

Lawrence Berkeley National Laboratory

LBL Publications

Title

Sensitivity Analysis of Genome-Scale Metabolic Flux Prediction.

Permalink

<https://escholarship.org/uc/item/9g493040>

Journal

Journal of computational biology : a journal of computational molecular cell biology, 30(7)

ISSN

1066-5277

Authors

Niu, Puhua
Soto, Maria J
Huang, Shuai
[et al.](#)

Publication Date

2023-07-01

DOI

10.1089/cmb.2022.0368

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Sensitivity analysis of genome-scale metabolic flux prediction

Puhua Niu¹, Maria J. Soto², Shuai Huang³, Byung-Jun Yoon^{1,4},
Edward R. Dougherty¹, Francis J. Alexander⁴, Ian Blaby², Xiaoning Qian^{1,4,*}

¹Texas A&M University, College Station, TX 77843, USA

²US Department of Energy Joint Genome Institute,
Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

³University of Washington at Seattle, Seattle, WA 98195, USA

⁴Brookhaven National Laboratory, Upton, NY 11973, USA

*To whom correspondence should be addressed

E-mail: xqian@ece.tamu.edu

May 23, 2023

Keywords: Uncertainty quantification, Bayesian network (BN) structure learning, regulated metabolic network modeling, metabolic engineering, optimal experimental design (OED)

Abstract: TRIMER, Transcription Regulation Integrated with MEtabolic Regulation, is a genome-scale modeling pipeline targeting at metabolic engineering applications. Using TRIMER, regulated metabolic reactions can be effectively predicted by integrative modeling of metabolic reactions with a Transcription Factor (TF)-gene regulatory network (TRN), which is modeled via a Bayesian network (BN). In this paper, we focus on sensitivity analysis of metabolic flux prediction ~~considering potential model~~ [for uncertainty quantification of BN structures for TRN modeling](#) in TRIMER. We propose a computational strategy to construct the uncertainty class of TRN models based on the inferred regulatory order uncertainty given transcriptomic expression data. With that, we analyze the prediction sensitivity of the TRIMER pipeline for the metabolite yields of interest. The obtained sensitivity analyses can guide Optimal Experimental Design (OED) to help acquire new data that can enhance TRN modeling and to achieve specific metabolic engineering objectives, including metabolite yield alterations. ~~We have performed simulation experiments to demonstrate the effectiveness of our developed sensitivity analysis strategy~~

and its potential to effectively guide OED. We have performed small- and large-scale simulated experiments, demonstrating the effectiveness of our developed sensitivity analysis strategy for BN structure learning to quantify the edge importance in terms of metabolic flux prediction uncertainty reduction and its potential to effectively guide OED.

1 Introduction

Optimal Experimental Design (OED) and control for complex biological systems have significant impact on developing new computational strategies in systems and synthetic biology (Balsa-Canto et al., 2021; Zhao et al., 2020) for targeted biochemical overproduction that may benefit human society, for example, in different energy-related and pharmaceutical applications (Ohta et al., 1991; Haro and de Lorenzo, 2001; Luengo et al., 2003; Barrett et al., 2006; Bro et al., 2006; Lü et al., 2011; Esvelt and Wang, 2013). In particular, metabolic engineering, which genetically redesigns microbial strains by gene or reaction knockouts, aims to optimize corresponding biological processes with respect to the desired engineering objective(s). Due to the demanding experimental cost and time to test different microbial strains *in vivo*, computational methods have been developed for *in silico* prediction of useful knockout strategies for beneficial mutants. [Mathematical models to systematically analyze genome-scale metabolic reaction networks have been developed to derive optimal intervention strategies to achieve the desired metabolic reaction fluxes, the turnover rates of the molecules through the corresponding metabolic pathways \(Varma and Palsson, 1994; Edwards and Palsson, 2000; Segre et al., 2002\).](#) However, many existing computational methods to obtain genetically engineered strains are based on genome-scale analysis at steady states assuming the static network models (Varma and Palsson, 1994; Edwards and Palsson, 2000; Segre et al., 2002; Burgard et al., 2003; Shlomi et al., 2005; Ren et al., 2013; Apaydin et al., 2016, 2017). Recent efforts have integrated genetic regulatory relationships involving transcriptional factors (TFs) that may regulate metabolic reactions to achieve more accurate and robust prediction of target metabolic behaviors under different conditions or contexts (Covert and Palsson, 2003b; Shlomi et al., 2007b; Covert et al., 2008b; Chandrasekaran and Price, 2010; Machado and Herrgård, 2014; Reed, 2017; Motamedian et al., 2017; Yu and Blair, 2019). To generalize these integrated hybrid models, we have developed Transcription Regulation Integrated with MEtabolic Regulation (TRIMER) (Niu et al., 2021) as a modeling pipeline targeting at metabolic engineering applications. Using TRIMER, regulated metabolic reactions can be effectively predicted by integrative modeling of metabolic reactions with a TF-gene regulatory network (TRN), where the TRN is modeled via a Bayesian network (BN) inferred from transcriptomic expression data. We have demonstrated promising metabolic flux prediction performances in both simulated and real-world microbial mutant design applications considering transcription regulation in

genome-scale metabolic prediction (Niu et al., 2021, 2022).

While all the existing efforts have demonstrated valid performances in selected model organisms with abundant data and careful manual curation, there is not much investigation on how model uncertainty, due to incomplete system knowledge and/or limited training data, may affect metabolic predictions. To the best of our knowledge, most of the existing works assume that the trained models are deterministic without considering potential model uncertainty. ~~In this paper, we propose to directly analyze how model uncertainty may affect the metabolic engineering performance.~~ Compared to the traditional ways where model uncertainty is often represented by how well models fit the training data, we propose to analyze how model uncertainty affects the metabolic engineering performance directly. ~~In particular, we focus on the sensitivity analysis in our TRIMER framework, by considering an uncertainty class of learned network models instead of deterministic network models to investigate the metabolic engineering performance under uncertainty.~~ In particular, we analyze metabolic flux prediction sensitivity with respect to the uncertainty class of BN structures for TRN modeling as one possible way of uncertainty quantification. Through a mathematical programming formulation of Bayesian network topological ordering, we construct uncertainty classes of BN models for TRN and analyze the metabolic prediction sensitivity of our TRIMER modeling pipeline. We evaluate the sensitivity of the TRIMER pipeline by comparing the ground-truth metabolite yield alterations with TF knockout mutations based on the BN uncertainty classes with the network edge space ranked by topological ordering. To be specific, we simulate both gene expression and metabolic flux data from a predefined ground-truth TRN-regulated metabolic network model, infer models from the simulated data, construct uncertainty classes, and then check the prediction sensitivity by checking the correlation of the predicted and ground-truth metabolite yields. The obtained sensitivity analyses can provide useful guidance for model learning, calibration, and OED for metabolic engineering, allowing biologists to better understand metabolism under perturbation and to take advantage of high-throughput genetic engineering for desired microbial strains with reduced cost.

