments. As stated in [19], this may be due to the lack of corresponding gene expression profiles measured with knockout mutants we considered, which is not required by PROM, IDREAM, or our TRIMER. Comparing TRIMER with PROM and IDREAM based on the same set of EGRIN-derived interactions, we can clearly see that TRIMER can more faithfully predict biomass production compared to PROM and IDREAM,

again due to better modeling global dependency among the interacting genes.

Based on the presented performance comparison results with both *E. coli* and yeast models and data, we have clearly shown that BN modeling transcription regulations in TRIMER can capture regulation relationships better and improve metabolic flux predictions for knockout mutants compared to the relative frequency based conditional probability estimation in PROM and IDREAM. What's more, TRIMER does not require significant tuning and is more user-friendly when being implemented for different model organisms, prediction tasks, and/or expression data. On the contrary, many existing hybrid models, including IDREAM and TRFBA, often require careful tuning when being applied to different models, tasks and data. Sometimes, it can be challenging, time-consuming, and requiring both biology and modeling expertise. Removing one of important bottlenecks in applying developed computational systems biology packages to solve real-world problems is to develop more flexible and user-friendly frameworks and tools.

4. CONCLUSIONS

Based on the performance comparison results with the presented experiments, we have shown that BN modeling transcription regulations in TRIMER can capture regulation relationships better and improve metabolic flux predictions for knockout mutants compared to the relative frequency based conditional probability estimation in PROM and its extensions. What's more, TRIMER does not require significant tuning and is more user-friendly when being implemented for different model organisms, prediction tasks, and/or expression data. By contrast, many existing hybrid models often require careful tuning when being applied to different models, tasks and data, which can be challenging, time-consuming, and requiring both biology and modeling expertise. One of important challenges in applying developed computational systems biology packages to solve real-world problems is to develop more flexible and user-friendly frameworks and tools, for which we have tried to consider when developing the TRIMER package.

Due to the inherent stochasticity and complexity of living systems, accurate inference of the transcription regulatory network model as well as the metabolic network model is practically challenging, especially when studying non-model organisms other than *E. coli* or *yeast* studied in this thesis. In order to better capture the potential model uncertainty and be able to make reliable predictions in the presence of substantial uncertainty, we may have to deal with an uncertainty class of network models rather than a single best model that are consistent with the available data and knowledge. This also enables closed-loop experimental design, where new experiments may be designed to reduce model uncertainty, the outcomes of the designed experiments may be used to update the uncertainty class, and where this experimental loop may be repeated. We leave this for our future research.

REFERENCES

[1] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. Ø. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature*, vol. 429, p. 9296, 2004.

[2] S.-M. Fendt, A. P. Oliveira, S. Christen, P. Picotti, R. C. Dechant, and U. Sauer, "Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast," *Molecular Systems Biology*, vol. 6, no. 1, p. 432, 2010.

[3] A. Varma and B. Ø. Palsson, "Metabolic flux balancing: Basic concepts, scientific and practical use," *Bio/technology*, vol. 12, no. 10, pp. 994–998, 1994.

[4] J. S. Edwards and B. Ø. Palsson, "The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences*, vol. 97, no. 10, pp. 5528–5533, 2000.

[5] C. L. Barrett, T. Y. Kim, H. U. Kim, B. Ø. Palsson, and S. Y. Lee, "Systems biology as a foundation for genome-scale synthetic biology," *Current Opinion in Biotechnology*, vol. 17, no. 5, pp. 488–492, 2006.

[6] D. Segre, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 15112–15117, 2002.

[7] A. P. Burgard, P. Pharkya, and C. D. Maranas, "Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization," *Biotechnology and Bioengineering*, vol. 84, no. 6, pp. 647–657, 2003.

[8] T. Shlomi, O. Berkman, and E. Ruppin, "Regulatory on/off minimization of metabolic flux changes after genetic perturbations," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7695–7700, 2005.

