Proteome Allocation Based Metabolic Modelling of Lactic Acid Bacteria Under Environmental Stress

Sizhe Qiu

Brasenose College

Department of Engineering Science

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Abstract

The response to environmental stress shapes the growth and metabolism of microorganisms under changing environmental conditions, and stress-induced secondary metabolism is a rich source of valuable bioproducts in various industries. Consequently, it's necessary to incorporate environmental stress into the modeling of microbial metabolism. However, due to the inherent limitations of conventional flux balance analysis (FBA), most existing FBA-based models cannot account for stress responses. Some researchers have proposed proteome resource allocation models incorporating the effect of environmental stress, but those theoretical models have not yet been implemented in any real-world scenarios with specific microbial systems.

With the aim to develop computational models for microbial metabolism under environmental stress, four research projects were conducted in this PhD study to analyze and model the effects of undissociated lactic acid, temperature changes, and acidic pH on microbial metabolism. The selected research object was lactic acid bacteria. Project 1 (**Chapter 3**) and Project 2 (**Chapter 4**) both focused on growth-coupled primary metabolism under stress. Project 1 built a dynamic community FBA model, integrated with constrained proteome allocation, for dominant species in the yogurt starter culture to accurately simulate the central carbon metabolism under the stress of accumulating undissociated lactic acid. Furthermore, the model could account for inter-species metabolic interactions. Subsequently, Project 2 developed a deep learning based predictor of temperature dependent enzyme k_{cat}, DLTKcat, to model microbial metabolism affected by temperature changes. The temperature sensitive proteome constrained FBA, performed with predicted k_{cat}, could capture the metabolic responses of lactic acid bacteria to temperature changes, though the quantitative accuracy remained to be improved.

Regarding stress-induced secondary metabolism, Project 3 (**Chapter 5**) revealed a trade-off in regulatory activities between primary metabolism and acid stress-induced exopolysaccharide (EPS) biosynthesis in *Lactiplantibacillus plantarum* from the independent component analysis of transcriptomic data. To quantitatively investigate acid stress-induced EPS production in *L. plantarum*, Project 4 (**Chapter 6**), with multi-omics data at different pH values, identified a proteome trade-off between primary metabolism and EPS biosynthesis, which led to the construction of regulatory proteome constrained flux balance analysis (RPCFBA). As the first mechanistic model that can simulate primary and secondary metabolism simultaneously, RPCFBA model showed good accuracy in predicting growth rates and EPS production fluxes of *L. plantarum* in the validation with experimental data.

Overall, FBA-based metabolic models developed in this PhD study demonstrated that modified proteome constrained FBA could overcome the limitations of conventional FBA on microbial metabolism under environmental stress. Moreover, such models have the potential to become computational tools to aid the control and engineering of complex microbial processes.

Acknowledgements

The PhD study in University of Oxford has been a long and fruitful journey, and the completion of my research has received help from many people. First of all, my supervisor, Prof. Aidong Yang, and my main collaborator, Prof. Hong Zeng, mentored and advised me throughout my PhD study. Their guidance was invaluable, and my research would have been directionless without their help. Also, I want to express my sincere gratitude to my other collaborators: Prof. Yanbo Wang, Prof. Bei Wang, Prof. Li Huang, Zhijie Yang, Xinyu Yang, Simiao Zhao, and Shishun Liang. Their support and contributions were important to my research projects.

Regarding the financial support, my parents were the main funders of my life in Oxford, and Brasenose College offered the Senior Hulme Scholarship to me. Additionally, the spiritual support from my family and friends (e.g., Lin Ban) was equally important.

In the end, I want to thank myself for choosing to do PhD study in University of Oxford and persisting through three years of numerous challenging tasks, which unveiled a new world to me. ***Descende, audax viator, et terrestre centrum attinges.***

Publications

1. Qiu S, Yang A, Yang X, Li W, Zeng H, Wang Y. Proteome trade-off between primary and secondary metabolism shapes acid stress induced bacterial exopolysaccharide production. bioRxiv. 2024. p. 2024.04.19.590233. doi:10.1101/2024.04.19.590233 (Preprint)
2. Qiu S, Huang Y, Liang S, Zeng H, Yang A. Systematic elucidation of independently modulated genes in Lactiplantibacillus plantarum reveals a trade-off between secondary and primary metabolism. Microb Biotechnol. 2024;17: e14425.
3. Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. Brief Bioinform. 2023;25. doi:10.1093/bib/bbad506
4. Qiu S, Zeng H, Yang Z, Hung W-L, Wang B, Yang A. Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. Biotechnol Bioeng. 2023;120: 2186–2198.
5. Qiu S, Yang Z, Zeng H, Wang B, Yang A. Dynamic microbial-community metabolic modeling for yogurt fermentation based on the metagenome of starter culture. In: Kokossis AC, Georgiadis MC, Pistikopoulos E, editors. Computer Aided Chemical Engineering. Elsevier; 2023. pp. 2619–2624.
6. Qiu S, Yang A, Zeng H. Flux balance analysis-based metabolic modeling of microbial secondary metabolism: Current status and outlook. PLoS Comput Biol. 2023;19: e1011391.