2 Background

In Niu et al. (2021, 2022), we have developed an integrated regulatory-metabolic hybrid network model and genome-scale metabolic analysis pipeline, Transcription Regulation Integrated with MEtabolic Regulation (or TRIMER). TRIMER enables condition-dependent genome-scale metabolic behavior modeling and provides *in silico* predictions of metabolic engineering tasks, such as knockout phenotype and knockout flux predictions. As a hybrid network model, TRIMER has a metabolic network module that predicts metabolic fluxes based on the classic flux balance analysis (FBA) framework (Covert and Palsson, 2003a;

Palsson, 2015; Edwards and Palsson, 2000; Lewis et al., 2010; Varma and Palsson, 1994). This well-known technique adopts Linear Programming (LP) with flux constraints introduced by metabolic regulation rules (i.e. a metabolic network) of a given organism, to model genome-scale steady-state metabolic status. For accurate and robust prediction of metabolic behaviors under different conditions, regulatory relationships, in the corresponding transcription factor regulatory network (TRN) between transcriptional factors (TF) and their target genes, are integrated with the FBA formulation as additional transcriptional constraints over fluxes (Shlomi et al., 2005; Jensen et al., 2011; Ren et al., 2013; Apaydin et al., 2017; Covert et al., 2008a; Shlomi et al., 2007a). Motivated by another genome-scale framework *PROM* (Probabilistic Regulation of Metabolism) (Chandrasekaran and Price, 2010), TRIMER adopts a TRN module that integrates the *a priori* known TF-gene interaction annotations and available gene/transcriptomic expression profiles to learn a corresponding Bayesian network (BN) for probabilistically modeling the TRN. The conditional probabilities $Pr(\text{gene(s)}|\text{TF(s)})$ can be inferred from the learned BN and used for the construction of regulatory constraints in the FBA formulation, predicting the metabolic behaviors under different conditions, for example, TF knockouts in metabolic engineering. A typical formulation that predicts the regulated metabolic fluxes is shown below:

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

where S is the stoichiometric matrix deduced from the given metabolic network, \vec{v} is a real-value vector representing metabolic fluxes, α and β can be considered as slack variables, κ is a hyper-parameter that controls the penalty of exceeding the fluxes bounds, lb'_i and ub'_i are regulatory flux upper/lower bounds computed based on $Pr(\text{gene(s)}|\text{TF(s)})$ and Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003). Readers can refer to Niu et al. (2021) for more details about flux bounds construction and how TRIMER predicts corresponding metabolic fluxes (Niu et al., 2022). [In practice, the Bayesian network is learned from a gene expression dataset with a limited number of data points, giving rise to potential uncertainty in BN structures, the estimated conditional probabilities \$Pr\(\text{gene\(s\)}|\text{TF\(s\)}\)\$, and consequently, the metabolic flux predictions. In this paper, we focus on how BN structure uncertainty impacts metabolic flux predictions and consider sensitivity analysis of metabolic flux predictions with respect to the learned BN structures as one type of uncertainty quantification. Our goal is to identify the most important edges in terms](#)

of reducing the metabolic flux prediction uncertainty, which can further guide the OED of BN modeling. While TRIMER is motivated by PROM, the adopted flux bounds in PROM are based on $Pr(\text{gene}|\text{TF})$ estimated by the relative frequency of the corresponding gene-TF state pairs in the given TF-gene interaction annotations. As there is no BN modeling involved and network dependency is not considered in PROM, the conducted sensitivity analysis in PROM was on frequentist estimation of the conditional probabilities given individual TF knockout (Chandrasekaran and Price, 2010), which is different from the BN structure uncertainty quantification of metabolic flux predictions in TRIMER investigated in this paper.

3 Materials and Methods

In this section, we first provide a brief overview of our proposed BN structure sensitivity analysis strategy, followed by detailed descriptions of all the steps of the procedure.

3.1 Overview: Sensitivity analysis of TRIMER

Sensitivity analysis of TRIMER is achieved by evaluating metabolic flux prediction performances based on the uncertainty classes of BNs modeling TRN. ~~In other words~~In particular, we focus on the model uncertainty of BN structures that directly impacts metabolic flux prediction. To construct an uncertainty class, the key idea is to first infer the BN topological ordering learned from gene expression data. We then grow various BN structures from the inferred node ordering to construct uncertainty classes at different perturbation levels. By statistical analysis of metabolic predictions with different uncertainty classes, we can have better understanding on how TRN modeling may affect the final metabolic predictions in the TRIMER hybrid model. With such sensitivity analyses ~~due to potential model uncertainty~~ as a way of the BN model uncertainty quantification, practical guidance can be obtained inform researchers for active learning and calibrating the modular components in TRIMER via optimal experiment design ~~with the corresponding uncertainty quantification~~, for example, defining the iterative structure updating policies for BN to improve metabolic prediction.

To develop such a sensitivity analysis capability in TRIMER, we first implement a BN topological order search algorithm to infer the network node ordering of BNs given transcriptomic expression data and the search space of BN edges. The topological ordering of BNs determines the parenthood of nodes: an ancestor node must be of higher order than its descendant nodes. Therefore, given node parent sets and a topological ordering of interest, the global optimal consistent structure can be obtained. To measure the uncertainty of the corresponding topological ordering, ordering samples are obtained by bootstrapping the adopted order search algorithms (Friedman et al., 1999). Accordingly, the statistics of bootstrapped ordering samples

can be computed to help derive the probability distributions and thereafter uncertainty quantification by constructing BN uncertainty classes.

The next question is how we may capture the potential model uncertainty from bootstrapped BN topological orderings and derive a rigorous mathematical framework to construct the classes of uncertain BN models for our proposed metabolic prediction sensitivity analysis. Following Xiao et al. (2018), we adopt a mathematical programming formulation to rank the pairwise ordering of nodes or directed edges in the BN modeling TF-gene regulatory relationships. The formulation aims at identifying the critical regulatory relationships for which reducing the ordering uncertainty can help significantly improve the BN model fitting to the given training dataset. Solving this problem, all valid edges in the given edge space for the BN are ranked by their contributions to the uncertainty reduction, where the uncertainty is captured by the corresponding covariance matrix of ordering scores computed from bootstrapped ordering samples. We then construct the BN uncertainty class of different sizes, where edges with higher rankings are more likely to be sampled to form the allowed edge space to derive the best BN structure consistent with the topological ordering. Finally, transcription-regulated genome-scale metabolic predictions with the BNs in the corresponding uncertainty class can be done following the TRIMER pipeline to investigate prediction sensitivity. The details of the TRIMER analysis pipeline can be referred to Niu et al. (2022).