[9] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. Ø. Palsson, "Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models," *Molecular Systems Biology*, vol. 6, no. 1, p. 390, 2010.

[10] P. Jensen, K. Lutz, and J. Papin, "TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks," *BMC System Biology*, vol. 5, no. 147, 2011.

[11] S. Ren, B. Zeng, and X. Qian, "Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints," *BMC Bioinformatics*, vol. 14, no. S2, p. S17, 2013.

[12] B. Palsson, *Systems Biology*. Cambridge University Press, 2015.

[13] M. Apaydin, L. Xu, B. Zeng, and X. Qian, "Robust mutant strain design by pessimistic optimization," *BMC Genomics*, vol. 18, no. 6, p. 677, 2017.

[14] M. W. Covert and B. O. Palsson, "Constraints-based models: regulation of gene expression reduces the steady-state solution space," *Journal of Theoretical Biology*, vol. 221, pp. 309–325, April 2003.

[15] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, "A genome-scale computational study of the interplay between transcriptional regulation and metabolism," *Molecular Systems Biology*, vol. 3, no. 1, p. 101, 2007.

[16] M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr, "Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*," *Bioinformatics*, vol. 24, pp. 2044–2050, 07 2008.

[17] D. Machado and M. Herrgård, "Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism," *PLOS Computational Biology*, vol. 10, no. 4, p. e1003580, 2014.

[18] J. L. Reed, "Genome-scale metabolic modeling and its application to microbial communities," in *The Chemistry of Microbiomes: Proceedings of a Seminar Series*, National Academies Press, 2017.

[19] E. Motamedian, M. Mohammadi, S. A. Shojaosadati, and M. Heydari, "TRFBA: An algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data," *Bioinformatics*, vol. 33, no. 7, pp. 1057–1063, 2017.

[20] H. Yu and R. H. Blair, "Integration of probabilistic regulatory networks into constraint-based models of metabolism with applications to Alzheimer's disease," *BMC Bioinformatics*, vol. 20, no. 386, 2019.

[21] S. Chandrasekaran and N. D. Price, "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences*, vol. 107, no. 41, pp. 17845–17850, 2010.

[22] Z. Wang, S. A. Danziger, B. D. Heavner, S. Ma, J. J. Smith, S. Li, T. Herricks, E. Simeonidis, N. S. Baliga, J. D. Aitchison, and N. D. Price, "Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast," *PLOS Computational Biology*, vol. 13, pp. 1–23, 05 2017.

[23] F. Shen, R. Sun, J. Yao, J. Li, Q. Liu, N. D. Price, C. Liu, and Z. Wang, "OptRAM: *In-silico* strain design via integrative regulatory-metabolic network modeling," *PLOS Computational Biology*, vol. 15, pp. 1–25, 03 2019.

[24] A. Varma and B. Ø. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110," *Applied and Environmental Microbiology*, vol. 60, no. 10, pp. 3724–3731, 1994.

[25] D. Koller and N. Friedman, *Probabilistic graphical models: Principles and techniques*. MIT press, 2009.

[26] R. Mahadevan and C. H. Schilling, "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models," *Metabolic Engineering*, vol. 5, no. 4, pp. 264–276, 2003.

[27] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Ø. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nature Biotechnology*, vol. 26, no. 9, pp. 1546–1696, 2008.

[28] A. Arkin, R. Cottingham, C. Henry, N. Harris, R. Stevens, S. Maslov, *et al.*, "KBase: The United States Department of Energy Systems Biology Knowledgebase," *Nature Biotechnology*, vol. 36, p. 566, 2018.

