



Gene Regulatory Network Inference Relating to Glycolysis in Escherichia coli with Causal Discovery Method Based on Machine Learning

Nakanishi, Akihito
Omino, Natsumi
Owa, Ren
Kinoshita, Hayato
Fukunishi, Hiroaki

(Citation)

Bacteria, 4(4):60

(Issue Date)

2025-11-13

(Resource Type)

journal article

(Version)

Version of Record

(Rights)

© 2025 by the authors. Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license

(URL)

<https://hdl.handle.net/20.500.14094/0100504333>



Article

Gene Regulatory Network Inference Relating to Glycolysis in *Escherichia coli* with Causal Discovery Method Based on Machine Learning

Akihito Nakanishi ^{1,*} , Natsumi Omino ², Ren Owa ³, Hayato Kinoshita ⁴ and Hiroaki Fukunishi ^{3,4,*} ¹ Graduate School of Science, Technology and Innovation, Kobe University, Kobe 657-8501, Hyogo, Japan² School of Bioscience and Biotechnology, Tokyo University of Technology, Tokyo 192-0982, Japan; b0a20043f8@edu.teu.ac.jp³ School of Computer Science, Tokyo University of Technology, Tokyo 192-0982, Japan; c0b21032d7@edu.teu.ac.jp⁴ Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Tokyo 192-0982, Japan; h.kinoshita980@gmail.com

* Correspondence: nakanishiah@phoenix.kobe-u.ac.jp (A.N.); fukunishiha@stf.teu.ac.jp (H.F.); Tel.: +81-78-803-6356 (A.N.); +81-3-637-2111 (H.F.)

Abstract

Escherichia coli LS5218 is an attractive host for producing polyhydroxybutyrate. The strain, however, strongly requires heterologous gene expressions like *phaC* for efficient production. For enhancing the production, the whole gene expressions relating to end product-producing flow should be optimized so that not only heterologous induced-genes but also other relating genes are comprehensively analyzed on the transcription levels, resulting in normally time-consuming mutant-creation. Additionally, the explanation for each transcriptional relationship is likely to follow the relationships on known metabolic pathway map to limit the consideration. This study aimed to infer gene regulatory networks within glycolysis, a central metabolic pathway in LS5218, using machine learning-based causal discovery methods. To construct a directed acyclic graph representing the gene regulatory network, we employed the NOTEARS algorithm (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning). Using transcription data of 264 time-resolved sampling points, we inferred the gene regulatory network and identified several distal regulatory relationships. Notably, *gapA*, a key enzyme controlling the transition between the preparatory and rewarding phases in glycolysis, was found to influence *pgi*, the enzyme at the pathway's entry point. These findings suggest that inferring such nonlocal regulatory interactions can provide valuable insights for guiding genetic engineering strategies.

Keywords: gene regulatory network; machine learning; causal discovery; *Escherichia coli*; approaching genetic engineering



Academic Editor: Margaret L. Britz

Received: 2 September 2025

Revised: 22 October 2025

Accepted: 30 October 2025

Published: 13 November 2025

Citation: Nakanishi, A.; Omino, N.; Owa, R.; Kinoshita, H.; Fukunishi, H. Gene Regulatory Network Inference Relating to Glycolysis in *Escherichia coli* with Causal Discovery Method Based on Machine Learning. *Bacteria* **2025**, *4*, 60. <https://doi.org/10.3390/bacteria4040060>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Escherichia coli is a Gram-negative bacillus belonging to facultative anaerobes. Although *E. coli* strains have a wide variety containing many pathogenic strains such as enterohemorrhagic *E. coli* O157 required for special hygiene precautions [1], many industrially important strains have contributed to the productions of useful substances by improving the technology related to scale-up and CO₂ utilization [2]. *E. coli* K-12 strain, which belongs to the *E. coli* taxonomic category Group A, has branched out into a large

number of mutant strains such as *E. coli* W1485 and *E. coli* LS5218 utilized as industrial producers and as parent strains genetically [3]. As a founder strain, *E. coli* W1485 diverged into *E. coli* K-12 W3110 and *E. coli* K-12 MG1655, which lack the F factor and λ , and these strains are used industrially as host strains for genetic recombination [4]. Stanley Norman Cohen and Herbert Boyer demonstrated producing recombinant DNA with plasmids and restriction enzymes, dramatically improving the industrial utility of *E. coli* [5]. Along with the improving stream, *E. coli* LS5218 genetically mutated on *fadR* (*fadR601*) [3,6–8] and *atoC* (*atoC(c)*) [8–10] has recently attracted attention from the perspective of industrial use [3]. Regarding the genes, *fadR* relates to the regulation of enzyme genes (*fadABDEIJ*) for β -oxidation [6,7] and *atoC* influences to the regulation of the *atoDAEB* operon [11,12]. These gene-disruptions enable *E. coli* LS5218 to produce a wide range of chain lengths of polyhydroxybutyrate (PHB) [8,13], which shows potential as a raw material in the production of biodegradable plastics [3,14]. The productivity of PHB does not meet the demand of polyhydroxybutyrate-production so that the strain strongly requires heterologous gene over-expressions such as *phaC*, *phaJ*, and *phaG* efficiently with genetic engineering [9,15]. For enhancing the production, the whole gene expression relating to the end product-producing flow should be optimized, meaning that not only heterologous induced-genes but also other relating genes ought to be comprehensively analyzed on the transcription levels. Even though the targeted genes were well improved on the transcription levels by heterologous over-expressions, if the whole gene was not well balanced on the transcription levels to express the enzymes making the metabolic flow into the end product, the trials of the genetic engineering are not finally suitable for the objective production. Therefore, not only the specific genes such as introduced genes but also the genes relating to the whole metabolic flow should be comprehensive analyzed on the transcription levels. However, analyses of the whole transcription levels typically require construction of the mutant strains modified on the targeted genes, facing problems of time-consuming and costly strain preparation steps. Although omics analyses such as transcriptomics [16] and metabolomics [17] have drastically improved to reveal the details of vital activities of *E. coli* in recent decades, these analyses are still time-consuming and costly. Additionally, the explanation for each transcriptional relationship is used to easily follow the relationships on the known metabolic pathway map and prone to limiting the scope of consideration to find gene regulatory networks.

Machine learning (ML) has expanded the academic field of bioinformatics, as it can analyze large-scale biological data and uncover patterns or insights that are difficult for humans to detect. Network models have been used as computational approaches to describe gene regulation by transcription factors. These models consist of nodes that represent genes and directed edges that represent regulatory interactions. In their review article, Chai et al. [18] compared six widely used computational approaches for constructing gene regulatory networks: Boolean networks (BNs), probabilistic Boolean networks (PBNs), Bayesian networks (BNs), dynamic Bayesian networks (DBNs), ordinary differential equation models (ODEs), and neural networks (NNs) such as recurrent neural networks (RNNs). They emphasized that the performance of each method depended on the nature of the data, the presence of noise, the size of the network, and the available computational resources. It was suggested that methods such as DBNs and NNs generally offered high performance at the expense of increased computational costs, whereas BNs and PBNs were computationally efficient but limited in performance.

Moeckel et al. [19] also surveyed previous approaches for inferring gene regulatory networks in recent years. In their review, gene regulatory network inference methods were systematically categorized into statistical, ML, and deep learning approaches. Statistical methods, such as correlation-based metrics, mutual information, and Bayesian networks,

were typically applied in an unsupervised learning framework using gene expression data. In contrast, supervised learning approaches including regression models, support vector machines (SVMs), and random forests (RFs) treated transcription factors and target gene relationships as binary classification tasks, enabling the inference of directed regulatory edges. Furthermore, the authors highlighted the growing role of deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), transformers, and diffusion-based architectures. These models were particularly effective at capturing nonlinear and high-order dependencies, making them well-suited for high-dimensional and complex datasets such as single-cell RNA sequencing and multi-omics data. Rather than recommending a single best method, the review emphasized the importance of selecting inference strategies based on the specific characteristics of the data and research objectives.

Li et al. [20] conducted gene regulatory network inference using dynamic Bayesian networks (DBNs), a causal discovery method based on directed acyclic graph (DAG) structures. The study utilized time-series gene expression data from *Drosophila melanogaster* covering various developmental stages. Each gene was modeled as a binary-valued random variable, with expression states classified as either repressed or active at each time point. Multiple subnetworks, ranging from 12 to 40 genes involved in muscle development, were constructed. Inference performance was evaluated based on the number of correctly identified edges, missed edges, false positives, recall, and precision. The highest performance was observed in the network with 12 nodes and 18 edges, achieving a recall of 64.2% and a precision of 82.5%. These results demonstrated the potential of a causal discovery method based on DAG structures for effective gene regulatory network inference.

This study aimed to infer gene regulatory networks within glycolysis as central metabolism in *E. coli* LS5218 by using causal discovery methods based on DAG structures using comprehensive data which changed over time in cultivation. In this study, we employed the NOTEARS algorithm (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning) [21] to construct directed acyclic graphs (DAGs). NOTEARS reformulates the structure learning problem of Bayesian networks, traditionally treated as a combinatorial optimization problem, into a continuous optimization problem. This transformation enables efficient exploration of the search space by leveraging gradient information, allowing for faster and more scalable structure learning. While conventional methods suffer from exponential increases in computational complexity as the number of nodes grows, NOTEARS remains applicable to large-scale networks. Furthermore, the continuous optimization framework facilitates the easy incorporation of penalty terms and complex constraints, providing flexibility for diverse applications. Recently, Zhu et al. [22] systematically compared causal discovery algorithms across three major categories—conditional independence test-based methods (PC), score-based methods (GES), and restricted structural equation model-based methods (LiNGAM)—as well as a suite of continuous optimization-based methods (NOTEARS, DAGMA, DECI) using actual gene regulatory network data. To evaluate algorithmic performance, they used large-scale observational gene expression datasets such as RNA-seq, single-cell RNA-seq (scRNA-seq), and Nanostring. As a result, they found that the PC algorithm excels in inferring DAG structure, but has the limitation that it can only estimate the direction of causality and cannot quantitatively estimate causal effect sizes. In contrast, NOTEARS-based continuous optimization algorithms enable both structure inference and quantitative estimation of causal effects. Notably, PC-NOTEARS (PCnt), which combines the structural inference power of PC with NOTEARS, demonstrated the best overall performance in both structural metrics and effect size estimation. These findings indicate that NOTEARS-based methods are superior in terms of computational efficiency, quantitative causal inference, and

applicability to biological research, making them a powerful choice for gene regulatory network discovery.