In summary, our proposed strategy is mainly comprised of three steps: 1) BN topological ordering, 2) uncertainty class construction, and 3) metabolic prediction sensitivity analysis. Figure 1 provides a high-level overview of the strategy, depicting the main workflow.

3.2 Order-based Tabu search

In the TRIMER pipeline (Niu et al., 2021), Bayesian network structure learning is by fitting the given gene expression profiles D for the corresponding gene set $X = \{X_n, n \in [1, N]\}$ and the given TF-gene interaction list E , where X and E are interpreted in BN as the set of network nodes and the search space of network edges respectively. Uncertainty quantification directly by analyzing factorized conditional probability distributions $P(G|D)$ is challenging as it depends on the BN graph structure G , which is combinatorial. We hence propose a topological order based uncertainty quantification strategy, for which we here first define the topological ordering based score, with the uncertainty modeled by a Gaussian distribution as detailed in Section 3.3. To be specific, to learn the BN topological ordering, denoted by \prec , our implementation follows the same idea in Teyssier and Koller (2012), where an order-based heuristic search was proposed. For candidate BN

Algorithm 1: Order-based Tabu search

Input : Dataset D , node set X , edge space E , network in-degree d , initial ordering \prec^0 , initial node scores s^0 , initial node delta-scores by switching orders of adjacent nodes δ^0

Output : Local optimal ordering \prec^*

Parameters: Maximal number of failures T , allowable number of failures before restart R , Tabu length L , score comparison threshold ϵ

```

1 Procedure Tabu_search:
2    $t \leftarrow 0, r \leftarrow 0, s, s^* \leftarrow s^0, \delta \leftarrow \delta^0$ 
3   while  $t < T$  do
4     if  $\sum_i s_i^* \leq \sum_i s_i$  then
5        $s^* \leftarrow s$ 
6     end
7      $op = \text{Tabu\_step}(\prec^t, \delta, \epsilon)$  //finding the swap operator  $op$  with the biggest score improvement//
8     if  $op = \emptyset$  then
9        $r \leftarrow r + 1$ 
10      if  $r \geq R$  then
11         $r \leftarrow 0$ 
12         $\prec^t \leftarrow \text{order\_initialize}(D, E)$  //order initialization//
13      end
14       $op \leftarrow \text{Tabu\_step}(\prec^t, \delta, -\infty)$  //finding the swap operator  $op$  with the smallest score
15      reduction//
16    end
17     $\prec^{t+1} \leftarrow \text{swap}(X_{t_j}, X_{t_{j+1}}; \prec^t)$ 
18    //updating node scores for ordering  $\prec^{t+1}$ //
19     $(s_j, s_{j+1}) \leftarrow (\text{score}(X_{t_{j+1}}, Pa_{\prec^{t+1}}^*(X_{t_{j+1}})), \text{score}(X_{t_j}, Pa_{\prec^{t+1}}^*(X_{t_j})))$ 
20    //updating node delta-scores for ordering  $\prec^{t+1}$ //
21     $(\delta_j^1, \delta_j^2) \leftarrow (-\delta_j^2, -\delta_j^1)$ 
22     $(\delta_{j+1}^1, \delta_{j+1}^2) \leftarrow (\text{score}(X_{t_j}, Pa_{\prec^{t+1}}^*(X_{t_j}) \cup X_{t_{j+2}}) - s_{j+1}, \text{score}(X_{t_{j+2}}, Pa_{\prec^{t+1}}^*(X_{t_j})) - s_{j+2})$ 
23     $(\delta_{j-1}^1, \delta_{j-1}^2) \leftarrow (\text{score}(X_{t_{j-1}}, Pa_{\prec^{t+1}}^*(X_{t_{j-1}}) \cup X_{t_{j+1}}) - s_{j-1}, \text{score}(X_{t_{j+1}}, Pa_{\prec^{t+1}}^*(X_{t_{j-1}})) - s_j)$ 
24     $t \leftarrow t + 1$ 
25  end
26  return  $\prec^*$ 
27 Function Tabu_step( $\delta, \epsilon$ ):
28  //  $H$ : a stack with the fixed length  $L$ , recording historical order operations//
29   $j \leftarrow \text{argmax}_{\{i|(X_{t_i}, X_{t_{i+1}}) \notin H\}} (\delta_i^1 - \delta_{i+1}^2)$ 
30  if  $\delta_j^1 - \delta_{j+1}^2 > \epsilon$  then
31    push  $(X_{t_j}, X_{t_{j+1}})$  into stack  $H$ 
32    return  $(X_{t_j}, X_{t_{j+1}})$ 
33  else
34    return  $\emptyset$ 
35  end
36 end

```

structures, the score of a given ordering \prec can be defined as:

$$\begin{aligned} score(\prec; D; E) &\triangleq score(G_{\prec}^*; D; E) \triangleq \max_{G_{\prec}} score(G_{\prec}; D; E), \\ score(G_{\prec}; D; E) &\triangleq \sum_i^N \left\{ score'(X_i; Pa_{\prec}(X_i, E); D) \right\}, \\ Pa_{\prec}(X_i, E) &\in \{Y : \forall y \in Y, y \prec X_i, (y, X_i) \in E; |Y| \leq d\}, \end{aligned} \quad (1)$$

where $X_i \in X$ denotes a node with its parent node set $Pa_{\prec}(X_i, E)$ consistent with the ordering \prec in the given edge space E , $y \prec X_i$ indicates that the order of y is higher than X_i , d is the predefined upper bound of node's in-degree, and $score(G_{\prec}; D; E)$ is a decomposable score function used in the BN structure learning, such as Bayesian Information Criterion (BIC) score or Bayesian Dirichlet (BD) score. In our implementation, we adopt the BD score. To find the best ordering, we use a heuristic search algorithm—multi-restart *Tabu* search—originally proposed in James et al. (2009) based on the decomposed node-wise score function $score'(\cdot; \cdot; \cdot)$. To perform the ordering search, we define a *swap* operator $swap(X_{t_j}, X_{t_{j+1}})$ over nodes with the adjacent ordering in the t -th iteration as:

$$(X_{t_1}, \dots, X_{t_j}, X_{t_{j+1}}, \dots, X_{t_N}) \rightarrow (X_{t_1}, \dots, X_{t_{j+1}}, X_{t_j}, \dots, X_{t_N}).$$

The best swapping operation is selected among all $n - 1$ candidate successors in the t -th iteration. This simplified search procedure based on the *swap* operator significantly reduces the computational cost of ordering comparison. Supposing that \prec^t is changed to a new ordering \prec^{t+1} by $swap(X_{t_j}, X_{t_{j+1}})$, the delta-score of the induced ordering, difference from the original ordering, only depends on the delta-score of X_{t_j} and $X_{t_{j+1}}$. In addition, the only new operators deduced from \prec^{t+1} are $swap(X_{t_j}, X_{t_{j+2}})$ and $swap(X_{t_{j-1}}, X_{t_{j+1}})$. In each iteration, we can find the optimal parent set for any node in $\mathcal{O}(f_{max}) = \mathcal{O}(N^d)$ as there are $f_{max} = \binom{N}{d}$ possible parent sets per node with the maximum in-degree d (Teyssier and Koller, 2012). With c iterations, the time complexity of the algorithm is $\mathcal{O}(cN^d)$. The pseudo-code of the implementation is provided in Algorithm 1.