Abbreviations

ABC: ATP-binding cassette

API: application programming interface

BGC: biosynthetic gene cluster

CDM: chemically defined medium

CFU: colony forming unit

CHiP-seq: chromatin immunoprecipitation followed by sequencing

CNN: convolutional neural network

CPI: compound protein interaction

CPS: capsular polysaccharide

dFBA: dynamic flux balance analysis

DW: dry weight

ECFP: Extended Connectivity Fingerprint

EPS: exopolysaccharide

FBA: flux balance analysis

GAT: graph attention network

GC: gas chromatography

GNN: graph neural network

GSMM: genome-scale metabolic model

GT: glycosyltransferase

ICA: independent component analysis

IM/iModulon: a set of independently modulated genes

LAB: lactic acid bacteria

LB: *Lactobacillus delbrueckii subsp. bulgaricus*

LC: liquid chromatography

LFC: log2 fold change

LL: *Lactococcus lactis*

LP: *Lactiplantibacillus plantarum*

Leaky ReLU: leaky Rectified Linear Unit

MAE: mean absolute error

MAG: metagenome assembled genome

MRS: De Man, Rogosa and Sharpe

MS: mass spectrometry

MSE: mean squared error

MUT: mutant

PCC: Pearson correlation coefficient

PCA: principal component analysis

R2: r-squared, the coefficient of determination

RMSE: root mean squared error

RNN: recurrent neural network

RPCFBA: regulatory proteome constrained flux balance analysis

RSM: response surface modeling

SMILES: simplified molecular-input line-entry system

ST: *Streptococcus thermophilus*

TF: transcriptional factor

TRN: transcriptional regulatory network

WT: wild type

Data availability statement

The code and data are openly available at:

1. Chapter 3: https://github.com/SizheQiu/MetaStLbCom.
2. Chapter 4: https://github.com/SizheQiu/DLTKcat.
3. Chapter 5: https://github.com/SizheQiu/LPiModulons.
4. Chapter 6: https://github.com/SizheQiu/LbPtEPS.

Table of Contents

Chapter 1 Introduction

1.1 Background

Microbial cellular metabolism requires enzyme proteins to catalyze most reactions to maintain the life activities of the cell (e.g., energy generation, protein translation, etc.), making proteins the most important cellular resources. The theoretical model of proteome resource allocation has been widely used to explain and model cellular growth and metabolism, since the interdependence between cellular growth and gene expression was discovered as a microbial growth law [1]. The proteome allocation theory models the competition for the limited cellular proteome resource among different proteome sectors (e.g., catabolism, anabolism), and thus characterizes the coordination of cellular pathways under different growth conditions [2]. For example, the proteome resource will be reallocated from ribosomal proteins to carbohydrate transport and energy metabolism at slow growth due to carbon starvation, whereas increased allocation to ribosomal proteins occurs at fast growth [3]. In contrast to the complex regulatory network, the proteome allocation principle provides a simplistic and generic model to explain the relationships of different cellular pathways by setting a global constraint on them.

With respect to modeling microbial metabolism, flux balance analysis (FBA) with genome-scale metabolic models (GSMMs) is the most commonly used method due to its low computational cost and relatively small requirement of data, in contrast to differential-equation-based models [4–6]. FBA uses linear optimization to compute metabolic fluxes with an objective function (usually the maximization of the cellular growth rate) [7] (**Figure 1.1**). As the distribution of proteome resources in different pathways governs cellular metabolism by constraining metabolic capacities of enzyme catalyzed reactions, the proteome allocation theory can be integrated into conventional FBA (loosely constrained by mass conservation) to tighten the

solution space of metabolic fluxes with proteome costs of reactions (formulated based on enzyme activities) and thus enhance the accuracy of simulation [8]. In coarse-grained proteome constrained FBA models, there are various case-specific methods to divide the flexible proteome into different sectors: for example, fermentation, respiration, and BM-sector (the remaining proteins for other cellular activities) [9], or C-sector (carbon intake and transport), E-sector (biosynthetic enzymes), and R-sector (ribosome-affiliated proteins) [8]. In recent years, proteome constrained FBA has been used to quantitatively model phenomena that cannot be captured by conventional FBA, such as the overflow metabolism in *E. coli* [8–10], and the preference of lactic acid production over acetic acid production in lactic acid bacteria [11].