The investigation required the sufficient differential data so that the transcriptional data of *E. coli* LS5218 cells were systematically collected every 5 min, resulting in 264 datum-sets of 33 genes, in total, as enough datum-source for revealing comprehensive gene regulatory networks. The gene regulatory networks were especially evaluated in the logarithmic growth phase and stationary phase. Preventing the influences from background information about the gene relationships on the known metabolic maps, the known metabolic pathway map was not given as background information to infer the gene regulatory networks under the ML in advance. The causal discovery method inferred many gene regulatory networks which were difficult to show through limited human consideration, meaning that this paper showed the unprecedented contents that the causal discovery method, without influences from information of the known metabolic pathway, inferred from the gene regulatory networks.

2. Materials and Methods

2.1. Cultivation and Collecting of *E. coli* Cells

Escherichia coli LS5218 [23] was cultured and collected for analysis. To collect cDNAs, the strain was cultivated in three steps in 10 mL of Luria–Bertani (LB) medium with rotation at 130 strokes per minute at 37 °C. Firstly, as a recovery-cultivation, the strain was picked up from its glycerol stock and cultivated for 6 h; secondly, as a pre-cultivation, the cultivated strain was also cultured for 6 h; finally, as a main cultivation, the strain was inoculated at 0.01 of an initial OD₆₀₀, and then started the cultivation. Cells in the main cultivation were collected every 5 min from 2 h to 24 h in order to obtain comprehensive data of gene transcription levels over time. Accurate repeating cultivations were performed to collect the cells under strict culture management.

2.2. Measurement of Gene Transcription Levels

Five mg of the cultured *E. coli* cells was harvested by centrifugation at 21,500× *g* for 5 min at 23 °C. The collected cells were added with 50 µL of QIAzol Lysis Reagent (Qiagen, Hilden, Germany) and shaken well for 5 min at 23 °C. The mixture was also added with 10 µL of chloroform and left for 3 min. Treated samples were centrifuged at 21,500× *g* for 15 min at 4 °C, and then the supernatants were transferred into RNase free microtubes. The supernatant was mixed with 25 µL of added isopropanol by pipetting and put at 23 °C for 10 min. The put mixture was centrifuged at 21,500× *g* for 15 min at 4 °C, and the precipitant was collected by discarding the supernatant. After rinsing the precipitants by an ethanol precipitation method, those were dried out with freeze dryer Refrigerated CentriVap Benchtop Vacuum Concentrator (Labconco Corporation, Kansas City, MO, USA). The total RNA solution was obtained by dissolving the precipitants completely with 10 µL of RNase free water. The complementary DNA (cDNA) was synthesized from the total RNA solution by using ReverTra Ace qPCR RT Master Mix with gDNA remover (Toyobo Co., Ltd., Osaka, Japan) and Thunderbird SYBR qPCR Mix (Toyobo Co., Ltd.). Quantitative polymerase chain reaction (qPCR) was performed for 33 genes relating to glycolysis with StepOne real-time qPCR system (Thermo Fisher Scientific, Waltham, MA, USA). After obtaining cycle threshold (Ct) values referring to the number of amplification cycles, each transcription level was normalized with a housekeeping gene *rrsA* by relative quantification [24]. The primer pairs for those genes were designed with Primer3Plus algorithm (<https://dev.primer3plus.com/index.html>, accessed on 1 October 2023) using gene information from the National Center for Biotechnology Information (NCBI) (Table S1 in Supporting File).

2.3. Machine Learning-Based Techniques for Inferring Gene Regulatory Networks

The culture phase of *E. coli* was broadly divided into the logarithmic growth phase (2–10 h) and stationary phase (10–24 h). Gene regulatory networks were inferred through a two-step process consisting of a machine learning (ML)-based causal discovery step and a path analysis step. This process was independently applied to gene transcription datasets obtained from the logarithmic growth phase and the stationary phase.

2.4. Causal Discovery Step

In the ML-based causal discovery method, causal relationships between variables in a dataset are represented as a directed acyclic graph (DAG). In a DAG, each variable corresponds to a node, and the directed edges represent causal relationships between nodes. By definition, DAGs are subject to structural constraints that prohibit cycles; that is, starting from any node, it is impossible to return to the same node. Such DAGs provide both a visual and theoretical representation of the underlying causal structure and can be regarded as generative models that describe how the observed data were produced. Based on the DAG framework, which assumes causal relationships among transcriptional variables, we adopted an ML-based causal discovery method to infer gene regulatory networks from gene transcription data.

To construct DAGs, we employed the NOTEARS algorithm (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning) [21]. NOTEARS reformulates the structure learning problem of Bayesian networks, traditionally treated as a combinatorial optimization problem, into a continuous optimization problem. NOTEARS evaluates network quality using an objective function that combines regression error with a penalty for violations of the DAG constraint. This transformation enables efficient gradient-based search of the structure space, significantly improving scalability compared to conventional methods. In this study, we utilized NOTEARS via CausalNex [25], a comprehensive Python 3 library for causal discovery. Specifically, we used the `from_pandas` function from the `causalnex.structure.pytorch` submodule, which provides a DAG structure learning algorithm based on NOTEARS. The algorithm estimates a weighted adjacency matrix, where each entry represents the strength of a directed edge between two variables. For the final output, the weighted adjacency matrix was binarized so that each element was represented as 1 or 0. This matrix represents the adjacency matrix. The detailed theoretical framework for DAG structure learning using this submodule is described in Supplementary S1 of the Supporting file. The constructed DAG, inferred via NOTEARS, serves as a candidate model for representing gene regulatory networks.

2.5. Path Analysis Step

The adjacency matrix obtained from the DAG in the causal discovery step was converted into a structural equation model (SEM) specification suitable for path analysis, which was subsequently performed. In this step, the SEM specification and the original gene transcription data were used as inputs to compute standardized path coefficients and fit indices.

While NOTEARS evaluates network quality using an objective function that combines regression error with a penalty for violations of DAG constraints, this criterion provides only a relative measure of optimization performance. The resulting structure depends on the optimization procedure and does not offer a framework for assessing statistical uncertainty or sampling variability. Consequently, NOTEARS alone cannot provide statistical significance tests or model fit indices. In contrast, path analysis offers a well-established statistical framework that enables quantitative evaluation of model adequacy, uncertainty, and reliability. For example, it allows statistical verification of model fit using fit indices

and provides significance testing for individual path coefficients. Moreover, unlike the interpretation of influence based on the weighted adjacency matrix produced by NOTEARS, path analysis allows for a clear distinction between direct and indirect effects. In addition, by using standardized coefficients, it becomes easier to compare effect sizes across different pathways.

For this purpose, we employed semopy (Structural Equation Model Optimization in Python) [26], a Python library for structural equation modeling. The theoretical framework for path analysis is described in Supplementary S2 of the Supporting File. To assess the discrepancy between the inferred DAG and the observed data, we evaluated multiple fit indices. A higher level of fit indicates that the path coefficients are valid and that the DAG adequately explains the data. The fit indices considered in this study include the Chi-square test (χ^2), Comparative Fit Index (CFI), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Normed Fit Index (NFI), and Root Mean Square Error of Approximation (RMSEA). Thresholds for good-fit and acceptable-fit were determined based on Table 1 [27], which presents standard criteria commonly used in studies employing path analysis. For visualization of the path diagram, standardized solutions were used for the path coefficients.

Table 1. Criteria for fit indices.

Fit Index *	Good-Fit	Acceptable-Fit
χ^2/df	$0 \leq \chi^2/\text{df} < 2$	$2 \leq \chi^2/\text{df} \leq 5$
CFI	$0.95 \leq \text{CFI} \leq 1.00$	$0.90 \leq \text{CFI} < 0.95$
GFI	$0.95 \leq \text{GFI} \leq 1.00$	$0.90 \leq \text{GFI} < 0.95$
AGFI	$0.90 \leq \text{AGFI} \leq 1.00$	$0.85 \leq \text{AGFI} < 0.90$
NFI	$0.95 \leq \text{NFI} \leq 1.00$	$0.90 \leq \text{NFI} < 0.95$
RMSEA	$0.00 \leq \text{RMSEA} \leq 0.05$	$0.05 < \text{RMSEA} \leq 0.08$

* The abbreviation df represents the degrees of freedom.

2.6. Optimization of DAG Model

In the causal discovery step of DAG structure learning, the regularization parameters λ_1 and λ_2 for NOTEARS, along with the threshold t for generating the adjacency matrix, were treated as hyperparameters. Hyperparameter tuning is crucial for improving the inference of DAG models. We optimized λ_1 , λ_2 , and t using Bayesian optimization, an adaptive experimental design method that sequentially selects candidate parameter sets based on prior evaluations. Bayesian optimization was performed using the Python library Optuna [28], which employs the tree-structured Parzen estimator (TPE) algorithm [29] as its underlying method. In this study, the optimization was performed within a multi-objective framework, aiming to simultaneously minimize RMSEA and maximize AGFI among the fit indices in path analysis step.