3.3 Order initialization

Initialization is crucial to guarantee satisfactory performance of heuristic search as the exact optimality is not ensured due to both the constrained search space and the greedy nature. In this paper, the TF-gene interaction list is not just used to help define the edge space but also as the prior knowledge of node ordering. For example, the ordering of the TF nodes cannot be lower than its regulated gene nodes. Otherwise, the

corresponding TFs cannot be the parent nodes of its regulated genes in BNs. In the following, we will introduce an order initialization algorithm via extracting the node ordering prior knowledge from a given TF-gene interaction list.

In the extreme case, the edge space E defined by a given interaction list corresponds to a directed acyclic graph (DAG), for which the optimal Bayesian network topological ordering is just any ordering consistent with the graph structure. However, in practice, the edge space almost always corresponds to a directed graph with cycles. While the optimal node ordering can not be fully determined directly read off from the graph structure, it is still possible to obtain partial prior knowledge about it. For a directed graph $G = (X, E)$, its strongly connected components (SCCs) denoted as $C = \{C_m : m \in [1, M]\}$ are defined to be the maximal sets of nodes such that for each set, every pair of nodes within the set are reachable from each other. A graph of C can be denoted as $G^c = (C, E^c)$, where an edge exists between two SCCs if there is at least one edge between two nodes belonging to the two SCCs respectively. By the definition of the SCC, G_c must be a DAG. Therefore, G_c determines a component-wise ordering \prec^c , which provides us partial knowledge about how to initialize the node ordering \prec^0 to inform the following order-based heuristic search algorithm. The globally optimal node-wise ordering \prec^* must be consistent with the \prec^c ordering while relative orders among nodes within the same SCC are still undetermined. To obtain an appropriate initial node ordering \prec^0 , G_c and the corresponding ordering \prec^c are first identified from $G = (X, E)$, where SCCs are identified by Tarjan's algorithm (Tarjan, 1972) with the computational complexity $\mathcal{O}(|E| + |X|)$. Next, node ordering \prec^0 consistent with \prec^c can be found easily, where relative orders of nodes belonging to the same SCC are randomly generated. The pseudo-code of the initialization algorithm is shown in Algorithm 2 and the corresponding workflow is illustrated in Figure 6.

Algorithm 2: Order initialization from a given edge space

Input : node set X , edge space E
Output: node order $\prec^0 = (X_{0_1}, \dots, X_{0_N})$

```

1 Procedure order_initialize:
2   // random_perm: function to randomly permute a given ordering//
3   Identify  $C = \{C_m : m \in [1, M]\}$  with  $C_m = \{X_{m_j} : j \in [1, |C_m|]\}$  from  $G = (X, E)$  by Tarjan's
   algorithm
4   Construct  $E^c = \{(C_i, C_j) : i, j \in [1, M], i \neq j; \exists x \in C_i, y \in C_j, (x, y) \in E\}$  based on  $E$  and  $C$ 
5   Identify an ordering  $\prec^c = (C_{s_1}, \dots, C_{s_m}, \dots, C_{s_M})$  consistent with  $G_c = (C, E_c)$ 
6    $i \leftarrow 0$ 
7   for  $m \in [1, M]$  do
8      $(X_{0_{i+1}}, X_{0_{i+2}}, \dots, X_{0_{i+|C_{s_m}|}}) \leftarrow \text{random\_perm}((X_{(s_m)_1}, X_{(s_m)_2}, \dots, X_{(s_m)_{|C_{s_m}|}}))$ 
9      $i \leftarrow i + |C_{s_m}|$ 
10  end
11  return  $\prec^0$ 
12 end

```

3.4 Semidefinite programming formulation for uncertainty class construction

To quantify the uncertainty of topological ordering, we use a multivariate Gaussian random vector $\phi \in \mathbb{R}^{|X|} \sim \mathcal{N}(\mu, \Lambda^{-1})$ as the numerical score representation of an ordering \prec , where each element represents the score of the corresponding node and the nodes of higher order are supposed to have larger scores in ϕ . By *bootstrapping* the previously described order-based search algorithm, we can estimate the corresponding distribution parameters of $\mathcal{N}(\mu, \Lambda^{-1})$ by the bootstrapped samples of ϕ . *Bootstrapping* here means repeatedly perturbing D and applying the order search algorithm on the perturbed data sets to obtain a set of perturbed local optimal orderings. To establish the relationship between node-wise orders and edges, we associate edge $E_k = (X_i, X_j), E_k \in E$ with a real-value random variable $y_k \sim \mathcal{N}(\phi_i - \phi_j, \gamma^{-1})$ to represent the pairwise order difference, where γ is a hyper-parameter. Values of y_k can indeed be interpreted as the confidence of the corresponding pairwise ordering: the larger the value of y_k is, the more confident we are to support the ordering induced by edge E_k . As proposed in Xiao et al. (2018), a binary matrix $B \in \{-1, 0, 1\}^{|E| \times |X|}$ can be used to collectively represent all edges in E , where for $E_k = (x_i, x_j)$:

$$B_{k,l} = \begin{cases} 1, & \text{if } l = i \\ -1, & \text{if } l = j \\ 0, & \text{otherwise.} \end{cases}$$