[29] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, S. Magnusdóttir, C. Y. Ng, G. Preciat, A. agare, S. H. J. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, H. M. Le, D. Ma, Y. Sun, L. Wang, J. T. Yurkovich, M. A. P. Oliveira, P. T. Vuong, L. P. El Assal, I. Kuperstein, A. Zinovyev, H. S. Hinton, W. A. Bryant, F. J. Aragón Artacho, F. J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M. A. Saunders, C. D. Maranas, N. E. Lewis, T. Sauter, B. Ø. Palsson, I. Thiele, and R. M. T. Fleming, "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0," *Nature Biotechnology*, vol. 14, no. 3, pp. 1750–2799, 2019.

[30] A. N. Brooks, D. J. Reiss, A. Allard, W.-J. Wu, D. M. Salvanha, C. L. Plaisier, S. Chandrasekaran, M. Pan, A. Kaur, and N. S. Baliga, "A system-level model for the microbial regulatory genome," *Molecular Systems Biology*, vol. 10, no. 7, p. 740, 2014.

[31] D. J. Reiss, C. L. Plaisier, W.-J. Wu, and N. S. Baliga, "cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism," *Nucleic Acids Research*, vol. 43, no. 13, pp. e87–e87, 2015.

[32] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*," *Genome Biology*, vol. 7, no. 5, pp. 1–16, 2006.

[33] R. Nagarajan, M. Scutari, and S. Lèbre, "Bayesian Networks in R with Applications in Systems Biology 2013," *New York, NYSpringer-Verlag*.

[34] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[35] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2002.

[36] M. Scutari and R. Nagarajan, "Identifying significant edges in graphical models of molecular networks," *Artificial Intelligence in Medicine*, vol. 57, no. 3, pp. 207–217, 2013.

[37] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeida, J. S. García-Sotelo, K. Alquicira-Hernández, L. J. Muñiz Rascado, P. Peña Loredo, C. Ishida-Gutiérrez, D. A. Velázquez-Ramírez, V. D. Moral-Chávez, C. Bonavides-Martínez, C.-F. Méndez-Cruz, J. Galagan, and J. Collado-Vides, "RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli K-12*," *Nucleic Acids Research*, vol. 47, no. D1, pp. D212–D220, 2019.

[38] A. Irrthum, L. Wehenkel, P. Geurts, *et al.*, "Inferring regulatory networks from expression data using tree-based methods," *PloS ONE*, vol. 5, no. 9, p. e12776, 2010.

[39] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "TIGRESS: Trustful inference of gene regulation using stability selection," *BMC Systems Biology*, vol. 6, no. 1, pp. 1–17, 2012.

[40] I. T. Young, "Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources," *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 935–941, 1977.

[41] S. Højsgaard *et al.*, "Graphical independence networks with the gRain package for R," *Journal of Statistical Software*, vol. 46, no. 10, pp. 1–26, 2012.

[42] J. Carrera, R. Estrela, J. Luo, N. Rai, A. Tsoukalas, and I. Tagkopoulos, "An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*," *Molecular System Biology*, vol. 10, no. 7, p. 735, 2014.

[43] S. Chandrasekaran and N. D. Price, "Metabolic constraint-based refinement of transcriptional regulatory networks," *PLoS Computational Biology*, vol. 9, no. 12, p. e1003370, 2013.

[44] D. Abdulrehman, P. T. Monteiro, M. C. Teixeira, N. P. Mira, A. B. Lourenco, S. C. Dos Santos, T. R. Cabrito, A. P. Francisco, S. C. Madeira, R. S. Aires, *et al.*, "YEASTRACT: Providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D136–D140, 2010.

[45] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. Datsenko, M. Tomita, B. Wanner, and H. Mori, "Construction of *Escherichia coli* K12 inframe, singlegene knockout mutants: the Keio collection," *Molecular Systems Biology*, vol. 2, p. 2006.0008, 2006.

[46] E. L. Chant and D. K. Summers, "Indole signalling contributes to the stable maintenance of *Escherichia coli* multicopy plasmids," *Molecular Microbiology*, vol. 63, no. 1, pp. 35–43, 2007.