Figure 1 presents a conceptual diagram of the optimization process. This combined process of causal discovery step and path analysis step was repeated until the number of trials specified by Bayesian optimization was reached. The predefined search ranges for the hyperparameters were set as $\lambda_1 = 0.01\text{--}0.1$, $\lambda_2 = 0.01\text{--}0.1$, and $t = 0.2\text{--}0.6$. The number of optimization iterations was set to 300. DAGs that satisfied $\text{RMSEA} < 0.15$ and $\text{AGFI} > 0.90$ were recorded as candidate gene regulatory networks, and their adjacency matrices, fit indices, and path coefficients were stored. To prevent overfitting, networks exhibiting perfect fit (e.g., $\text{RMSEA} = 0.0$ and $\text{AGFI} = 1.0$) were excluded using Optuna's pruning mechanism. When a trial was pruned, its computation was terminated early, and a new parameter set was sampled for the next trial, ensuring efficient exploration of the search space.

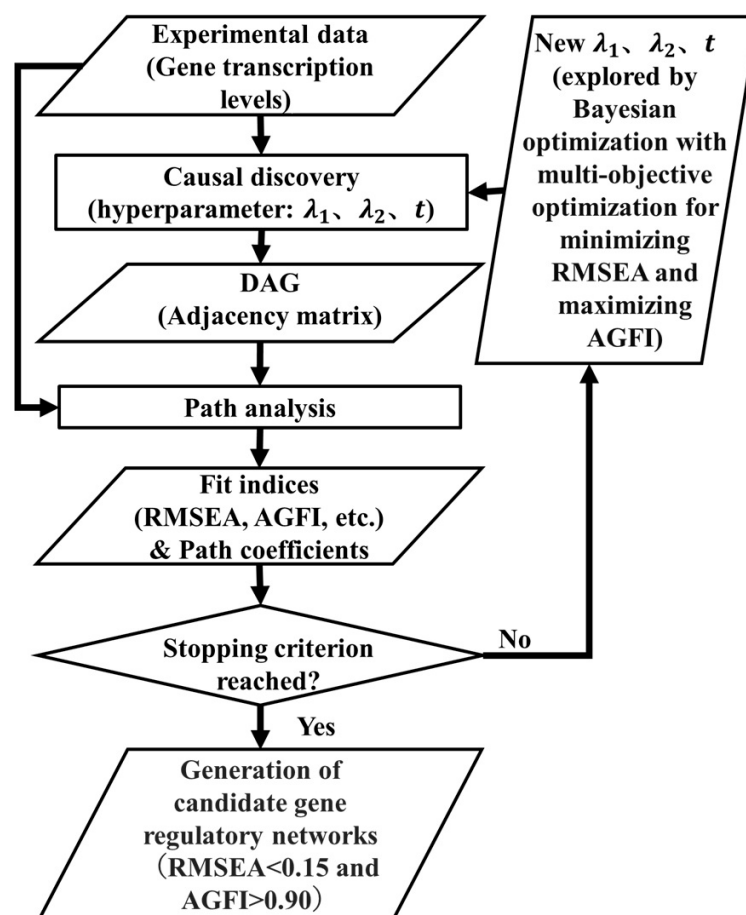


Figure 1. Overview of hyperparameter optimization for generating candidate gene regulatory networks. Hyperparameters include L1 regularization weight λ_1 and L2 regularization weight λ_2 and threshold t for generating the adjacency matrix from the weighted adjacency matrix.

3. Results

3.1. Collection of Transcription Levels

Escherichia coli stably cultured in LB medium were collected every 5 min for 2–24 h and showed high reproducibility of the growth. After converting the transcriptional data of enzyme genes related to glycolysis into numerical data for the ML, the data were used for analyzing the gene regulatory networks in the logarithmic growth phase (2–10 h) and stationary phase (10–24 h). Stable Ct values of *rrsA* were obtained at each sampling point.

3.2. Fit Indices of Inferred Gene Regulatory Networks

In the process of inferring gene regulatory networks, not only the network with the highest fit indices but also multiple top candidate networks were recorded. Rather than automating the inference process entirely, we aimed to support life science researchers in their interpretative evaluation. By presenting multiple high-ranking candidate networks, researchers are afforded the opportunity to assess them comprehensively and potentially derive novel biological insights.

Accordingly, gene regulatory networks satisfying the criteria of $\text{RMSEA} < 0.15$ and $\text{AGFI} > 0.90$ were selected as top candidates. The optimized hyperparameter values (λ_1 , λ_2 , and t) used to infer networks in both the logarithmic growth and stationary phases are listed in Table 2. These values fall well within the predefined search ranges and are not concentrated near the boundaries, suggesting that the current search space is appropriately defined.

Table 2. Optimized hyperparameter values for inferring gene regulatory networks in (A) logarithmic growth phase and (B) stationary phase.

Network *	Hyperparameters		
	λ_1	λ_2	t
(A) Logarithmic growth phase			
1	0.032782	0.032834	0.413330
2	0.034076	0.032098	0.325320
3	0.033515	0.034091	0.361556
4	0.017767	0.033161	0.415738
5	0.038384	0.042828	0.322086
6	0.053201	0.020239	0.409680
7	0.022126	0.033681	0.415102
8	0.054958	0.033772	0.479814
(B) Stationary phase			
1	0.017297	0.020722	0.437612
2	0.049361	0.014726	0.386253
3	0.049095	0.021144	0.434112
4	0.012722	0.011658	0.446718

* The network numbers match the network numbers in Table 3 and Figure 2.

The fit indices for the gene regulatory networks inferred during the logarithmic growth and stationary phases are summarized in Table 3. In the logarithmic growth phase, the top five networks satisfied the criteria for either good-fit or acceptable-fit across all indices. The sixth and seventh networks did not satisfy the acceptable-fit threshold for RMSEA but met the criteria for the remaining indices. The eighth network failed to satisfy the acceptable-fit criteria for both χ^2/df and RMSEA, yet still fulfilled the good-fit criteria for the other indices. In the stationary phase, the top two networks met either good-fit or acceptable-fit criteria across all indices. The third and fourth networks did not meet the acceptable-fit thresholds for χ^2/df and RMSEA but satisfied the good-fit criteria for the remaining indices. Statistical fit indices indicate how well the inferred DAG reproduces the observed data. Although excellent statistical fit can be useful for supporting the discussion of potential causal relationships, it should be noted that it does not guarantee biological validity. Therefore, the results were interpreted with caution in Section 4.

Table 3. Fit indices for the inferred gene regulatory networks in (A) logarithmic growth and (B) stationary phases.

Network	Fit Indices *					
	χ^2/df	CFI	GFI	AGFI	NFI	RMSEA
(A) Logarithmic growth phase						
1	1.392	0.993	0.978	0.965	0.978	0.058
2	1.583	0.978	0.944	0.926	0.944	0.071
3	1.828	0.974	0.945	0.916	0.945	0.084
4	1.949	0.972	0.945	0.922	0.945	0.090
5	2.071	0.972	0.948	0.922	0.948	0.096
6	2.451	0.971	0.953	0.915	0.953	0.112
7	2.879	0.971	0.958	0.916	0.958	0.127
8	3.476	0.976	0.967	0.934	0.967	0.146
(B) Stationary phase						
1	1.222	0.999	0.995	0.991	0.995	0.032
2	2.752	0.992	0.987	0.977	0.987	0.091
3	3.342	0.983	0.976	0.959	0.976	0.106
4	5.064	0.982	0.978	0.959	0.978	0.139

* The table is sorted in ascending order of RMSEA. Bold and underlined values indicate a good-fit, while bold values indicate an acceptable-fit. The abbreviation df represents the degrees of freedom. The network numbers match the network numbers in Table 2 and Figure 2.

The corresponding path diagrams for the inferred gene regulatory networks in the logarithmic growth and stationary phases are shown in Figure 2. We added standardized path coefficients obtained through path analysis to these diagrams, thereby providing quantitative information on both the magnitude and direction (positive or negative) of gene-to-gene influence. This addition enables the separate interpretation of direct and indirect effects, while also allowing for the comparison of effect sizes across different pathways.