In cases when only a subset of E is considered, a binary matrix $Diag(v)B$ is used as the corresponding matrix representation, where $v \in \{0, 1\}^{|E|}$ is a binary vector, $Diag(v)$ denotes the corresponding diagonal matrix. Therefore, $y \sim \mathcal{N}(Diag(v)B\phi, \Gamma^{-1})$ represents the pairwise ordering confidence about all the edges of interest, where Γ is a hyper-parameter covariance matrix. It should be pointed that $P(\phi|\mu, \Lambda^{-1})$ is a conjugate prior of $P(y|Diag(v)B\phi, \Gamma^{-1})$ as they are both Gaussian. Therefore, it can be easily verified that (Xiao et al., 2018):

$$P(\phi|y, B^*) = N(\mu', (\Lambda')^{-1}), \quad \Lambda' = ((B^*)^\top Diag(v)B^* + \Lambda), \quad (2)$$

where $B^* = QB$ and $Q^\top Q$ is the Cholesky factorization of Γ . It can be observed that Λ' quantifies the effect of the edge set of interest over the uncertainty of ordering ϕ . Therefore, a straightforward idea to rank edges is by their contribution to the uncertainty reduction. To be more specific, the ranking is achieved by solving

a semidefinite programming problem (SDP) proposed in Xiao et al. (2018), which is defined as:

$$\begin{aligned} \max_{\eta} \quad & \lambda_1 \left(\Lambda + \sum_{i=1}^{|E|} v_i (b_i^T b_i) \right) \\ \text{s.t.} \quad & \mathbf{1}^T v \leq \tau \\ & \mathbf{0} \leq v \leq \mathbf{1}. \end{aligned} \tag{3}$$

In the formulation above, $\lambda_1(\cdot)$ denotes the smallest nonzero eigenvalue of the matrix, $b_i \in \{-1, 0, 1\}^{1 \times |X|}$ denotes a row vector corresponding to the i_{th} row of B^* , and τ is a scalar representing the size of a selected edge subset. The values of the elements in the resulting vector v help identify the top τ edges in items of uncertainty reduction. By solving the SDP repeatedly for $|E|$ times with τ increasing from 1 to $|E|$, the growing ordering of selected edges implies a ranking of edges in E . It should be pointed out that v is relaxed to $[0, 1]^{|X|}$ to guarantee the convexity of the SDP. While this relaxation leads to potential interpretation ambiguity of the solution as the number of non-zero elements in v can be higher than τ . However, we can still select the top τ edges based on the magnitudes of values in v . To construct uncertainty classes of BN models for TRN in TRIMER, ranked edges are assumed to comply with a distribution defined as follows:

$$p(i) = \frac{2}{|E|(|E| + 1)} (|E| - \text{rank}(i)), \tag{4}$$

where i denotes the index of an edge in E and $\text{rank}(i)$ denotes its rank in terms of uncertainty reduction. We then construct a BN uncertainty class in the following way: By the distribution defined above, we first draw multiple edge sample sets of the same size, denoted as $S = \{S_l | l \in [1, L]\}$. For the corresponding sampled set, the best BN structure is identified by $G^{S_l} = \text{argmax}_{G \prec^\mu} \text{score}(G \prec^\mu; D; S_l)$, where \prec^μ is deduced from the numerical mean μ of the ordering samples.

3.5 TRIMER as a simulator

As described in Niu et al. (2021), TRIMER can serve as a simulator of gene expression and metabolic flux data. Given a ground-truth BN model for TRN with appropriate conditional probability tables for each node in the BN, gene expression datasets can be simulated by drawing samples from the distribution described by the BN. Moreover, conditional probabilities with respect to TFs and target genes can be inferred from the BN and used for constructing regulatory flux constraints for genome-scale metabolic predictions when they are integrated with the available metabolic reaction network model of the organism under study. Adding these new constraints into the corresponding FBA formulation for the metabolic network, condition-dependent

metabolic states of the organism can be simulated and treated as the ground truth.

For sensitivity analysis, we simulate gene expression data for BN uncertainty class construction and ground-truth metabolic fluxes to estimate the biomass for different TF-knockout *E. coli* strains as described in Niu et al. (2021). We investigate the model uncertainty by computing the corresponding Pearson correlation coefficients (PCCs) between ground-truth biomass fluxes and the predicted fluxes by TRIMER based on BN uncertainty classes, of which the BN structures may significantly deviate from the ground-truth model that simulates the data.

4 Results

In this section, we present the experimental results based on two simulated datasets to demonstrate the effectiveness of proposed regulatory order based sensitivity analysis strategy for TRIMER.

4.1 Sensitivity analysis results

In our experiments, simulated ground-truth TRIMER models are used to generate gene expression data as well as metabolic flux data (Niu et al., 2021), which are used as the training datasets for BN learning and the ground-truth metabolic flux predictions for sensitivity analysis and performance evaluation. We focus on biomass prediction under multiple TF-knockouts while the proposed methods can be used for other metabolite yield predictions based on the problems of interest. We here use the two same simulated TRIMER models for *E. coli* with iAF1260 (King et al., 2016) as their genome-scale metabolic network model as described in Niu et al. (2021). One model is based on a small-scale TRN and the other is based on a large genome-scale TRN as detailed in the following subsections. More detailed descriptions on the TRIMER models, data sources, software requirements, hardware setups, [as well as run-time statistics of each TRIMER component can be found in Niu et al. \(2021, 2022\)](#). [All the reported experiments are implemented on a PC with Intel i7 processor and 16GB RAM.](#)

4.1.1 Small-scale model sensitivity analysis

For the small-scale model, its corresponding TF-gene regulatory network (TRN) contains 50 nodes (12 TFs and 38 regulated genes) with 118 randomly generated edges. It is assumed that regulated genes can not be the parents of TFs in the BN model. The ground-truth metabolic fluxes in this model are simulated for all the 12 TF single-knockout conditions. Besides, we generate a gene expression dataset of 1000 samples from the ground-truth BN model. [In light of experimental results reported in \(Niu et al., 2021\), a dataset of 1000 training samples is believed to be adequate to guarantee reasonable order learning to achieve desired](#)

metabolic flux prediction performance. Then we bootstrap the Tabu order-based search for 100 times over the dataset obtaining 100 ordering samples, which can finish around 8 to 12 hours where one time running of Tabu order-based search typically takes five to eight minutes. The covariance matrix of ordering scores based on bootstrapped ordering samples is shown in Figure 2(a). As shown in the figure, the variances corresponding to TFs with indices from 1 to 12 are much smaller than the ones for regulated genes indexed from 13 to 50. This is reasonable as the regulated genes are mostly downstream and their orderings can be more uncertain compared to TFs.

Next, edges in the predefined edge space based on the available prior knowledge for the TRN are ranked by their contribution to uncertainty reduction during the ordering posterior covariance updates via the mathematical programming formulation detailed in *Materials and Methods*. As the edge space is relatively small in this small-scale example, the edge ranking can be completed within a few seconds. Ten BN uncertainty classes, each of which contains ten graph structures consistent with the ordering numerical mean, are constructed from sampled edge sets of sizes ranging from 10% to the whole edge space. Finally, the TRIMER model with constructed BN uncertainty classes is built and biomass predictions are made following the TRIMER pipeline. In our experiments, we calculate Pearson correlation coefficients (PCCs) between predictions and simulated ground-truth fluxes to evaluate TRIMER’s performance under a specific BN configuration in the uncertainty class. We have also performed t-tests between the PCC values of metabolic flux predictions obtained based on the uncertainty class covering the whole edges space and the corresponding PCC values from the other uncertainty classes. The corresponding p-values are calculated to show the statistical significance of performance change. The final metabolic prediction performance by TRIMER with different BN uncertainty classes is illustrated in Figure 3(a) for this experiment.