In spite of successful applications on many scenarios as discussed above, most existing proteome allocation based metabolic models did not consider the effect of environmental stress (e.g., temperature changes [12]) and stress induced secondary metabolism (e.g, high population density induced antibiotics production [13]), which are important for the quantitative understanding of microbial physiology and precise modeling and control of microbial processes. At the molecular level, environmental stress such as acidic pH or high temperature affect enzyme activities, and thus cast an impact on the microbial growth and metabolism [14,15]. Sometimes, environmental stress can trigger secondary metabolism, e.g. the elevated temperature can stimulate the production of jadomycin B in *Streptomyces venezuelae* [16,17]. Furthermore, the response to environmental stress is a significant factor in industrial microbial processes such as the fermentation by lactic acid bacteria. For instance, the accumulation of undissociated lactic acid produced by lactic acid bacteria inhibits the bacterial growth in cheese fermentation [18], and the acidic pH can induce the biosynthesis of EPS in *L. plantarum* [19]. Consequently, the mechanistic modeling of microbial metabolism in response to environmental stress is required to better analyze microbial phenotypes under various conditions and control microbial processes in real-world scenarios.

Currently, there exist several mechanistic models proposed to characterize stress responses. For example, Özcan et al., 2021 used a feed-back inhibition function on upper bounds of carbon and nitrogen source uptake fluxes to accurately model the effect of undissociated lactic acid [18], though the model was not based on constrained proteome allocation. Regarding proteome resource allocation, Synthetic Chemostat Model (SCM) [20] and Grime's competitor-stress-ruderal triangle model (C-S-R) [21][22] suggested that the proteome resource would be allocated from primary metabolism to a proteome sector defined for stress resistance. Although neither SCM nor C-S-R has been implemented in the form of proteome allocation based metabolic modeling, they provided potentially applicable directions to enable the modeling of microbial growth and metabolism in response to environmental stress.

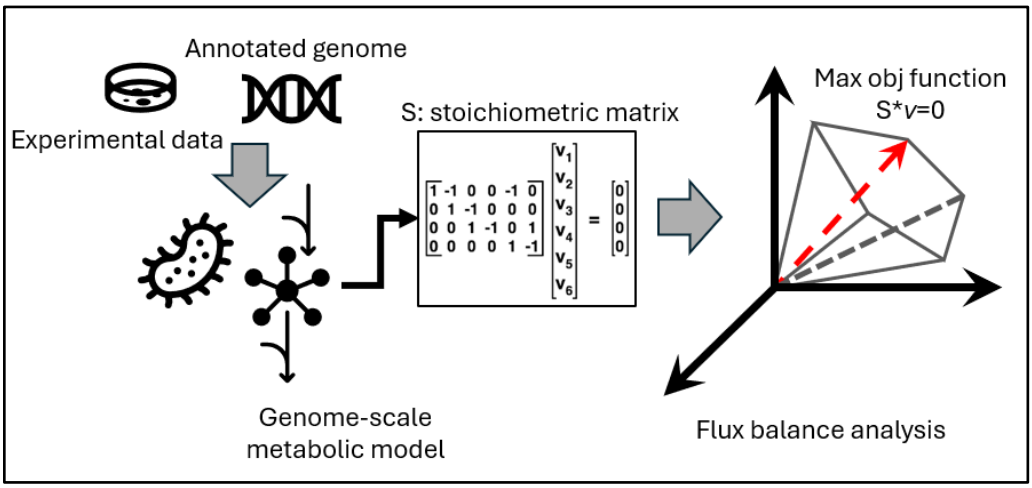

Figure 1.1. The workflow of FBA-based metabolic modeling. (1) Genome-scale metabolic model reconstruction from the annotated genome and experimental data (e.g., biomass composition). (2) Optimization of the objective function with the mass conservation constraint ($S * v = 0$, S: the stoichiometric matrix, v: metabolic reaction fluxes).