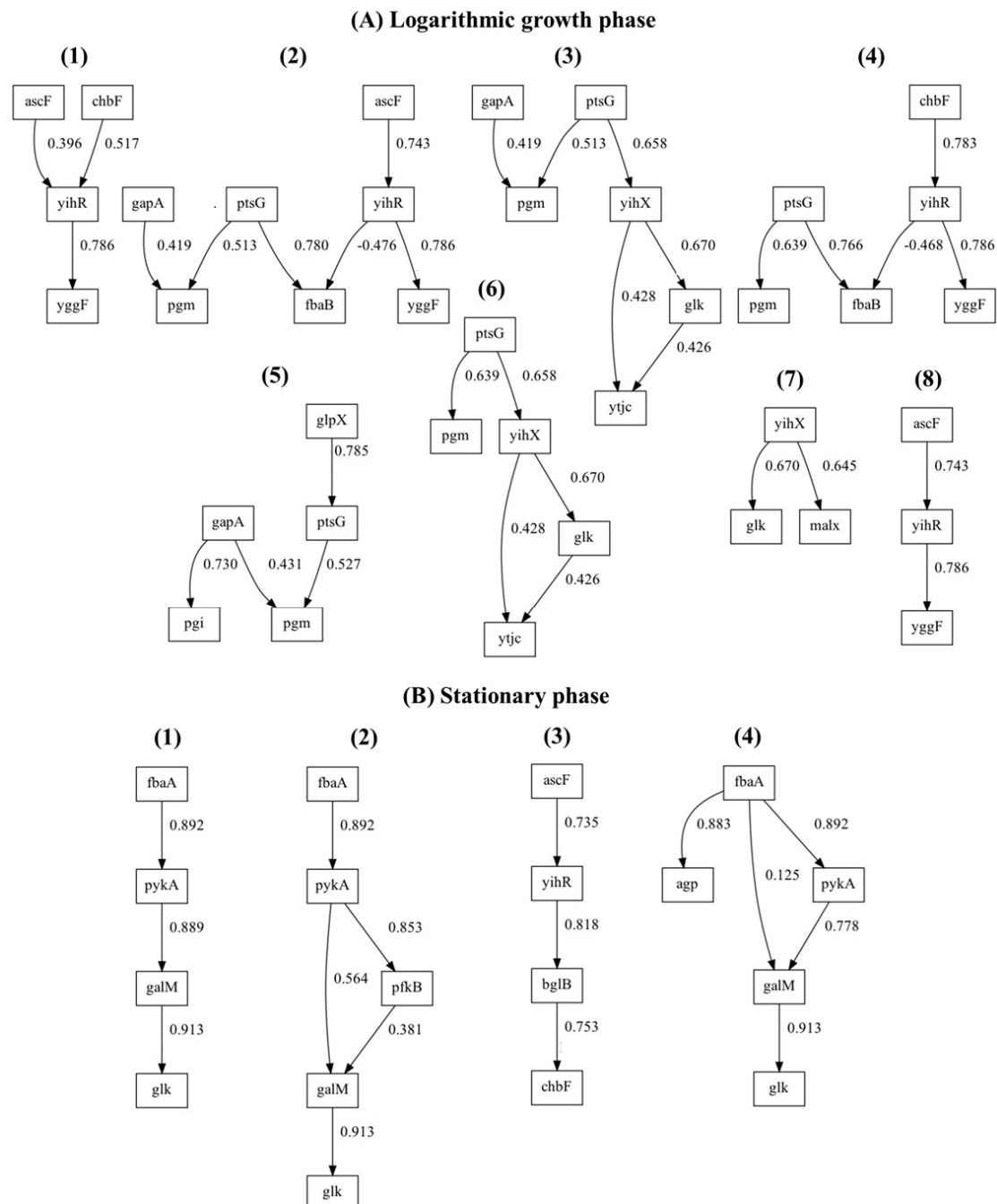
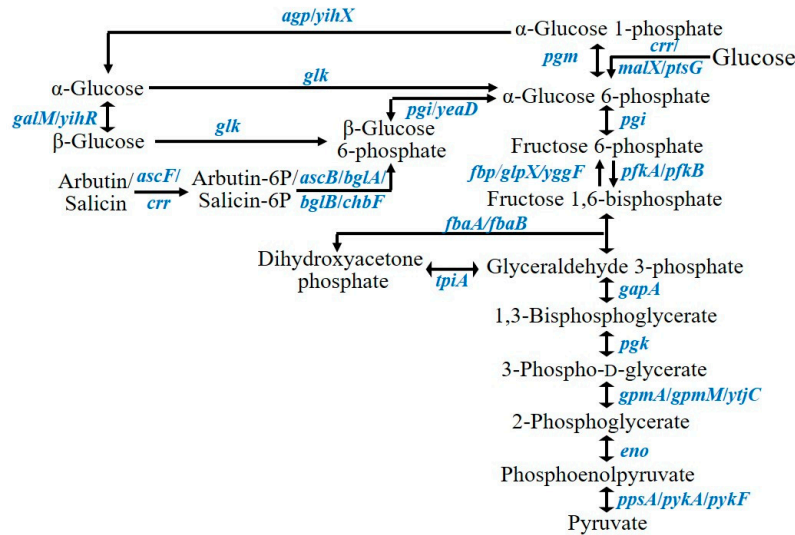


Figure 2. Path diagrams of inferred gene regulatory networks. The diagrams correspond to (A) the logarithmic growth phase and (B) the stationary phase of *Escherichia coli*. Each node represents a gene, and standardized path coefficients are indicated on the edges. Numbers in parentheses denote the network numbers listed in Tables 2 and 3.

3.3. Analysis of the Inferred Gene Regulatory Networks on a Metabolic Pathway Map

In the logarithmic growth phase, the inferred gene regulatory networks were categorized into three types as follows: (I) *yihR*→*yggF* (+), (II) *ptsG*→*pgm* (+), and (III) *yihX*→*glk* (+) (Figure 3A).

Metabolic pathway map as a background



(A) Logarithmic growth phase

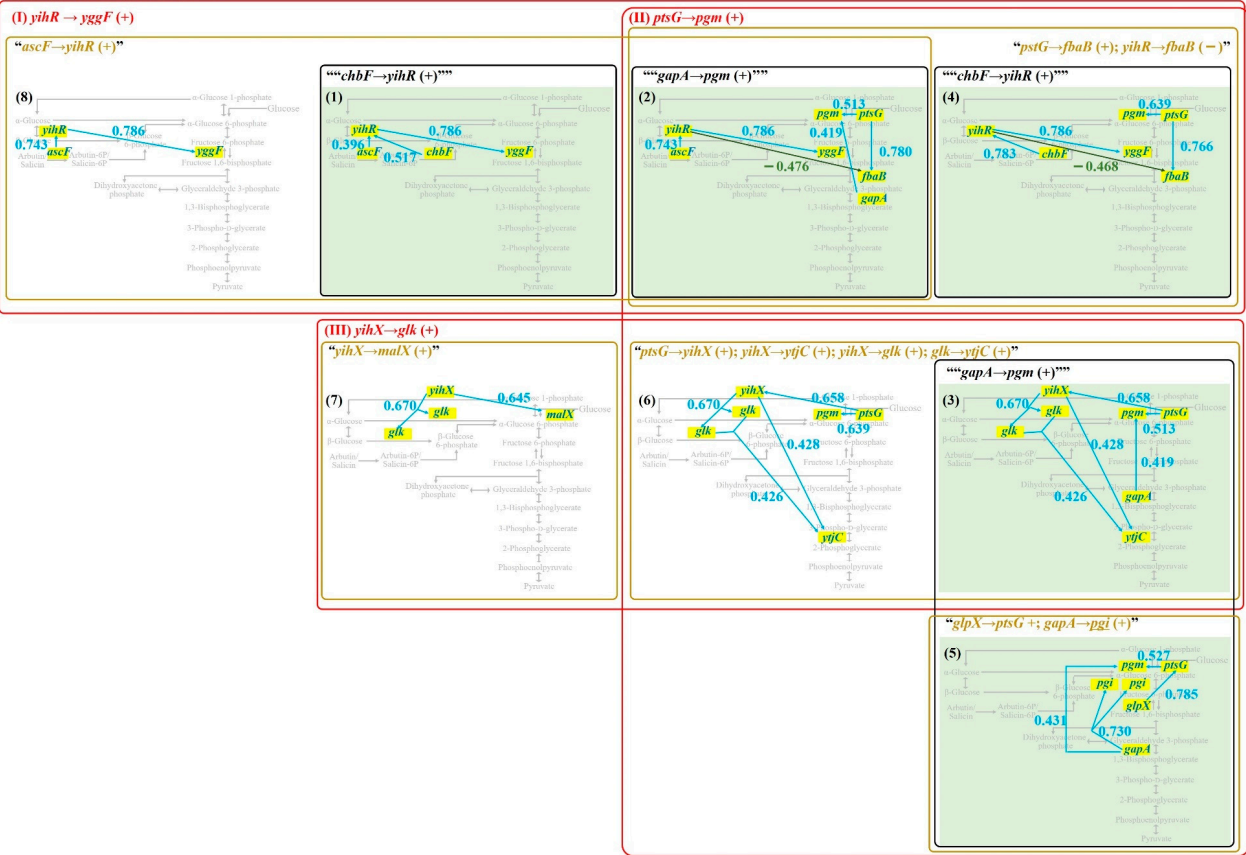


Figure 3. Cont.

(B) Stationary phase

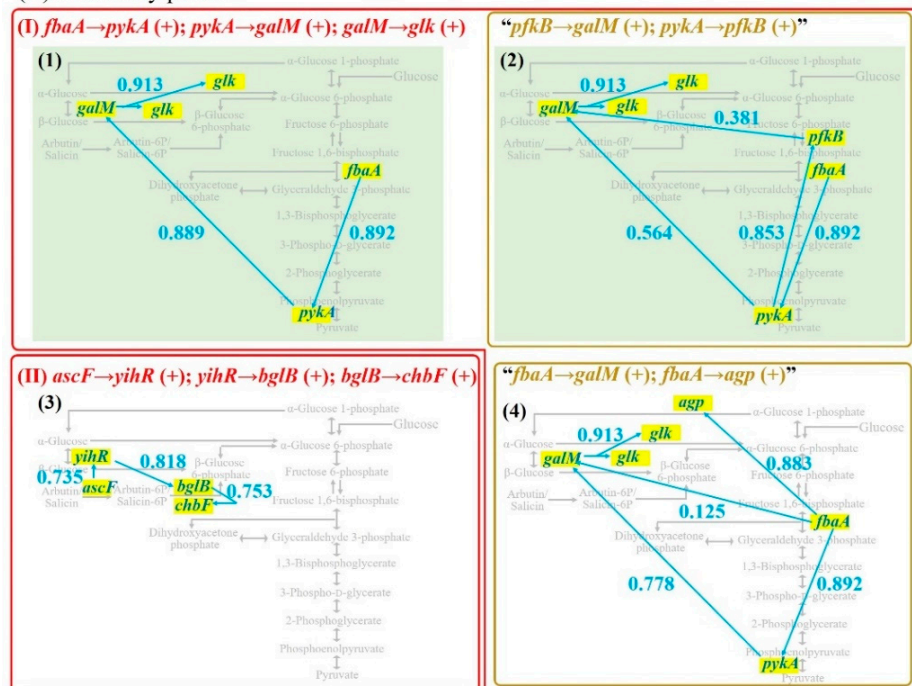


Figure 3. Inferred gene regulatory network of gene transcription levels in glycolysis. The results of ML based on increases and decreases in gene transcription levels when *Escherichia coli* LS5218 was cultured in LB medium show (A) eight inferred gene regulatory networks in the logarithmic growth phase (2–10 h) and (B) four in the stationary phase (10–24 h). The direction of relationship in gene transcription levels is indicated by arrows, the strength of relationship is indicated by numbers as standardized path coefficients above the arrows, and positive and negative numbers are indicated in blue and green, respectively. Inferred gene regulatory networks under the condition of Root Mean Square Error of Approximation (RMSEA) < 0.10 and RMSEA < 0.15 are shown with and without colored backgrounds.