Increasing the size of edge sample sets from 10% to the whole edge space corresponds to decreasing the uncertainty of BN classes by extending the edge space. As the edge space is extended by the top regulatory relationships in terms of BN topological ordering uncertainty reduction, BN classes of high uncertainty can still maintain essential edges. From the plot, one obvious trend is when the uncertainty class is constructed with the allowed BN edge space covering more than 70% of the complete edge space, biomass predictions for TF knockouts closely approach the simulated ground truth results with the average PCC higher than 0.95. In general, the prediction performance improves when considering larger BN structure space as expected. More critically, the prediction sensitivity due to the model uncertainty decreases with the increasing edge search space. With the sampled ranked edges, it is clear that the corresponding edges ranked in the lower 30% may not have much influence on the model prediction and sensitivity performances. On the other hand, when the constructed BN models miss highly ranked edges, metabolic predictions can be quite sensitive to potential model uncertainty even if their model predictions are satisfactory, as demonstrated when the uncertainty

class construction covers the top ranked 40% to 70% edges of the complete edge space. From the plots for small-scale experiments, the prediction performance improves when considering a larger BN structure space. The prediction sensitivity due to the model uncertainty decreases as expected. More critically, one obvious trend is that TRIMER can still achieve relatively good performance when we enforce sparser BN structures, i.e. the uncertainty class construction covers the 40% to 70% edges of the complete edge space. This is because the most important edges found by the proposed SDP are assigned with the highest probabilities for edge sampling to construct the uncertainty class. However, the prediction sensitivity, illustrated by the quantile bars in the plot, is high as these important edges can be missing during edge sampling. On the other hand, when the uncertainty class is constructed with the allowed BN edge space covering more than 70% of the complete edge space, biomass flux predictions for all the TF-knockouts closely approach the simulated ground-truth results with the average PCC higher than 0.95, showing that the corresponding edges ranked in the lower 30% may not have much influence in both model prediction and sensitivity performances. All these trends are further confirmed when we perform t-tests based on the PCC values across different uncertainty class construction settings as illustrated in the plot.

To further evaluate the prediction sensitivity of BN models for the small-scale TRIMER model, we conduct an experiment to investigate how its performance varies by changing the size of training gene expression datasets to construct the BN uncertainty classes. We have generated five expression datasets with sizes ranging from 200 to 1000. For each dataset, we evaluate the performance of TRIMER with the uncertainty classes constructed from the corresponding sampled edge sets whose sizes are fixed to the half of the complete edge space. The experimental results are depicted in Figure 4, [where we provide the t-test p-values based on the corresponding PCC values between the last condition and each of the other conditions](#). Overall, with the increasing training samples for BN learning, both the average prediction accuracy, measured by PCC, and sensitivity, shown in the quantile bar plot, improve in general. The prediction performance may not improve much in average when we have more than 400 training samples. However, when investigating prediction sensitivity to the inherent model uncertainty with our regulatory order based uncertainty class construction, our results indicate that more training samples may need to guarantee the desired robust predictions. In our experiment, we observe that we require 1000 gene expression samples to achieve accurate and stable predictions.

We conduct another experiment to further investigate how flux predictions may be affected by the noise in gene expression data for BN training. We randomly flip the corresponding ON/OFF states of 10%, 20%, \dots , 50% genes in a simulated dataset of 1000 samples, resulting in five perturbed gene expression datasets at different noise levels. For each dataset, we evaluate the performance of TRIMER with the uncertainty classes constructed as previously described based on topological ordering derived from the cor-

responding perturbed gene expression data. For this experiment, we fix the sampled edge set size to be the half of the complete edge space. The experimental results are shown in Figure 5, where t-tests are performed between the uncertainty class constructed from non-perturbed dataset and each of the other uncertainty classes from datasets perturbed at different levels. From the plot, perturbing 10% or 20% of training data does not affect the performances much, the performance declines significantly when more than 30% of gene expression states are flipped, with the corresponding p-value equal to 0.0018 showing the statistical significance of the performance change. As the gene expression noise is reflected in the uncertainty classes of learned BN models, the observed trends of flux predictions with different gene expression noise levels are similar as in the sensitivity analysis with different perturbation levels to BN structures directly.

All these experiments with this small-scale ground-truth model have demonstrated that our topological ordering based sensitivity analysis strategy can appropriately quantify the model uncertainty, identify important edges that directly impact our objective of reliably predicting metabolic fluxes in the TRIMER pipeline. When needed, it can help more effectively refine or calibrate the BN model, for example, by growing the BN edge space based on the ranked edges, optimal experiment design (OED) in the TRIMER pipeline, where BN can be calibrated and updated iteratively based on the proposed sensitivity analysis. On the contrary, blind random edge sampling may not give rise to an informative performance plot.

4.1.2 Large-scale model sensitivity analysis

For the large-scale model, we consider a genome-scale TRN with 1509 edges randomly selected from the edge space containing 3704 edges in the annotated interaction list for *E. coli* in EcoMAC (Carrera et al., 2014). When constructing the BN uncertainty classes in this experiment, we only focus on an edge subspace comprised of 1533 edges, which connect genes directly regulating the reactions involving biomass production in the iAF1260 metabolic network model. We obtain 100 ordering samples by bootstrapping the order search over the simulated gene expression dataset of 1000 samples, similarly as described in the small-scale model experiment. In this set of experiments, the cyclic graph deduced from the interaction list is close to an acyclic graph with only 20 SCCs containing more than two nodes, for which one time bootstrapping of Tabu order-based search can be completed within one to two minutes. The corresponding covariance matrix is shown in Figure 2(b). To construct the BN uncertainty classes, we first fix the BN structure for the genes that are not associated with biomass-related reactions to the optimal structure consistent with the derived mean topological ordering. The run-time of edge ranking for this larger graph is around ten minutes. We then sample different sizes of edge sets from the focused edge subspace. The corresponding BN edge space for the uncertainty classes also grows from containing 10% of the focused edge subspace to the complete space under consideration. Figure 3(b) illustrates the performance under uncertainty.