1.2 Motivation and objectives

Notwithstanding the importance of stress responses in microbial processes, most proteome constrained FBA models cannot account for the effect of environmental stress and

stress induced secondary metabolism (**section 1.1**). The modeling framework of existing proteome constrained FBA models shared the limitations of conventional FBA: (1) the regulatory activities upstream to cellular metabolism are not mechanistically represented; (2) kinetic parameters are not available for some essential metabolic enzymes under different environmental conditions; (3) two major objectives of microbial life activity, growth and stress resistance, cannot be captured simultaneously; and (4) regarding stress-induced secondary metabolism, the genome-scale pathway reconstruction is primitive for some secondary metabolites (e.g., EPS). With an aim to address those existing issues, this PhD study explored the incorporation of the effect of environmental stress into the proteome constrained FBA for lactic acid bacteria (e.g., *L. plantarum*). In addition, this study also intended to provide insights, through quantitative analysis of multi-omics data, on the microbial growth strategy to balance growth-coupled primary metabolism and stress-induced secondary metabolism.

More specifically, this PhD study comprises four individual research projects (**Figure 1.2**), and their objectives are enumerated as follows:

1. Project 1 attempted to construct a dynamic proteome constrained community-level FBA model for yogurt fermentation, and model the product inhibition of undissociated lactic acid and inter-species metabolic interactions.
2. Project 2 endeavored to develop a deep learning-based predictor of temperature dependent enzyme k_{cat} and model the effect of temperature changes on microbial growth and metabolism using temperature sensitive proteome constrained FBA with predicted enzyme k_{cat}.
3. Project 3 aimed to discover the relationship between growth-coupled primary metabolism and acid stress response in *L. plantarum* through inferring the transcriptional regulatory network (TRN) via independent component analysis of transcriptomic data.
4. Project 4 intended to investigate the relationship between primary metabolism and acid stress induced EPS biosynthesis in *L. plantarum* HMX2 through multi-omics analysis,

and construct a regulatory proteome constrained FBA (RPCFBA) model to simulate primary and secondary metabolism simultaneously.

1.3 Thesis structure

This thesis is presented in the integrated format and consists of seven chapters in total, including introduction (Chapter 1), literature review (Chapter 2), four research chapters (Chapter 3-6) and conclusions (Chapter 7). Each research chapter is a published peer-reviewed journal paper or a manuscript under review. For consistency, the original sections, figures, tables and equations in the published paper/manuscript are re-numbered in this thesis. The contents of 7 chapters are outlined as follows:

1. Chapter 1 illustrates the background, motivation and aims of research projects in this PhD thesis.
2. Chapter 2 critically reviews existing flux balance analysis-based modeling methods of microbial metabolism under environmental stress, and provides suggestions that can potentially improve the predictive power.
3. Chapter 3 describes the construction of a dynamic metagenome-scale metabolic model for two dominant species in the yogurt bacterial community, which quantitatively captures the metabolic dependence between two species and the inhibitory effect of undissociated lactic acid on bacterial growth.
4. Chapter 4 presents a deep learning-based predictor of temperature dependent enzyme k_{cat}, DLTKcat, and examines its application on modeling the response of bacterial metabolism to temperature changes.
5. Chapter 5 details the derivation of iModulons of *L. plantarum* through independent component analysis of its transcriptomics data, and how the analysis of iModulons reveals a trade-off between regulatory activities of primary and secondary metabolism.

6. Chapter 6 elaborates the identification of a proteome trade-off between primary metabolism and acid stress induced EPS biosynthesis in *L. plantarum* HMX2 via multi-omics analysis, and the construction of a regulatory proteome constrained FBA model that can accurately simulate growth rates and EPS production fluxes.
7. Chapter 7 summarizes the conclusions of this PhD thesis and recapitulates the contributions of four research projects. The limitations of this thesis are also discussed, along with potential directions for future research.

The relationships of research projects are graphically abstracted in **Figure 1.2.** Project 1 provided Projects 2 and 4 the mathematical formulation of constrained proteome allocation specialized for lactic acid bacteria. The deep learning-based predictor developed in Project 2 was used in Project 4 to estimate missing enzyme k_{cat} values. The transcriptomic trade-off between primary and secondary metabolism identified in Project 3 laid the groundwork for Project 4 to further investigate the acid stress induced proteome allocation in *L. plantarum*. Overall, four research projects attempted to model and analyze lactic acid bacterial metabolism under four major types of environmental stress, i.e., nutrient limitation, product inhibition, temperature, and low pH.