The (I) $yihR \rightarrow yggF (+)$ group contained four networks under RMSEA < 0.15 and three networks under RMSEA < 0.10. The group was additionally categorized as “ $ascF \rightarrow yihR (+)$ ” by binding the outermost groups. Moreover, in the “ $ascF \rightarrow yihR (+)$ ” group, relationships of “ $chbF \rightarrow yihR (+)$ ” and “ $gapA \rightarrow pgm (+)$ ” emerged.

The (II) $ptsG \rightarrow pgm (+)$ group encompassed five networks under RMSEA < 0.15 and four networks under RMSEA < 0.10. The group was divided into three groups, “ $ptsG \rightarrow fbaB (+)$; $yihR \rightarrow fbaB (-)$ ”, “ $ptsG \rightarrow yihX (+)$; $yihX \rightarrow ytiC (+)$; $yihX \rightarrow glk (+)$; $glk \rightarrow ytiC (+)$ ”, and “ $glpX \rightarrow ptsG$; $gapA \rightarrow pgi (+)$ ”, by binding the outermost groups. As the other group, “ $gapA \rightarrow pgm$ ” was also exhibited lying on the part of “ $ptsG \rightarrow yihX (+)$; $yihX \rightarrow ytiC (+)$; $yihX \rightarrow glk (+)$; $glk \rightarrow ytiC (+)$ ” and “ $glpX \rightarrow ptsG$; $gapA \rightarrow pgi (+)$ ”.

The (III) $yihX \rightarrow glk (+)$ group covered three networks under RMSEA < 0.15 and a network under RMSEA < 0.10. The group was also divided into two groups: “ $yihX \rightarrow malX (+)$ ” and “ $ptsG \rightarrow yihX (+)$; $yihX \rightarrow ytiC (+)$; $yihX \rightarrow glk (+)$; $glk \rightarrow ytiC (+)$ ”.

In the stationary phase, the inferred gene regulatory networks were categorized into two types, (I) $fbaA \rightarrow pykA (+)$; $pykA \rightarrow galM (+)$; $galM \rightarrow glk (+)$ and (II) $ascF \rightarrow yihR (+)$; $yihR \rightarrow bglB (+)$; $bglB \rightarrow chbF (+)$, by enclosing the outermost groups (Figure 3B).

The (I) $fbaA \rightarrow pykA (+)$; $pykA \rightarrow galM (+)$; $galM \rightarrow glk (+)$ group contained three networks under RMSEA < 0.15 and two networks under RMSEA < 0.10. The group was additionally categorized into two groups: “ $pfkB \rightarrow galM (+)$; $pykA \rightarrow pfkB (+)$ ” and “ $fbaA \rightarrow galM (+)$; $fbaA \rightarrow agp (+)$ ”.

The (II) *ascF*→*yihR* (+); *yihR*→*bglB* (+); *bglB*→*chbF* (+) group encompassed a network under RMSEA < 0.15 and no network under RMSEA < 0.10.

3.4. Analysis of Gene Loci on a Genome

The loci of 33 genes used in this study were arranged based on genomic information obtained from the NCBI (Figure 4). Ten and six of genes showed relationships only in the logarithmic and stationary phases, respectively, and four of genes showed relationships in both phases, resulting in a total of twenty genes. No gene loci, which were close enough on the genome map, suggested the relationships displayed in Figure 3.

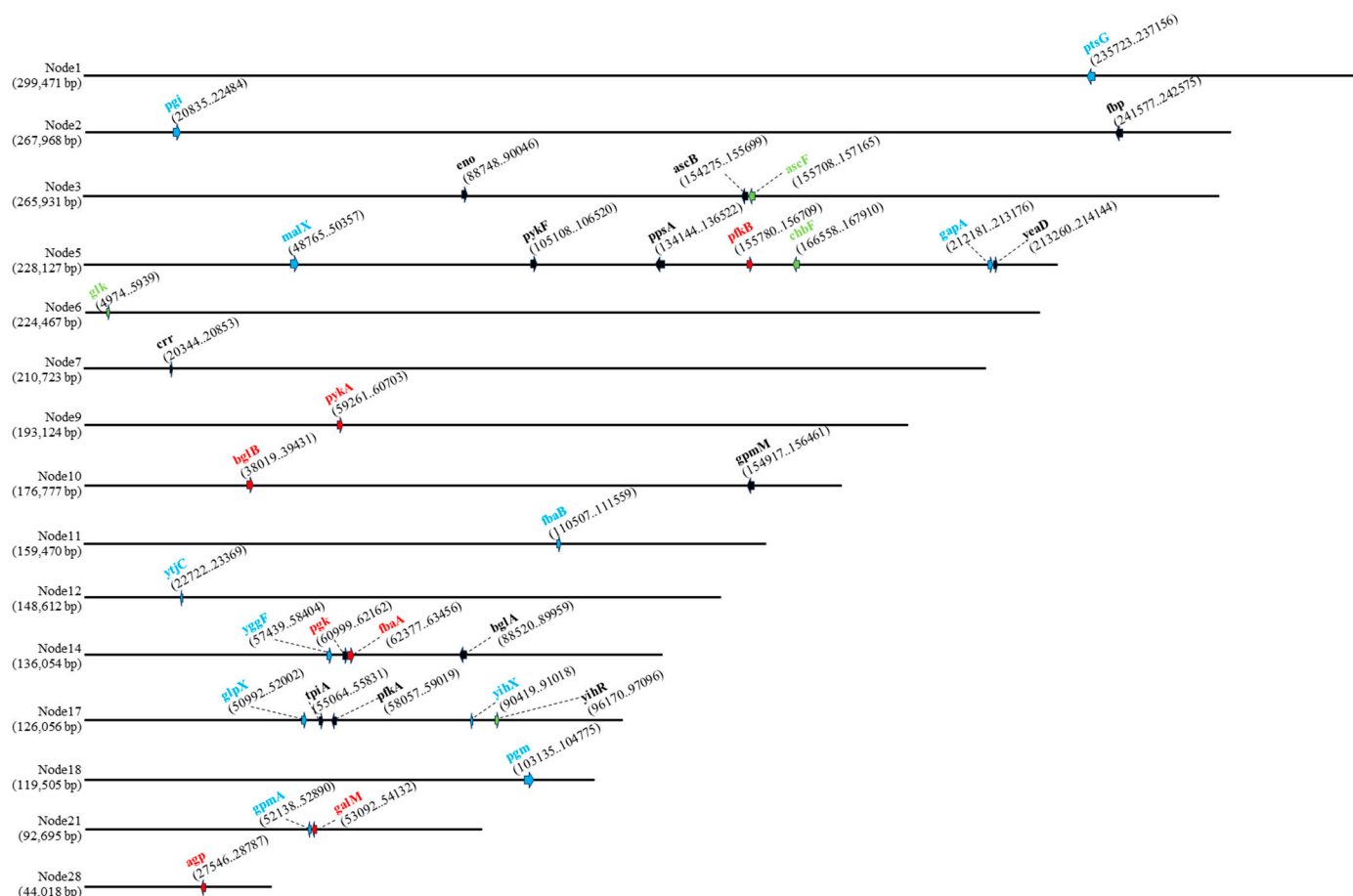


Figure 4. Each gene location on the genome in this study. The loci of genes analyzed in this study were placed on a map designed with data sources. Blue, red, green, and black arrows indicate loci of genes shown only in the logarithmic growth phase, stationary phase, both phases, and non-emerging phase, respectively. Numbers indicate the locus site numbers in the genome.

4. Discussion

The explanations for each transcriptional relationship have been used to easily follow the relationship on the known metabolic pathway map [30], and are prone to limit the scope of consideration. In this study, the ML model intentionally did not use the relationships on the known metabolic maps, namely meaning that the relationships among the gene transcription levels were not revised by the well-known information of the metabolic flow map when the ML model analyzed the strength of relationship. Relationships indicated by increased and/or decreased transcription levels of genes were evaluated using path analysis. The claims in this paper were based on “current knowledge of microbiology” that glycolysis is an essential metabolic system involved in energy production and material production in life, and assumes that “glycolysis is an important metabolic pathway and

therefore a regulatory relationship between genes certainly should exist even though it is difficult to detect". Furthermore, followed by the assumption, the claims in this paper were also based on "the certainty of the algorithm", and assume that "a regulatory network between genes can be derived". The reason why humans have difficulty detecting transcriptional regulation between genes was due to the need to comprehensively analyze large amounts of data, and so, to avoid this difficulty, we conducted a search based on a rational algorithm constructed by humans. In this report, we were only able to find relationships that are difficult for humans to perceive based on an algorithm shown by ML, so the results of this study cannot escape the framework of assumptions based on empirical methods. The analyses using an algorithm obtained by ML could suggest transcriptional regulation effectively. The dynamics of the transcription levels among genes were analyzed in the logarithmic growth phase and stationary phases, and the inferred gene regulatory networks were evaluated as shown below.