APPENDIX A

EXAMPLE OF CONSTRUCTING TRANSCRIPTIONAL CONSTRAINTS

A.1 Examples of inferring conditional probabilities given BN

We provide two examples for the two ways of applying transcriptional regulations based on BN-inferred conditional probabilities as explained in the main text.

We provide an example to illustrate the operation based on (2.1). In this example, the reaction r is catalyzed by two genes A and B according to the GPR rules in the metabolic model.When their regulating TF is knockout, we can obtain a probability vector: $[p(A = 1|TF = 0),\, p(B = 1|TF = 1)]^{\mathrm{T}}$. The corresponding reaction flux upper/lower bounds for reaction r are set to be:

$$\begin{aligned} ub_r &= v_{max}(r) \times \min\{p(A = 1|TF = 0), p(B = 1|TF = 0)\} \\ &\text{if} \quad v^0(r) > 0; \\ lb_r &= -v_{max}(r) \times \min\{p(A = 1|TF = 0), p(B = 1|TF = 0)\} \\ &\text{if} \quad v^0(r) < 0, \end{aligned}$$

where $v^0(r)$ is the wild-type flux for reaction r.

We now give the example to illustrate the operation based on (2.2). In this example, the reaction r is associated with a GPR rule, $(A\ and\ B) or C$. The corresponding GPR rule values and three gene states are illustrated in Table A.1. We can see that only four of the sixteen possible state combinations render that the GPR rule to be false. Thus, the upper or lower bounds with respect to r will be computed as:

Table A.1: GPR rules for gene state profiles of three genes: A, B, and C.

GPR value	0	1	0	1	0	1	1	1
A	0	0	0	0	1	1	1	1
B	0	0	1	1	0	0	1	1
C	0	1	0	1	0	1	0	1

$$
\begin{aligned}
ub_r =& v_{max}(r) \times (p(A=0,B=0,C=0|TF=0) \\
&+ p(A=0,B=1,C=0|TF=0) \\
&+ p(A=1,B=0,C=0|TF=0)) \quad \text{if } v^0(r) > 0; \\
lb_r =& - v_{max}(r) \times (p(A=0,B=0,C=0|TF=0) \\
&+ p(A=0,B=1,C=0|TF=0) \\
&+ p(A=1,B=0,C=0|TF=0)) \quad \text{if } v^0(r) < 0.
\end{aligned}
$$

A.2 Example of adding constraints for flux prediction

In TRIMER, metabolic flux prediction allows adding new constraints based on the specified data structures in the main text. If we want to add three additional vectors $\vec{a}$, $\vec{b}$ and $\vec{c} \in R^N$ with constraints, $Q\vec{a} - U\vec{b} = \vec{c}\,(U, Q \in R^{N \times N})$, to a data structure $trimer$, it can be simply achieved by the following command lines:

```
trimer = add_column(trimer,
[A;B;C],vartype', 'c');
trimer = add_constraint(trimer, linalg(Q,A, U,B, '=', C)).
```

APPENDIX B

ADDITIONAL FEATURES OF TRIMER

B.1 Refine TRIMER with given phenotypes

In TRIMER, we provide a way to refine the current metabolic model given a minimum growth rate. This can help to remove or adjust regulatory bounds that over-constrain the prediction model when TFs are knocked out. These bounds can be decided by solving the following optimization problem:

$$
\begin{aligned}
\min_{y} \quad & \sum_{i \in I} (y_i) \\
\text{s.t.} \quad & S\vec{v} = 0; \\
& \text{biomass}(\vec{v}) > v_{growth} \\
& lb_i \leq v_i \leq ub_i, \quad \forall i \in I \\
& v_i - (ub_i^{init} - ub_i) y_i \leq ub_i, \quad \forall i \\
& v_i - (lb_i^{init} - lb_i) y_i \geq lb_i, \quad \forall i,
\end{aligned}
$$