We observe similar trends as in the small-scale model sensitivity analysis. In general, both metabolic prediction and sensitivity performances improve with the growing edge space. Due to the integration of the EcoMAC interaction list as prior knowledge for defining the BN edge space, we can achieve satisfactory prediction performance, $PCC > 0.95$, when we cover more than 50% of the focused edges. On the other hand, to achieve stable predictions, we may need to cover more than 70% of the defined edge space. [In summary, all the experimental results again verify the effectiveness of our proposed method in quantifying the metabolic flux prediction uncertainty with respect to BN structures and identifying important edges that may directly affect metabolic predictions.](#)

5 Conclusion

From our experimental results, it can be observed that when the BN models deviate more from the optimal BN by missing highly ranked critical edges, the prediction accuracy, measured by correlation between ground-truth biomass fluxes and predicted fluxes of the perturbed BN models, does decrease. More critically, when constructing such uncertainty model classes and investigating prediction sensitivity, our results indicate that we may need better prior knowledge and more training data to achieve both accurate and stable predictions. Our sensitivity analyses also indicate that reliable uncertainty quantification may require more data. While many existing works have reported valid performances in selected experiments, there may still be potential overfitting risks with predictions not easy to generalize when having slightly perturbed systems.

Our topological ordering based sensitivity analysis also helps identify the set of edges, for which the corresponding uncertainty reduction can significantly help model prediction and improve sensitivity. Such a capability can lead to new uncertainty quantification formulations, which may enable optimal experimental design strategies for active model learning (Zhao et al., 2021a,b,c) and more robust intervention strategies in metabolic engineering, which we leave for future research.

Funding

This presented material is 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-0012704.

Acknowledgments

PN and XQ are partially supported by the National Science Foundation under Grants CCF-1553281. This work has been supported by the DOE Joint Genome Institute (<http://jgi.doe.gov>) by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

Contribution statement

Puhua Niu: Methodology, Investigation, Software, Writing - Original Draft, Writing - Review & Editing; **Maria J. Soto**: Investigation, Validation, Writing - Review & Editing; **Shuai Huang**: Conceptualization, Methodology, Writing - Review & Editing; **Byung-Jun Yoon**: Conceptualization, Methodology, Investigation, Formal Analysis, Writing - Review & Editing, Funding Acquisition; **Edward R. Dougherty**: Conceptualization, Investigation, Formal Analysis, Writing - Review & Editing, Funding Acquisition; **Francis J. Alexander**: Conceptualization, Investigation, Formal Analysis, Writing - Review & Editing, Resources, Funding Acquisition; **Ian Blaby**: Conceptualization, Investigation, Validation, Writing - Review & Editing, Funding Acquisition, Resources, Supervision; **Xiaoning Qian**: Conceptualization, Methodology, Investigation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Funding Acquisition, Resources, Supervision.

Author disclosure statement

The authors declare that they have no conflicting financial interests.

1. **Figure 1**: Schematic illustration of the proposed sensitivity analysis workflow: 1) Gene expression data with the prior knowledge on regulatory interactions are used to infer the topological orderings of nodes in Bayesian networks (BNs). 2) The uncertainty class of BNs is constructed based on the uncertainty of the topological ordering due to incomplete knowledge and/or limited data. 3) TRIMER pipeline is used with the uncertainty class to analyze the metabolic flux prediction sensitivity.
2. **Figure 2**: Estimated covariance matrices of topological ordering scores based on bootstrapped samples for (a) small-scale and (b) large-scale network models to help rank the pairwise TF-gene ordering corresponds to edges in the BN uncertainty class. Warmer colors indicate higher uncertainty regarding

the pairwise orderings, hence specifying the corresponding edges may help significantly reduce the BN model uncertainty. The matrices are then used to inform the Semi-Definite programming (SDP) formulation to help quantify the Bayesian network edge importance.

3. **Figure 3:** Box plots of Pearson correlation coefficient (PCC) between the metabolic prediction by TRIMER with constructed BN uncertainty classes and the ground truth. Results are shown for (a) small-scale and (b) large-scale BN uncertainty classes as a function of the number of sampled edges and the average size of the BNs in the uncertainty classes. For both small- and large-scale experiments, we can see that both the prediction performance and prediction sensitivity of the uncertainty classes considering more than 70% of the edge space do not have significant difference from the uncertainty class considering the whole edge space. The p-values of the t-tests based on PCC values between the last uncertainty class covering the whole edge space and each of the other uncertainty classes are also shown in the plot to illustrate the statistical significance of the corresponding performance change.

~~The metabolic prediction performance improves when the uncertainty class is constructed from a larger edge space. Furthermore,~~ The metabolic prediction sensitivity, illustrated by the quantile bars in the plot, decreases in general as the additional edges included in the edge space are less critical to achieve robust predictions. It can be observed that in spite of high sensitivity, TRIMER’s performance can still be relatively high when we enforce sparser BN structures. This demonstrates the effectiveness of the proposed SDP for edge ranking as the most important edges found by the SDP are assigned with the highest sampling probabilities for uncertainty class construction. When considering these edges in constructing the uncertainty classes, metabolic flux prediction performance and sensitivity are relatively stable.

4. **Figure 4:** Box plots of Pearson correlation coefficient (PCC) between the metabolic prediction by TRIMER and the ground truth. Results are shown for uncertainty classes constructed based on simulated gene expression datasets of different sizes. As statistical supports, the p-values of the t-tests between the PCCs in the last condition and each PCC set of the other conditions are also shown in the plot to illustrate the statistical significance of the corresponding performance change. Note that the prediction performance significantly improves when we have 400 training gene expression data points. After that, the prediction performance slowly increases with the increasing number of training data points.
5. **Figure 5:** Box plots of Pearson correlation coefficient (PCC) between the metabolic flux predictions by TRIMER and the ground truth. Results are shown for the uncertainty classes constructed based on gene expression data perturbed at different levels ρ , the flipping probability to perturb gene expression

states. The p-values of the t-tests between the PCCs from the constructed uncertainty class from the non-perturbed gene expression dataset and those using the datasets perturbed at different levels are also shown in the plot to illustrate the statistical significance of the corresponding performance change. Note that the prediction performance drops significantly at 30%. The p-values of the t-tests performed between the PCCs of the uncertainty class constructed from the non-perturbed dataset and each PCC set of the other uncertainty classes from datasets perturbed at different levels are also shown in the plot to illustrate the statistical significance of the corresponding performance change. Note that the prediction performance drops significantly at 30%.

6. **Figure 6:** Illustration of the order initialization workflow. First, the directed acyclic graph (DAG) of strongly connected components (SCCs) is identified from the given directed graph; then a node-wise ordering consistent with the SCC ordering is randomly selected.