In order to perform a better resolution analysis in ML to obtain causal relationships, a dataset obtained from each condition was required that was at least eight to ten times the feature amount. In this case, for the analysis of the gene regulatory network, if the feature amount was a gene and each condition was the culture time, and samples were collected every 5 min during culture, a total of 264 datasets could be collected from 2 to 24 h of culture. By calculating backwards from this, it was determined that approximately 28 to 34 genes could be targeted as features. Considering this number of genes and the need to target fundamental metabolic pathways related to substance production in substance-producing *E. coli* strains such as *E. coli* LS5218, we concluded that the optimal number of genes would be approximately 33, which covers the entire glycolytic pathway. The inferred gene regulatory networks in logarithmic growth phase were divided into three type-groups of networks, (I) *yihR*→*yggF* (+), (II) *ptsG*→*pgm* (+), and (III) *yihX*→*glk* (+), by categorizing groups showing similarity (Figure 3A).

(I) *yihR*→*yggF* (+) was shown as an inferred gene regulatory network derived from the results of ML. *yihR* is an aldose epimerase and *yggF* is a fructose-1,6-bisphosphatase. Between fructose 6-phosphate and fructose 1,6-bisphosphate, the metabolic reaction is controlled not by the same enzyme which can convert in both directions but by the different enzymes which can, respectively, perform as one way of reversed reaction. *yggF*, *fbp*, and *glpX* encode hydrolysis enzymes converting fructose 1,6-bisphosphate to fructose 6-phosphate, while *pfkA* and *pfkB* encode phosphorylation enzymes reacting fructose 6-phosphate to fructose 1,6-bisphosphate. The one-way enzymes can control the metabolic flux into an upstream or downstream of glycolysis [31,32]. Although the relationship from *yihR* to *yggF* has been difficult to explain with current knowledge, if the results of the path analysis are reliable and there is indeed some relationship, the glucose isomerization reaction could be a factor that controls the direction of glycolysis in cells.

Additionally, specific relationship was shown as "*ascF*→*yihR* (+)" in this group (I). *ascF* codes for a beta-glucoside specific phosphotransferase system (PTS) enzyme IIBC component, and the protein *ascF* participates in transporting β -glucose into cells and phosphorylating the transported β -glucose. The *yihR* codes aldose epimerase converting α -glucose and β -glucose, which is similar to the metabolic reaction involving *ascF*. Therefore, the relationship between *ascF* and *yihR* could be acceptable with understanding, and the relationship possibly indicates that the cells respond to stabilize the structure of glucose to phosphorylate it. This relationship may be necessary for cells to metabolize glucose for growth because it was observed in the logarithmic growth phase. To date, there are no detailed reports on the effects of *ascF* on *yihR*. Moreover, the relationships of "*chbF*→*yihR* (+)" and "*gapA*→*pgm* (+)" emerged as relationships in the group of "*ascF*→*yihR* (+)". In the gene regulatory network of "*chbF*→*yihR* (+)", *yihR* could also

be positively controlled by *chbF*. *chbF* is a 6-phospho-beta-glucosidase, and its metabolism is also close to the reaction by *yihR* on the metabolic pathway map. *yihR* might be the important enzyme having a functional role for glucose metabolism that is much more than our consideration because the *yihR*-relationships were repeatedly indicated by other gene regulatory networks. In addition, the effect from *chbF* to *yihR* has not been reported, meaning the novel knowledge in this study. In the gene regulatory network of “*gapA*→*pgm* (+)”, *gapA*, which encodes an enzyme controlling the mid-glycolysis, showed a positive relationship to *pgm*, which encodes an enzyme reacting with the head-glycolysis. *gapA*, which functions at the entrance of the pay-off phase in glycolysis, is one of the key enzymes extremely important for glycolysis metabolism. *pgm* is an enzyme that exchanges the positions of the phosphate groups of glucose 6-phosphate and glucose 1-phosphate. Interestingly, *gapA* transcription positively controlled not only mid-glycolysis but also upstream glycolysis via *pgm* in the gene regulatory network of our study. To the best of our knowledge, how the transcription and expression of *gapA* directly affects that of *pgm* has not been reported.

(II) *ptsG*→*pgm* (+) were also shown as an inferred gene regulatory network derived from the results of ML. *ptsG* codes for a glucose-specific PTS enzyme IIBC component, which is responsible for the phosphoenolpyruvate-dependent sugar phosphotransferase system (sugar PTS) and phosphorylates sugars in cells. In this metabolic pathway, *ptsG* and *pgm* are enzymes that control adjacent reactions. Therefore, cells may convert glucose 6-phosphate into glucose 1-phosphate after phosphorylating glucose taken into cells to glucose 6-phosphate as a manner even in a non-glucose-containing medium.

By enclosing the outermost group in (II) *ptsG*→*pgm* (+), “*ptsG*→*fbaB* (+); *yihR*→*fbaB* (−)”, “*ptsG*→*yihX* (+); *yihX*→*ytjC* (+); *yihX*→*glk* (+); *glk*→*ytjC* (+)” and “*glpX*→*ptsG*; *gapA*→*pgi* (+)” were identified as groups of gene regulatory networks. Interestingly, in the gene regulatory network of “*ptsG*→*fbaB* (+); *yihR*→*fbaB* (−)”, *ptsG* positively regulated *fbaB*, while *yihR* negatively regulated *fbaB*, which appeared to be in an ambivalent relationship. *fbaB*, a fructose-bisphosphate aldolase class I, controls decomposing fructose 1,6-bisphosphate into dihydroxyacetone phosphate and glyceraldehyde 3-phosphate as a forward reaction and synthesizing dihydroxyacetone phosphate and glyceraldehyde 3-phosphate into fructose 1,6-bisphosphate as a reverse reaction, respectively. It was unclear why the transcription levels of enzyme genes involved in phosphorylation of glucose and conversion of α -glucose to β -glucose, such as *ptsG* and *yihR* upstream of glycolysis metabolism, competed with the level of the enzyme gene controlling metabolites such as *fbaB* in mid-glycolysis, and no literature has been published on this point. At the viewpoint of gene regulatory networks being dependent on our results, in “*ptsG*→*yihX* (+); *yihX*→*ytjC* (+); *yihX*→*glk* (+); *glk*→*ytjC* (+)”, transcriptions of enzyme-encoding genes such as *ptsG*, *yihX*, and *glk* were, respectively, connected with relation to dephosphorylation and phosphorylation in upstream of glycolysis, and possibly affected transcription of *ytjC* that encodes a phosphate group isomerase apart from the upstream. Enzymes from *ptsG* to *glk* via *yihR* function in phosphorylation and dephosphorylation on glucose; in contrast, *ytjC* catalyzes the phosphate intermolecular-transferring reaction between 3-phosphoglyceric acid and 2-phosphoglyceric acid as 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase. How the changes in the transcription levels of enzyme genes at the entrance of glycolysis may affect the level in the pay-off phase has not been reported as far as we know. Our results suggest the possibility that glycolysis is controlled in large frames rather than old-fashioned consideration. The relationships of “*glpX*→*ptsG*; *gapA*→*pgi* (+)” were also exhibited in group (II) as gene regulatory networks dependent on our results, interestingly, indicating that the enzyme genes of *glpX* and *gapA* in mid-glycolysis could regulate those of *ptsG* and *pgi* in the upstream. The control of glucose-flow may be regulated in mid-

glycolysis on the levels of gene transcription. Additionally, “*gapA*→*pgm* (+)” appeared to be an important factor because three of the networks were shown in four networks that exhibited RMSEA < 0.10 as a highly level definition. In this study, the relationships between transcriptional levels of genes in the glycolysis pathway, which are difficult to detect but strongly predicted to be related in life science, were analyzed based on the “certainty of the algorithm” of ML, and as a result, new transcriptional networks were derived. The reason why it is difficult for humans to detect transcriptional controls between genes is due to the difficulty of comprehensively analyzing large amounts of data, so the significance of this study was that it showed the gene regulatory networks based on a reasonable algorithm constructed by humans to avoid this difficulty.

(III) *yihX*→*glk* (+) emerged as an inferred gene regulatory network group derived from the results of ML. Regarding *yihX* as a dephosphoenzyme and *glk* as a phosphoenzyme relating the pathway from glucose 1-phosphate to glucose 6-phosphate via α -glucose, the *yihX* controlling positively to *glk* seemed to be a salvage of glucose 6-phosphate. If the reaction from glucose 6-phosphate to glucose 1-phosphate does not occur, the reaction from glucose-1-phosphate to glucose 6-phosphate via α -glucose is unnecessary to generate the flow into the main stream of glycolysis. The strains in this study were cultivated in LB medium so that glucose would not be taken up from outside the cells, indicating the possibility of high salvage as the recycle of carbon sources based on inferred gene regulatory network.