where ub^{init} and lb^{init} are the initial bounds from the original wild-type model in `COBRA` and v_{growth} is the minimum growth-rate requirement specified by the user. Suppose the threshold for a lethal KO is marked with 0.05 times the wild-type biomass flux. In the experiments of biomass prediction based on the experimental data in [1], we predict that the phenotype of arcA KO is lethal when we have sFBA with the TRIMER-C model. However, the actual phenotype of arcA KO is non-lethal according to the experimentally measured fluxes. This may indicate that some estimated conditional probabilities for constructing flux constraints are too small and some reactions affected by arcA KO are over-constrained. Via the optimization problem above, we can identify which reactions are over-constrained for TF(s) knockouts for given non-lethal phenotypes. Based on this, we can further adjust the values of conditional probabilities corresponding to these reactions to

Table B.1: Reactions that are over-constrained with the corresponding inferred and adjusted probabilities.

reaction index	probabilities	probabilities adjusted
ACONTa	0.0393	0.9
ACONTb	0.0393	0.9
CS	0.1534	0.9
biomass flux	0.0193	0.4027

make the predicted phenotypes to better match the experimentally observed phenotypes and also, the predicted fluxes to be closer to the experimentally observed fluxes. The name abbreviations of the reactions that are over-constrained and their corresponding condition probabilities for arcA KO is shown in Table B.1. It is clear that these probability values are all small and result in the predicted phenotype to be lethal. We adjust all these probabilities to be 0.9 and the predicted biomass flux becomes 0.4027, which is very close to the experimentally measured flux.

B.2 Integrating TRIMER with TIGER

In TRIMER, we have also programmed a simulation pipeline that can simulate knockout mutant metabolic fluxes in various growth conditions by borrowing the modules in `TIGER` [10], instead of only being capable simulating aerobic and anaerobic glucose minimal medium conditions as in PROM [21]. We adopt a Boolean model to simulate the feedback regulatory rules as implemented in `TIGER`. As TRIMER adopts the similar data structure as `TIGER`, which is known to be a platform to integrate `COBRA` models with these Boolean transcription regulations, TRIMER allows the user to build a hybrid model that integrates probabilistic TRN, Boolean feedback rules, and `COBRA` metabolic models into a single unified pipeline, making it possible to simulate knockout mutants in various growth conditions. We have simulated 125 growth conditions for 15 TF KOs based on the *E. coli* iAF1260 model and compared the performance of this hybrid model and that by PROM using the phenotype datasets, originally given in [21], as the ground truth. The parameter settings of growth conditions can be found in [1]. For the `TIGER` part of the hybrid model to model Boolean regulations interfacting the TRN and metabolic model, we have adopted the

iMC1010 Boolean network in [1]. The TRIMER part of the hybrid model is the same as PROM. Figure B.1 shows the results. The best performance of the hybrid model and PROM implentations are both achieved when the threshhold for lethal phenotypes is set to be 0.15 times the WT growth rate. As we can see, the predictions mainly differ in the growth condition with the growth media, 1,2-Propanediol L-Tartaric Acid, L-Tartaric Acid, and Guanine. With additional constraints introduced from the Boolean rules, many predicted phenotypes become lethal.