References

- Apaydin, M., Zeng, B., and Qian, X. A reliable alternative of optknock for desirable mutant microbial strains. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 573–576. IEEE, 2016.
- Apaydin, M., Xu, L., Zeng, B., and Qian, X. Robust mutant strain design by pessimistic optimization. *BMC Genomics*, 18(6):677, 2017.
- Balsa-Canto, E., Bandiera, L., and Menolascina, F. Optimal experimental design for systems and synthetic biology using AMIGO2. *Methods in Molecular Biology*, (2021;2229):221–239, 2021.
- Barrett, C. L., Kim, T. Y., Kim, H. U., Palsson, B. Ø., and Lee, S. Y. Systems biology as a foundation for genome-scale synthetic biology. *Current Opinion in Biotechnology*, 17(5):488–492, 2006.
- Bro, C., Regenberg, B., Förster, J., and Nielsen, J. *In silico* aided metabolic engineering of *saccharomyces cerevisiae* for improved bioethanol production. *Metabolic Engineering*, 8(2):102–111, 2006.
- Burgard, A. P., Pharkya, P., and Maranas, C. D. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–657, 2003.
- Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *E. coli*. *Molecular systems biology*, 10(7):735, 2014.

- Chandrasekaran, S. and Price, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 107(41):17845–17850, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1005139107. URL <https://www.pnas.org/content/107/41/17845>.
- Covert, M. W. and Palsson, B. O. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of theoretical biology*, 221(3):309–325, 2003a.
- Covert, M. W. and Palsson, B. O. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of Theoretical Biology*, 221(3):309–325, April 2003b. ISSN 0022-5193. doi: 10.1006/jtbi.2003.3071. URL <https://doi.org/10.1006/jtbi.2003.3071>.
- Covert, M. W., Xiao, N., Chen, T. J., and Karr, J. R. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044–2050, 2008a.
- Covert, M. W., Xiao, N., Chen, T. J., and Karr, J. R. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044–2050, 07 2008b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn352. URL <https://doi.org/10.1093/bioinformatics/btn352>.
- Edwards, J. S. and Palsson, B. Ø. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533, 2000.
- Esvelt, K. M. and Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Molecular Systems Biology*, 9(1):641, 2013.
- Friedman, N., Goldszmidt, M., and Wyner, A. Data analysis with Bayesian networks: a bootstrap approach. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 196–205, 1999.
- Haro, M.-A. and de Lorenzo, V. Metabolic engineering of bacteria for environmental applications: construction of *Pseudomonas* strains for biodegradation of 2-chlorotoluene. *Journal of Biotechnology*, 85(2): 103–113, 2001.
- James, T., Rego, C., and Glover, F. Multistart tabu search and diversification strategies for the quadratic assignment problem. *IEEE TRANSACTIONS ON SYSTEMS, Man, And Cybernetics-part a: systems and humans*, 39(3):579–596, 2009.
- Jensen, P., Lutz, K., and Papin, J. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC System Biology*, 5(147), 2011. URL <https://doi.org/10.1186/1752-0509-5-147>.

- King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016.
- Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., and Palsson, B. Ø. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(1):390, 2010. doi: <https://doi.org/10.1038/msb.2010.47>. URL <https://www.embopress.org/doi/abs/10.1038/msb.2010.47>.
- Lü, J., Sheahan, C., and Fu, P. Metabolic engineering of algae for fourth generation biofuels production. *Energy & Environmental Science*, 4(7):2451–2466, 2011.
- Luengo, J. M., Garcia, B., Sandoval, A., Naharro, G., and Olivera, E. R. Bioplastics from microorganisms. *Current Opinion in Microbiology*, 6(3):251–260, 2003.
- Machado, D. and Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLOS Computational Biology*, 10(4):e1003580, 2014. URL <https://doi.org/10.1371/journal.pcbi.1003580>.
- Mahadevan, R. and Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, 2003. doi: 10.1016/j.ymben.2003.09.002. URL <https://www.sciencedirect.com/science/article/pii/S1096717603000582>.
- Motamedian, E., Mohammadi, M., Shojaosadati, S. A., and Heydari, M. TRFBA: An algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinformatics*, 33(7):1057–1063, 2017.
- Niu, P., Soto, M. J., Yoon, B.-J., Dougherty, E. R., Alexander, F. J., Blaby, I., and Qian, X. Trimer: Transcription regulation integrated with metabolic regulation. *Iscience*, 24(11):103218, 2021.
- Niu, P., Soto, M. J., Yoon, B.-J., Dougherty, E. R., Alexander, F. J., Blaby, I., and Qian, X. Protocol for condition-dependent metabolite yield prediction using the trimer pipeline. *STAR protocols*, 3(1):101184, 2022.
- Ohta, K., Beall, D., Mejia, J., Shanmugam, K., and Ingram, L. Metabolic engineering of *klebsiella oxytoca* m5a1 for ethanol production from xylose and glucose. *Applied and Environmental Microbiology*, 57(10):2810–2815, 1991.

- Palsson, B. *Systems Biology*. Cambridge University Press, 2015.
- Reed, J. L. Genome-scale metabolic modeling and its application to microbial communities. In *The Chemistry of Microbiomes: Proceedings of a Seminar Series*. National Academies Press, 2017.
- Ren, S., Zeng, B., and Qian, X. Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints. *BMC Bioinformatics*, 14(S2):S17, 2013.
- Segre, D., Vitkup, D., and Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.
- Shlomi, T., Berkman, O., and Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences*, 102(21):7695–7700, 2005.
- Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3(1):101, 2007a.
- Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*, 3(1):101, 2007b. doi: <https://doi.org/10.1038/msb4100141>. URL <https://www.embopress.org/doi/abs/10.1038/msb4100141>.
- Tarjan, R. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.
- Teyssier, M. and Koller, D. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.
- Varma, A. and Palsson, B. Ø. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/technology*, 12(10):994–998, 1994.
- Xiao, C., Jin, Y., Liu, J., Zeng, B., and Huang, S. Optimal expert knowledge elicitation for Bayesian network structure identification. *IEEE Transactions on Automation Science and Engineering*, 15(3):1163–1177, 2018.
- Yu, H. and Blair, R. H. Integration of probabilistic regulatory networks into constraint-based models of metabolism with applications to Alzheimer’s disease. *BMC Bioinformatics*, 20(386), 2019.
- Zhao, G., Qian, X., Yoon, B.-J., Alexander, F., and Dougherty, E. Model-based robust filtering and experimental design for stochastic differential equation systems. *IEEE Transactions on Signal Processing*, 68: 3849–3859, 2020.

- Zhao, G., Dougherty, E., Yoon, B.-J., Alexander, F., and Qian, X. Efficient active learning for Gaussian process classification by error reduction. In *35th International Conference on Neural Information Processing Systems (NeurIPS)*, 2021a.
- Zhao, G., Dougherty, E., Yoon, B.-J., Alexander, F., and Qian, X. Bayesian active learning by soft mean objective cost of uncertainty. In *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021b.
- Zhao, G., Dougherty, E., Yoon, B.-J., Alexander, F., and Qian, X. Uncertainty-aware active learning for optimal Bayesian classifier. In *9th International Conference on Learning Representations (ICLR)*, 2021c.