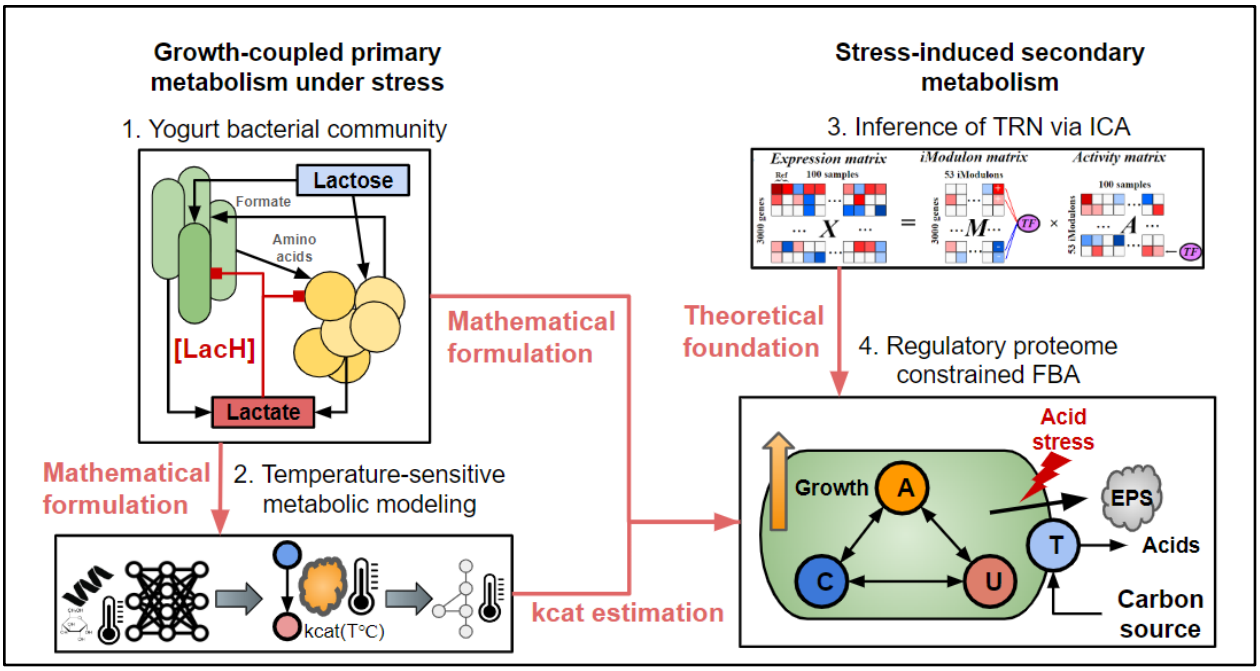

Figure 1.2 Graphical abstract of 4 research projects in this PhD study. (1) Project 1 (**Chapter 3**): Dynamic community-level metabolic model of the yogurt bacterial community. LacH: undissociated lactic acid. (2) Project 2 (**Chapter 4**): Deep learning-based predictor of temperature dependent enzyme k_{cat}. (3) Project 3 (**Chapter 5**): Inference of the TRN of *L. plantarum* using independent component analysis. TRN: transcriptional regulatory network. (4) Project 4 (**Chapter 6**): The construction of the regulatory proteome constrained FBA model to simulate acid stress induced EPS production in *L. plantarum* HMX2. EPS: exopolysaccharide; A: anabolism; C: catabolism; T: transporter proteins; U: secondary metabolism (EPS biosynthesis).

References

1. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330: 1099–1102.

2. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Mol

Syst Biol. 2015;11: 784.

3. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R. Proteome reallocation in Escherichia coli with increasing specific growth rate. Mol Biosyst. 2015;11: 1184–1193.

4. Nilsson A, Nielsen J, Palsson BO. Metabolic Models of Protein Allocation Call for the Kinetome. Cell Syst. 2017;5: 538–541.

5. Foster CJ, Wang L, Dinh HV, Suthers PF, Maranas CD. Building kinetic models for metabolic engineering. Curr Opin Biotechnol. 2021;67: 35–41.

6. Damiani C, Di Filippo M, Pescini D, Maspero D, Colombo R, Mauri G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. Bioinformatics. 2017;33: i311–i318.

7. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248.

8. Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained Allocation Flux Balance Analysis. PLoS Comput Biol. 2016;12: e1004913.

9. Zeng H, Yang A. Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways. BMC Syst Biol. 2019;13: 3.

10. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, et al. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 2015;528: 99–104.

11. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM, Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758.

doi:10.1002/bit.27605

12. Wendering P, Nikoloski Z. Model-driven insights into the effects of temperature on metabolism. Biotechnol Adv. 2023;67: 108203.

13. Liu X, Bimerew M, Ma Y, Müller H, Ovadis M, Eberl L, et al. Quorum-sensing signaling is required for production of the antibiotic pyrrolnitrin in a rhizospheric biocontrol strain of Serratia plymuthica. FEMS Microbiol Lett. 2007;270: 299–305.

14. Bendig T, Ulmer A, Luzia L, Müller S, Sahle S, Bergmann FT, et al. The pH-dependent lactose metabolism of Lactobacillus delbrueckii subsp. bulgaricus: An integrative view through a mechanistic computational model. J Biotechnol. 2023;374: 90–100.