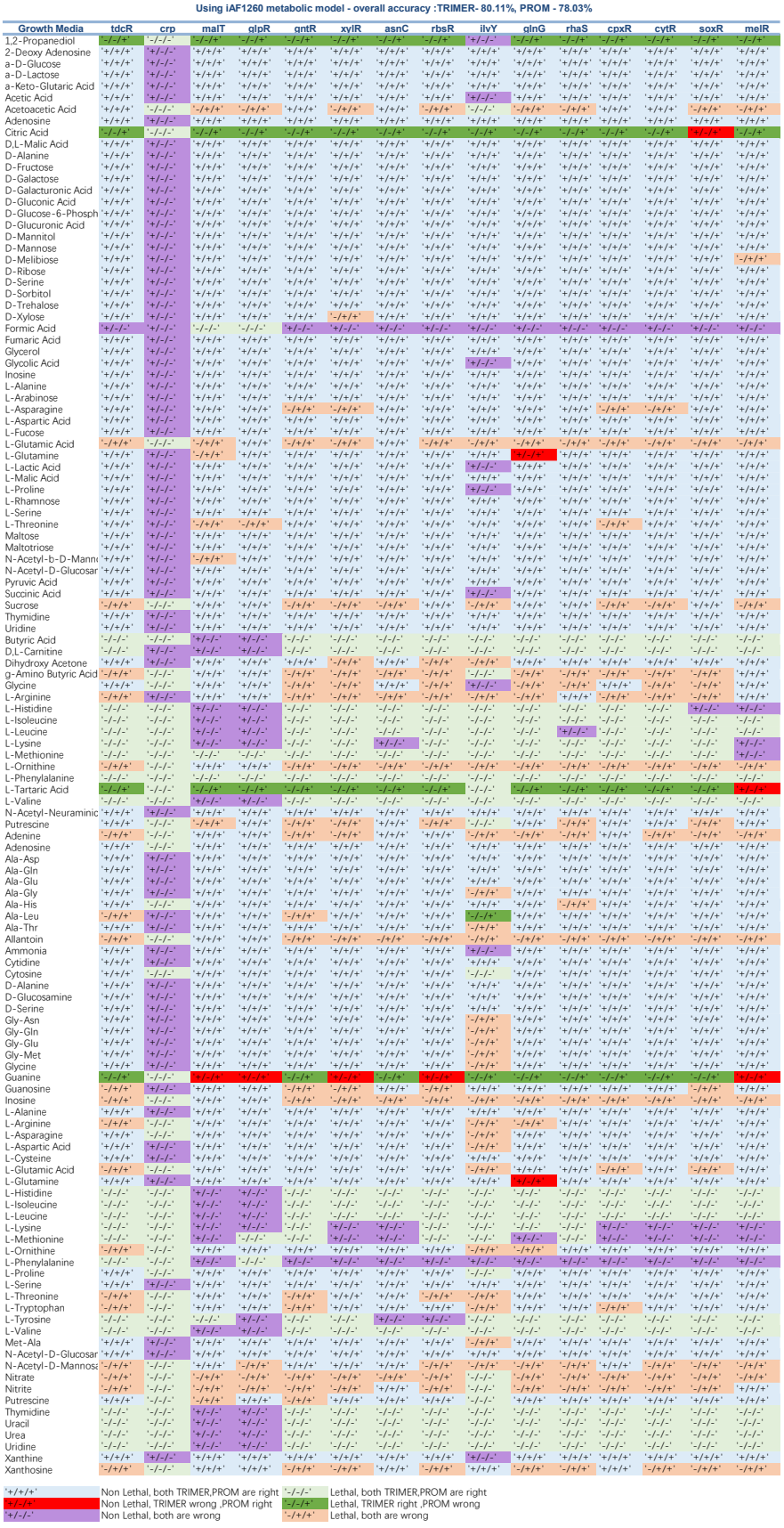


Figure B.1: Phenotype prediction comparison between TRIMER and PROM.

APPENDIX C

COLLECTED EXPERIMENTAL DATA

Table C.1: Total indole concentrations of E. coli transcription factor deletants in LB media

Strain	Mean absorbance1	STDEV	mmols of Indole
WT	0.102	0.006	0.1031
crp	0.043	0.005	0.0000
tnaA	0.045	0.006	0.0000
lldR	0.074	0.006	0.0492
glcC	0.076	0.009	0.0531
nadR	0.080	0.006	0.0595
relB	0.081	0.006	0.0614
fabR	0.081	0.003	0.0620
arsR	0.081	0.007	0.0627
acrR	0.082	0.003	0.0640
fhlA	0.085	0.003	0.0697
araC	0.085	0.005	0.0704
marR	0.086	0.003	0.0710
uxuR	0.087	0.013	0.0736
metR	0.087	0.009	0.0742
cytr	0.088	0.013	0.0761
dsdX	0.089	0.010	0.0768
allR	0.089	0.004	0.0781
exuR	0.090	0.007	0.0787
soxR	0.090	0.004	0.0793