Moreover, in group (III), two relationships of “*yihX*→*malX* (+)” and “*ptsG*→*yihX* (+); *yihX*→*ytjC* (+); *yihX*→*glk* (+); *glk*→*ytjC* (+)” emerged by enclosing the outermost groups as the results of ML. In “*yihX*→*malX* (+)”, enhanced activity of *malX* may intend to increase the pool of glucose 6-phosphate. The enzyme *malX*, which is related to phosphorylation of glucose in the maltose/glucose PTS EIICB component, works synergistically with *glk*. If *yihX* does not function, there is no need to strengthen the function of *glk*; therefore, the role of *yihX* remains unclear. It could also be said that the evaluation using the ML model in this study was valuable in the point that raised doubt about the functionality of *yihX*. The same group-trends of “*ptsG*→*yihX* (+); *yihX*→*ytjC* (+); *yihX*→*glk* (+); *glk*→*ytjC* (+)” were observed in group (II), in which the transcription of *ptsG* may positively regulate that of *pgm*, as well as in the group in which the transcription of *yihX* controlled that of *glk*. Both *pgm* and *glk* are catalytic enzymes involved in glucose 6-phosphate so that those gene improvements looked like making the glucose-circuit. Interestingly, *ptsG* linked to the enzyme genes in upstream of glycolysis with relationships, and finally affected the gene of phosphate isomeric enzyme *ytjC* in mid of the pay-off phase. These results could indicate that glucose 6-phosphate plays an important role in the phosphorylation and control of metabolic flow in glycolysis.

The inferred gene regulatory networks in the stationary phase were divided into two type-groups of networks, (I) *fbaA*→*pykA* (+); *pykA*→*galM* (+); *galM*→*glk* (+) and (II) *ascF*→*yihR* (+); *yihR*→*bglB* (+); *bglB*→*chbF* (+), by enclosing similar groups (Figure 3B).

In the group (I) *fbaA*→*pykA* (+); *pykA*→*galM* (+); *galM*→*glk* (+), *fbaA* is a fructose-bisphosphate aldolase class II involved in the cleavage reaction of fructose-bisphosphate; *pykA* is a pyruvate kinase that performs the reaction between phosphoenolpyruvate and pyruvate; *galM* is a galactose mutarotase that transforms the α - and β -type of glucose; and *glk* is a glucokinase that phosphorylates glucose. In transcription, the inferred gene regulatory networks showed that *fbaA* in the preparatory phase of glycolysis was related to *pykA* in the endpoint of the pay-off phase of glycolysis, that *pykA* contacted *galM* to isomerize glucose before phosphorylation, and that *galM* influenced *glk* in phosphorylating glucose. Therefore, *fbaA* may be transcribed in anticipation of both the end and beginning

of glycolysis, indicating the possibility that controlling the transcription levels of *fbaA* could be a key to regulating the whole glycolysis in stationary phase.

In addition, the groups of “*pfkB*→*galM* (+); *pykA*→*pfkB* (+)” and “*fbaA*→*galM* (+); *fbaA*→*agp* (+)” showed similar gene regulatory networks as each other. In “*pfkB*→*galM* (+); *pykA*→*pfkB* (+)”, *pykA* positively controlled *galM* via *pfkB* as 6-phosphofructokinase 2; in “*fbaA*→*galM* (+); *fbaA*→*agp* (+)”, *fbaA* also regulated *galM* via *pykA*, and *fbaA* directly controlled *galM*. These relationships reinforce the possibility that *fbaA* transcription in the mid-glycolysis is a key factor to control both the end and beginning of glycolysis. Additionally, in “*fbaA*→*galM* (+); *fbaA*→*agp* (+)”, the gene transcription of *fbaA* positively controls that of *agp* as a glucose-1-phosphatase/inositol phosphatase according to the analytical results of ML. The enzyme *agp* could promote dephosphorylation of α -glucose 1-phosphate and be against the stream entering glucose into glycolysis like *yihX* shown in Figure 3A. The two genes of *yihX* and *agp* were possibly switched to control the dephosphorylation pathway in logarithmic growth phase and stationary phase.

In group “(II) *ascF*→*yihR* (+); *yihR*→*bglB* (+); *bglB*→*chbF* (+)”, the inferred gene regulatory networks showed that the relationships of the transcription levels of *ascF* and *yihR*, which were also observed in logarithmic phase, and the connection relied on those of *bglB* as 6-phospho-beta-glucosidase B and *chbF* as 6-phospho-beta-glucosidase. These causalities possibly indicate relationships between gene groups involved in sugar phosphorylation-glucose isomerization and those involved in the upstream pathway of the glycolysis system. Reactions in the pathway were directly involved with each other and are grouped together, and these results may mean the promotion that phosphorylated sugar sources were introduced into the glycolytic pathway. However, unlike in the logarithmic phase, the cascade did not attempt to affect the gene in the preparatory phase of the glycolysis system like *yggF*. As expected, the causality graph showed that the trends were different between logarithmic growth phase and stationary phase.

Many of the new regulatory relationships identified in this study may not reflect direct transcriptional regulation rather than indirect metabolic coupling. However, this report was significant in the suggestion that the regulatory relationships identified in this study may have a consequential effect, even if indirectly. This could be useful for future considerations when modifying the relevant genes through genetic recombination. The well-known functions of global transcription factors include the regulation of gene expression related to the catabolic pathway of cAMP receptor protein (CRP) [33] and the regulation of gene expression under anaerobic conditions in fumarate and nitrate reductase (FNR) [34]. While it is difficult to pinpoint specifics because of the condition-differences in the experiments, the analytical evaluation method of ML used in this study could suggest the possibility of investigating the functions of global transcription factors such as CRP and FNR by adjusting and comparing culture conditions in the future.

In bacteria, highly related genes exist as operons and are closely arranged on the genome in general. Of the 33 genes examined in this study, the gene pairs of *ascB-ascF*, *gapA-yeaD*, *pgk-fbaA*, and *gpmA-galM* are, respectively, adjacent on the genome; however, no loci of the 20 genes relating to the gene regulatory networks were close arranged on the genome (Figure 4). This result strongly suggests that the gene regulatory networks determined in this study were not regulated as the operons but were controlled independently according to cell requirements. The causal discovery method used in this study may be useful for evaluating gene transcription relationships that are difficult to clarify using conventional methods. In 2017, Fang et al. reconstructed and evaluated the global transcriptional regulatory network in *E. coli* based on existing information [35]. In addition, in 2005, Levine and Davidson used a similar approach to evaluate gene regulatory networks in sea urchin embryos and *Drosophila* dorsal-venter by integrating past literature

information [36]. Although they did evaluate gene regulatory networks during embryo development, their report was not based on a continuous and unified wet experimental system like ours, so there were concerns about some discrepancies. In addition, Liu et al. evaluated the hierarchical relationships between transcription factors, associated proteins, and their target genes in *E. coli* using a comprehensive regulatory network [37]. However, although experiments were conducted with different strains, the evaluation was limited to those where the OD₆₀₀ reached to 0.5–0.6, which was excellent for evaluating hierarchical structures, but is not appropriate for evaluating the sequential controls of transcription relationships during *E. coli*-cultivation. Although we have not absolutely conducted a hierarchical analysis, we have focused on the gene transcription level and performed a very detailed analysis that has never been seen before, resulting in the reports with high resolution. Our report is, academically, a very significant result in the evaluation of responses during cultivation, and we believe that the results of this research will provide information that is fully useful for controlling transcription levels through genetic engineering.

5. Conclusions

The ML-based gene regulatory network inference of the levels of comprehensive gene transcription demonstrated in this study may be useful for improving the production of targeted substances by modifying gene transcription levels through genetic engineering. In applied microbiology, to improve the production of target substances, metabolic flow related to production can be improved by controlling gene transcription based on the pathway map with genetic engineering. Control of transcription in industrial microorganisms leads to high production of targeted substances depending on the enhanced metabolic flows with optimized gene transcription; however, unfavorable control can decrease production when the flow is stifled through inadequate transcription. These processes can be affected by low biomass productivity, activation of other metabolic flows, and suppressed expression of genes other than the target gene. Although critical issues regarding unsuitable transcription of well-known gene relationships on the metabolic map can be resolved, unknown gene relationships cannot be addressed. The ML model, which inferred the relationships among gene transcription levels in this study, evaluated those with increased and/or decreased values of relationship. This function showed its possibility to be a useful tool for predicting comprehensive gene-level optimizations related to the metabolic flow of material production. Previously, the creation of recombinant strains was required to determine the relationships among gene transcription levels, which is time-consuming. However, strain creation could be unrequired to evaluate the relationships when using the ML model in the case that ML could comprehensively explore changes in gene transcription levels from multiple angles with more progress based on our research in near future. We demonstrated the possibility of using information on the relationships among gene transcriptions suggested by the ML model to assist genetic engineering. It may be possible to use ML models for accurate analysis at low cost in a rapid manner in the industry by expanding omics information.

As the use of ML becomes more widespread, the risk of taking ML results at face value increases. However, information yielded by ML models are not always true and must be carefully considered. Therefore, it is important for life science researchers to consider the results and make final decisions. In a causal discovery method, multiple top candidate gene regulatory networks were inferred based on the fit indices. These types of analyses will enable life science researchers, for whom experimental validation is difficult, to conduct more in-depth studies. To the best of our knowledge, a causal discovery method has not been utilized to provide insights into gene regulatory networks at a level comparable to experimental validation. We demonstrated a new form of collaboration between ML

models and life sciences, in which the quality of consideration is greatly improved by presenting multiple candidates to life science researchers.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/bacteria4040060/s1>, Table S1: Information of genes and primer-pairs; Supplementary S1: DAG structure learning algorithm based on NOTEARS using the causalnex.structure.pytorch submodule; Supplementary S2: Theoretical framework of path analysis.