15. Li G, Hu Y, Jan Zrimec, Luo H, Wang H, Zelezniak A, et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. Nat Commun. 2021;12: 190.

16. Lind AL, Smith TD, Saterlee T, Calvo AM, Rokas A. Regulation of Secondary Metabolism by the Velvet Complex Is Temperature-Responsive in Aspergillus. G3 . 2016;6: 4023–4033.

17. Yoon V, Nodwell JR. Activating secondary metabolism with stress and chemicals. J Ind Microbiol Biotechnol. 2014;41: 415–424.

18. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. Biotechnol Bioeng. 2021;118: 223–237.

19. Nguyen P-T, Nguyen T-T, Vo T-N-T, Nguyen T-T-X, Hoang Q-K, Nguyen H-T. Response of Lactobacillus plantarum VAL6 to challenges of pH and sodium chloride

stresses. Sci Rep. 2021;11: 1301.

20. Panikov NS. Genome-Scale Reconstruction of Microbial Dynamic Phenotype: Successes and Challenges. Microorganisms. 2021;9. doi:10.3390/microorganisms9112352

21. Grime JP. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory. Am Nat. 1977;111: 1169–1194.

22. Bruggeman FJ, Teusink B, Steuer R. Trade-offs between the instantaneous growth rate and long-term fitness: Consequences for microbial physiology and predictive computational models. Bioessays. 2023;45: e2300015.

Chapter 2 Literature review: Use flux balance analysis to model microbial metabolism under environmental stress

Rewritten based on Qiu, S., Yang, A., Zeng, H. (2023). Flux balance analysis-based metabolic modeling of Microbial Secondary Metabolism: Current status and outlook. PLOS Computational Biology, 19(8). https://doi.org/10.1371/journal.pcbi.1011391

Sizhe Qiu, as the first author, conducted the literature review by conceptualizing the scope and outline, curating the literature, and producing the first draft of this review article.

Summary

Microbial stress responses configure the growth and metabolism of microorganisms under varying conditions, and stress-induced secondary metabolism is an important source of natural products widely used in various areas such as pharmaceutics and food additives. However, due to the limitations of conventional flux balance analysis (FBA), the quantitative modeling of microbial metabolism under stress is still poorly established. This chapter discusses current efforts on modeling primary metabolism under stress as well as stress induced secondary metabolism using FBA-based modeling techniques. Additionally, potential extensions of FBA are suggested to incorporate stress responses for the improvement of metabolic modeling of microorganisms. As this chapter posits, a FBA-based modeling framework capturing the effect of environmental stress will facilitate the quantitative study of microbial physiology and in-silico design of engineering strategies for microbial processes.

2.1 Introduction

The growth and metabolism of microorganisms are subject to the impact of changing environmental conditions, such as temperature and pH. With respect to primary metabolism that is coupled with cellular growth, enzymes, that are responsible for the cellular activity, are inhibited under stressed conditions (e.g., high temperature [1] or acidic pH [2]), and thus the catabolism for energy generation and anabolism for biomass formation are also inhibited [3]. Additionally, environmental stress can induce microbial secondary metabolism, which mediates microbial adaptation to the living environment [4]. Though unessential for cellular growth and reproduction from the simple mass balance point of view, secondary metabolism is a rich source of valuable natural products, which contribute to pharmaceutical, cosmetic, food and agricultural industries since the discovery of penicillin [5]. For example, lactic acid bacteria derived exopolysaccharides (EPSs) are used as food additives [6]. In general, the study of microbial metabolism under stress is both important for the advancement of the understanding of microbial physiology and analysis and engineering of the production of secondary metabolites of industrial values.

As microorganisms are widely used in industrial processes of various areas [7], many computational models for microbial metabolism have emerged in recent years. Two most commonly used modeling methods are differential-equation-based models [8,9] and flux balance analysis (FBA)-based models [10]. Admittedly, differential equation-based models can be easily connected with process control, but most of them treat the whole metabolic network as a “black box” [11,12]. On the other hand, fine-grained “white-box” differential equation-based models [9] are typically costly to construct due to limited data availability [13,14] and complex enzyme kinetic mechanisms [15]. Alternatively, the reconstruction of a genome-scale metabolic model (GSMM) is fast and mostly automated [16], and the major analytical approach of a GSMM, i.e., FBA using linear programming is less computationally expensive than solving a

large-scale differential equation set [17,18]. Therefore, this chapter focuses on how FBA has been used to model microbial metabolism under stress.