gadX	0.091	0.012	0.0819
fnr	0.094	0.003	0.0870
mprA	0.095	0.010	0.0883
nanR	0.095	0.005	0.0896
stpA	0.096	0.009	0.0902
nhaR	0.096	0.013	0.0915
kdgR	0.097	0.005	0.0922
idnR	0.097	0.008	0.0922
melR	0.097	0.005	0.0928
ada	0.098	0.012	0.0941
metJ	0.098	0.015	0.0941
mlrA	0.098	0.007	0.0941
galS	0.098	0.019	0.0954
tyrR	0.099	0.010	0.0967
ilvY	0.099	0.017	0.0967
xapR	0.100	0.017	0.0979
zntR	0.100	0.012	0.0979
rstA	0.100	0.012	0.0979
nagC	0.100	0.006	0.0992
csiR	0.100	0.004	0.0992
hupA	0.101	0.010	0.1011
trpA	0.101	0.006	0.1011
leuO	0.101	0.008	0.1011
ebgR	0.102	0.018	0.1024

lacI	0.102	0.004	0.1031

soxS	0.103	0.017	0.1050
rtcR	0.104	0.007	0.1069
rbsR	0.105	0.011	0.1076
narL	0.105	0.014	0.1082
gntR	0.105	0.008	0.1082
chbR	0.106	0.021	0.1095
lysR	0.106	0.018	0.1108
glnG	0.106	0.005	0.1108
lrp	0.107	0.014	0.1127
cbl	0.107	0.011	0.1127
rhaS	0.107	0.010	0.1127
sgrR	0.108	0.010	0.1140
yehT	0.108	0.010	0.1140
envR	0.108	0.015	0.1146
glpR	0.109	0.001	0.1159
fadR	0.110	0.010	0.1172
xylR	0.110	0.013	0.1178
uidR	0.110	0.012	0.1184
torR	0.111	0.029	0.1191
oxyR	0.111	0.011	0.1197
rhaR	0.111	0.014	0.1204
rpiR	0.111	0.007	0.1204
appY	0.112	0.017	0.1210
creB	0.112	0.013	0.1210
hcaR	0.112	0.004	0.1210
slyA	0.112	0.003	0.1210

prpR	0.112	0.011	0.1217
feaR	0.112	0.010	0.1217
srlR	0.112	0.016	0.1223
dcuR	0.112	0.014	0.1223
argP	0.113	0.009	0.1242
caiF	0.113	0.028	0.1242
nac	0.114	0.020	0.1249
cynR	0.114	0.012	0.1249
betI	0.114	0.013	0.1249
aidB	0.114	0.008	0.1249
fis	0.114	0.016	0.1255
adiY	0.114	0.009	0.1255
trpC	0.114	0.010	0.1255
galR	0.114	0.003	0.1261
fliZ	0.114	0.009	0.1261
argR	0.115	0.014	0.1268
tdcA	0.115	0.015	0.1274
evgA	0.115	0.017	0.1274
gadE	0.115	0.009	0.1274
gadW	0.115	0.010	0.1274

gutM	0.116	0.002	0.1293
cdaR	0.116	0.011	0.1300
ycfQ	0.116	0.009	0.1300
arcA	0.117	0.007	0.1319
marA	0.117	0.007	0.1319