Author Contributions: A.N.: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing—original draft, Writing—review and editing, Visualization, Supervision, Project administration, Funding acquisition. N.O.: Validation, Formal analysis, Investigation, Data curation, Writing—original draft. R.O.: Formal analysis, Investigation, Software. H.K.: Investigation, Software. H.F.: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing—original draft, Writing—review and editing, Visualization, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Tokyo University of Technology Special Grant-in-Aid for joint project as research funding (funding data: 1 August 2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supporting File.

Acknowledgments: The authors deeply appreciate Riri Matsumoto and Sanshiro Yarita for their kind help.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaper, J.B.; Nataro, J.P.; Mobley, H.L. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2004**, *2*, 123–140. [CrossRef]
2. Cordell, W.T.; Avolio, G.; Takors, R.; Pfleger, B.F. Milligrams to kilograms: Making microbes work at scale. *Trends Biotechnol.* **2023**, *41*, 1442–1457. [CrossRef]
3. Rand, J.M.; Gordon, G.C.; Mehrer, C.R.; Pfleger, B.F. Genome sequence and analysis of *Escherichia coli* production strain LS5218. *Metab. Eng. Commun.* **2017**, *5*, 78–83. [CrossRef]
4. Browning, D.F.; Hobman, J.L.; Busby, S.J.W. Laboratory strains of *Escherichia coli* K-12: Things are seldom what they seem. *Microb. Genom.* **2023**, *9*, mgen000922. [CrossRef] [PubMed]
5. Cornelis, P. Expressing genes in different *Escherichia coli* compartments. *Curr. Opin. Biotechnol.* **2000**, *11*, 450–454. [CrossRef] [PubMed]
6. Fujita, Y.; Matsuoka, H.; Hirooka, K. Regulation of fatty acid metabolism in bacteria. *Mol. Microbiol.* **2007**, *66*, 829–839. [CrossRef]
7. Agnew, D.E.; Stevermer, A.K.; Youngquist, J.T.; Pfleger, B.F. Engineering *Escherichia coli* for production of C 12—C 14 polyhydroxyalkanoate from glucose. *Metab. Eng.* **2012**, *14*, 705–713. [CrossRef] [PubMed]
8. Rhie, H.G.; Dennis, D. Role of *fadR* and *atoC*(con) mutations in poly(3-hydroxybutyrate-co-3-hydroxyvalerate) synthesis in recombinant *pha+* *Escherichia coli*. *Appl. Environ. Microbiol.* **1995**, *61*, 2487–2492. [CrossRef]
9. Lioliou, E.E.; Mimitou, E.P.; Grigoroudis, A.I.; Panagiotidis, C.H.; Panagiotidis, C.A.; Kyriakidis, D.A. Phosphorylation activity of the response regulator of the two-component signal transduction system AtoS-AtoC in *E. coli*. *Biochim. Biophys. Acta.* **2005**, *1725*, 257–268. [CrossRef]
10. Theodorou, E.C.; Theodorou, M.C.; Samali, M.N.; Kyriakidis, D.A. Activation of the AtoSC two-component system in the absence of the AtoC N-terminal receiver domain in *E. coli*. *Amino Acids.* **2011**, *40*, 421–430. [CrossRef]
11. Jenkins, L.S.; Nunn, W.D. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in *Escherichia coli*: The *ato* system. *J. Bacteriol.* **1987**, *169*, 42–52. [CrossRef]
12. Matta, M.K.; Lioliou, E.E.; Panagiotidis, C.H.; Kyriakidis, D.A.; Panagiotidis, C.A. Interactions of the antizyme AtoC with regulatory elements of the *Escherichia coli* *atoDAEB* operon. *J. Bacteriol.* **2007**, *189*, 6324–6332. [CrossRef]
13. Tappel, R.C.; Wang, Q.; Nomura, C.T. Precise control of repeating unit composition in biodegradable poly(3-hydroxyalkanoate) polymers synthesized by *Escherichia coli*. *J. Biosci. Bioeng.* **2012**, *113*, 480–486. [CrossRef]

14. Goto, S.; Hokamura, A.; Shiratsuchi, H.; Taguchi, S.; Matsumoto, K.; Abe, H.; Tanaka, K.; Matsusaki, H. Biosynthesis of novel lactate-based polymers containing medium-chain-length 3-hydroxyalkanoates by recombinant *Escherichia coli* strains from glucose. *J. Biosci. Bioeng.* **2019**, *128*, 191–197. [CrossRef]
15. Wang, Q.; Tappel, R.C.; Zhu, C.J.; Nomura, C.T. Development of a new strategy for production of medium-chain-length polyhydroxyalkanoates by recombinant *Escherichia coli* via inexpensive non-fatty acid feedstocks. *Appl. Environ. Microbiol.* **2012**, *78*, 519–527. [CrossRef]
16. Jiao, J.; Lv, X.; Shen, C.; Morigen, M. Genome and transcriptomic analysis of the adaptation of *Escherichia coli* to environmental stresses. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2132–2140. [CrossRef] [PubMed]
17. Long, C.P.; Antoniewicz, M.R. Metabolic flux analysis of *Escherichia coli* knockouts: Lessons from the Keio collection and future outlook. *Curr. Opin. Biotechnol.* **2014**, *28*, 127–133. [CrossRef] [PubMed]
18. Li, P.; Zhang, C.; Perkins, E.J.; Gong, P.; Deng, Y. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinform.* **2007**, *8*, S13. [CrossRef]
19. Moeckel, C.; Mouratidis, I.; Chantzi, N.; Uzun, Y.; Georgakopoulos-Soares, I. Advances in computational and experimental approaches for deciphering transcriptional regulatory networks: Understanding the roles of cis-regulatory elements is essential, and recent research utilizing MPRA, STARR-seq, CRISPR-Cas9, and machine learning has yielded valuable insights. *Bioessays* **2024**, *46*, e2300210.
20. Chai, L.E.; Loh, S.K.; Low, S.T.; Mohamad, M.S.; Deris, S.; Zakaria, Z. A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* **2014**, *48*, 55–65. [CrossRef] [PubMed]
21. Zheng, X.; Aragam, B.; Ravikumar, P.K.; Xing, E.P. DAGs with NO TEARS: Continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9472–9483.
22. Zhu, Y.; Benos, P.V.; Chikina, M. A hybrid constrained continuous optimization approach for optimal causal discovery from biological data. *Bioinformatics* **2024**, *40* (Suppl. 2), ii87–ii97. [CrossRef]
23. Spratt, S.K.; Ginsburgh, C.L.; Nunn, W.D. Isolation and genetic characterization of *Escherichia coli* mutants defective in propionate metabolism. *J. Bacteriol.* **1981**, *146*, 1166–1169. [CrossRef]
24. Livak, K.J.; Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *Methods* **2001**, *25*, 402–408. [CrossRef]
25. McKinsey & Company. CausalNex, GitHub Repository. Available online: <https://github.com/mckinsey/causalnex> (accessed on 31 January 2025).
26. Georgy, M. Semopy, GitLab Repository. Available online: <https://gitlab.com/georgy.m/semopy> (accessed on 31 January 2025).
27. Koçoglu, E.; Tekdal, D.; Çetinkaya, N. Pedagogical literacy scale: A scale development study. *Educ. Res. Rev.* **2002**, *17*, 176–186. [CrossRef]
28. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
29. Ozaki, Y.; Tanigaki, Y.; Watanabe, S.; Onishi, M. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference, New York, NY, USA, 8–12 July 2020; pp. 533–541.
30. Yang, D.; Park, S.Y.; Park, Y.S.; Eun, H.; Lee, S.Y. Metabolic engineering of *Escherichia coli* for natural product biosynthesis. *Trends. Biotechnol.* **2020**, *38*, 745–765. [CrossRef] [PubMed]
31. Cabrera, R.; Baez, M.; Pereira, H.M.; Caniuguir, A.; Garratt, R.C.; Babul, J. The crystal complex of phosphofructokinase-2 of *Escherichia coli* with fructose-6-phosphate. *J. Biol. Chem.* **2011**, *286*, 5774–5783. [CrossRef] [PubMed]
32. Torres, J.C.; Guixé, V.; Babul, J. A mutant phosphofructokinase produces a futile cycle during gluconeogenesis in *Escherichia coli*. *Biochem. J.* **1997**, *327*, 675–684. [CrossRef] [PubMed]
33. Lauritsen, I.; Frendorf, P.O.; Capucci, S.; Heyde, S.A.H.; Blomquist, S.D.; Wendel, S.; Fischer, E.C.; Sekowska, A.; Danchin, A.; Nørholm, M.H.H. Temporal evolution of master regulator Crp identifies pyrimidines as catabolite modulator factors. *Nat. Commun.* **2021**, *12*, 5880. [CrossRef]
34. Myers, K.S.; Yan, H.; Ong, I.M.; Chung, D.; Liang, K.; Tran, F.; Keleş, S.; Landick, R.; Kiley, P.J. Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. *PLoS Genet.* **2013**, *9*, e1003565. [CrossRef]
35. Fang, X.; Sastry, A.; Mih, N.; Kim, D.; Tan, J.; Yurkovich, J.T.; Lloyd, C.J.; Gao, Y.; Yang, L.; Palsson, B.O. Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 10286–10291. [CrossRef] [PubMed]

-
36. Levine, M.; Davidson, E.H. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4936–4942. [[CrossRef](#)] [[PubMed](#)]
 37. Liu, R.; Liang, L.; Freed, E.F.; Choudhury, A.; Eckert, C.A.; Gill, R.T. Engineering regulatory networks for complex phenotypes in *E. coli*. *Nat. Commun.* **2020**, *11*, 4050. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.