Despite the success of FBA in the field of systems and synthetic biology [19–22], the limitations of conventional FBA on stress responses have been perceived by researchers in recent years, and were partially described in several review articles. For example, Mohite et al., 2019 pointed out the shortcomings of different FBA-based modeling of the stress-induced production of antibiotics in actinomycetes [23]; Panikov, 2021 indicated that conventional FBA neglected stress resistance as an equivalently important objective as cellular growth [24]. Nevertheless, there still lacks a systematic review on metabolic modeling of microorganisms under stressed conditions. Therefore, this chapter intends to holistically and generically elucidate the major challenges faced by conventional FBA, and then critically assess various FBA-based modeling techniques for microbial metabolism under stress and stress-induced secondary metabolism. Finally, suggestions are given to potential extensions of FBA to improve the prediction accuracy of microbial stress responses. In brief, this chapter aims to clearly put forth the problem statement through the overview of current progress of FBA-based modeling techniques on microbial metabolism under stress and advise researchers on plausible future directions.

2.2 Challenges faced by conventional flux balance analysis

Flux balance analysis (FBA) models cellular metabolism by utilizing linear programming to compute metabolic fluxes, optimizing an objective function, usually biomass formation (v_{growth}) or a tailored objective ($c_{1\times R} * [v_1, v_2, \ldots, v_R]^T$) (Eq. 1), v stands for the reaction flux. The linear programming is solved in a constrained solution space of mass conservation (Eq. 2) and upper/lower bounds of reaction fluxes, $v_{i,max}$ and $v_{i,min}$ (Eq. 3) [10]. Eq. 2 is based on the pseudo-steady state assumption, and hence, FBA is often applicable only for modeling different

stabilized steady states but not transient phases where intracellular metabolite concentrations will change [24].

$$Maximize\ \ objective\ =\ v_{growth}\ or\ c_{1\times R} * [v_1, v_2, \ldots, v_R]^T\ (Eq.1)$$

$$Mass\ conservation: S_{M\times R} * [v_1, v_2, \ldots, v_R]^T\ =\ 0\ (Eq.2)$$

$$Upper/lower\ bounds: v_{i,min} \leq v_i \leq v_{i,max}\ (Eq.3)$$

v_{growth} is the biomass formation rate normalized to one-gram dry weight, also considered as the growth rate. $S_{M\times R}$ is the stoichiometric matrix of the metabolic network with M metabolites and Rreactions. A lot of toolboxes have been developed for conducting FBA, such as COBRA Toolbox [25], MetaFlux [26] and FBA-SimVis [27], and it has become increasingly common to use FBA to quantitatively study cellular metabolism, e.g. elucidation of metabolic responses to different culture conditions [28,29], identification of gene knockout to improve metabolite production [30], prediction of metabolite cross-feeding in microbial communities [31].

Although FBA has been successful in modeling microbial growth and primary metabolism, its applications in predicting the effect of environmental stress on primary metabolism and stress-induced secondary metabolism face challenges from several different aspects. The conventional FBA is tailored to simulate the growth and metabolism of microorganisms in the exponential phase, which is not ideal for simulating microbial metabolism under stress where the microorganisms are usually in the stationary or late growth phase [23,24]. Ergo, the first apparent challenge is the unsuitable objective function. Generally, the maximization of biomass formation rate is used as the objective function, based on the assumption that cells use the available nutrients most efficiently for biomass formation [32]. Consequently, the change in metabolic fluxes caused by the environmental stress, which are irrelevant with biomass formation, cannot be modeled due to the default setting of conventional FBA [23]. Besides, the stress resistance, sometimes considered as an equally important objective of microorganisms [24], is ignored in conventional FBA. Theoretically, the optimality

principle used to predict microbial metabolism should keep the balance between biomass formation and minimization of death under stress conditions. Another inherent shortcoming of conventional FBA is its lack of characterization of gene expression regulation. The metabolic switch is mediated by gene expression regulation, and therefore, the simulation of microbial metabolism affected by environmental stress will become more accurate if FBA incorporates the gene regulatory network (GRN) [33]. Without incorporating stress responses, FBA would not be able to accurately simulate microbial growth and metabolism under changing environmental conditions.

2.3 FBA-based modeling techniques for primary metabolism under stress

To address the limitations of conventional FBA discussed in **section 2.2**, some existing FBA-based approaches incorporated the effect of environmental stress into the modeling of microbial primary metabolism: 1. Set the model parameter as a function of the stress; 2. Integrating gene expression data in FBA. The existing modeling techniques and their applications are summarized in **Table 2.1**, as well as their predictive powers.