bglJ	0.117	0.004	0.1319
trpL	0.118	0.013	0.1326
treR	0.118	0.019	0.1332
phoP	0.118	0.021	0.1332
iscR	0.118	0.015	0.1338
paaX	0.118	0.011	0.1338
fur	0.118	0.008	0.1338
tnaC	0.120	0.015	0.1370
mngR	0.120	0.013	0.1370
cueR	0.121	0.003	0.1383
rob	0.121	0.016	0.1383
tnaB	0.121	0.012	0.1383
csgD	0.121	0.006	0.1390
asnC	0.121	0.011	0.1390
cadC	0.122	0.015	0.1402
mhpR	0.122	0.006	0.1402
yeiL	0.122	0.004	0.1402
cspA	0.122	0.007	0.1402
kdpE	0.122	0.006	0.1409
gatR	0.122	0.005	0.1415
bolA	0.124	0.010	0.1441
norR	0.124	0.009	0.1454
sdiA	0.125	0.014	0.1460
malI	0.125	0.017	0.1460
purR	0.125	0.008	0.1467
lrhA	0.125	0.010	0.1467

zur	0.126	0.010	0.1479
narP	0.126	0.015	0.1486
basR	0.126	0.018	0.1492
alaS	0.126	0.015	0.1492
atoC	0.128	0.022	0.1524
envY	0.129	0.021	0.1537
phoB	0.129	0.015	0.1537
uhpA	0.129	0.012	0.1537
fucR	0.129	0.007	0.1543
malT	0.129	0.002	0.1543
hdfR	0.129	0.011	0.1550
pdhR	0.130	0.014	0.1556
gcvA	0.130	0.011	0.1563
zraR	0.131	0.015	0.1582
trpR	0.133	0.013	0.1614
cusR	0.135	0.028	0.1659
hyfR	0.135	0.005	0.1659
baeR	0.135	0.006	0.1659
deoR	0.135	0.009	0.1665
yqhC	0.136	0.015	0.1672
pepA	0.136	0.015	0.1678
ompR	0.136	0.045	0.1684
yiaJ	0.136	0.017	0.1684
tdcR	0.137	0.009	0.1704
yjiE	0.138	0.016	0.1717
cpxR	0.139	0.013	0.1742

hipB	0.139	0.010	0.1742
ascG	0.140	0.009	0.1755
putA	0.140	0.016	0.1761
dinJ	0.142	0.022	0.1800
qseB	0.144	0.023	0.1832
agaR	0.144	0.008	0.1838
trpD	0.145	0.014	0.1845
trpE	0.146	0.011	0.1864
iclR	0.146	0.007	0.1870
dhaR	0.150	0.015	0.1954
cysB	0.155	0.001	0.2043
ihfA	0.158	0.024	0.2095
hns	0.160	0.008	0.2146
flhC	0.166	0.018	0.2255
trpB	0.168	0.017	0.2287
yqjI	0.175	0.022	0.2428
Standard Curve			
0 mmol	0.047	0.004	
0.2 mmol	0.184	0.002	
0.4 mmol	0.256	0.049	
0.6 mmol	0.358	0.033	
0.8 mmol	0.473	0.044	
1 mmol	0.522	0.028	
1.2 mmol	0.560	0.052	

Table C.2: Total indole concentrations of E. coli transcription factor deletants in M9 media

cpxR	0.039	0.002	0.0000
creB	0.038	0.001	0.0000
trpB (No growth)	0.000	0.000	0.0000
trpD No growth)	0.000	0.000	0.0000
trpE No growth)	0.000	0.000	0.0000
paaX	0.039	0.001	0.0000
trpA No growth)	0.000	0.000	0.0000
tnaA	0.038	0.001	0.0000
trpL	0.039	0.002	0.0000
tnaC	0.040	0.002	0.0000
tnaB	0.040	0.002	0.0000
dhaR	0.040	0.001	0.0000
Standard Curve			
0 mmol	0.043	0.000	
0.2 mmol	0.284	0.019	
0.4 mmol	0.403	0.037	
0.6 mmol	0.454	0.003	
0.8 mmol	0.480	0.040	
1 mmol	0.507	0.053	
1.2 mmol	0.502	0.060	