In conventional FBA, the growth function (the default objective) and constraints on exchange fluxes are deterministic to the result of metabolic flux simulation (**section 2.2**), and hence, some researchers incorporated environmental stress into those two parameters to quantitatively model the cellular metabolism under stress. Metris et al. 2012 set the concentrations of potassium, glutamate, trehalose and glycine betaine and ATP requirement in the growth function as variables dependent on medium osmolality [34]. Though the quantitative accuracy was low, Metris et al. 2012 could qualitatively predict the decrease of carbon source uptake rate and growth rate caused by the increase of medium osmolality. Similarly, Özcan et al. 2021 set the upper bounds of carbon and nitrogen source uptake fluxes as exponential

decay functions of the undissociated lactic acid concentration [35], which was adopted from Vereecken & Van Impe, 2002 [36] and Aghababaie et al., 2015 [37], and obtained high accuracy in modeling the growth kinetics, glucose consumption and lactic acid production of lactic acid bacteria. Incorporating the environmental stress into model parameters is intended to resolve the lack of characterization of regulation in conventional FBA, but this approach fails to represent the molecular mechanism of stress response due to the simplicity of regulatory terms.

Modeling techniques discussed above do not take into account the inherent lack of characterizing regulation in conventional FBA. Reactions in a GSMM, if not gap filled, have associated gene-protein-reaction (GPR) rules, allowing the integration of quantitative gene expression data, such as transcriptomics or proteomics. If gene expression levels are integrated into FBA, the predicted fluxes can reflect metabolic enzyme activities, and thus characterize regulations. The generic algorithm is to minimize the utilization of reactions with low gene expression levels and allow reactions with high gene expression levels to have higher absolute flux values. Up till now, there have been various developed techniques for integrating gene expression data into FBA, e.g. GIMME [38], E-flux [39], iMAT [40], but the application of this modeling technique on stress responses is quite scarce.

Montezano et al. 2015 defined an alternative objective function for FBA based on protein expression levels to model the growth and metabolism of *Mycobacterium tuberculosis* under the antibiotic stress by mefloquine [41]. This modeling technique demonstrated its good accuracy with a lower prediction error than E-flux [39]. Nonetheless, this technique relies on condition-specific gene expression data, which limits the scope of extrapolation. It can predict metabolism for various design settings of interest only if different gene expression data corresponding to the different settings are provided. Otherwise, this technique, to a large extent, can merely compute the flux distribution corresponding to a specific set of gene expression levels.

Table 2.1. Summary of modeling techniques for primary metabolism under stress

Stress	Organism	Modeling technique	Predictive power	Reference
Osmolality	*Escherichia coli*	Adjust the growth function based on medium osmolality	Qualitatively accurate but low quantitative accuracy	[34]
Oxidative stress by paraquat	*Escherichia coli*	Multi-objective optimization using epsilon-constraint method	Not quantitatively examined	[42]
Antibiotic stress by mefloquine	*Mycobacterium tuberculosis*	Define an alternative objective function based on proteomic data	Good accuracy but restricted by settings	[41]
Undissociated weak acid	*Lactococcus lactis subsp. cremoris, Lactococcus lactis subsp. lactis, Streptococcus thermophilus, Leuconostoc mesenteroides*	Adjust the upper bound of nutrient uptake based on the undissociated lactic acid concentration	Good accuracy and unrestricted by settings	[35]

2.4 FBA-based modeling techniques for stress-induced secondary metabolism

To address the limitations of conventional FBA on stress-induced secondary metabolism discussed in **section 2.2**, several techniques have been developed, including: 1. adding the targeted secondary metabolite into the growth function (biomass formation); 2. switching from the classical biomass formation objective to a secondary metabolism-associated objective; 3. estimating metabolic fluxes through sampling in a strictly constrained space; 4. Integrating gene expression data in FBA. The first three techniques aim to overcome the limitation caused by the unsuitable objective function, while the last one manages to incorporate gene expression

regulation. The existing modeling techniques and their applications are summarized in **Table 2.2**, and their predictive power is also discussed.

Firstly, the flux distribution obtained from conventional FBA highly depends on the specific objective function used. If maximizing biomass formation is selected as the objective function, as commonly done, FBA would predict zero flux through the secondary metabolic pathways as it does not contribute to cell growth [43]. To resolve this type of limitation, it has been proposed to add the secondary metabolite into the biomass formation reaction, so that FBA will optimize both the synthesis of biomass precursors and the target secondary metabolite. This modeling technique was used to simulate the antibiotic production in *Streptomyces coelicolor*: based on the observed antibiotics production rate, the amount of antibiotics was increased dynamically when the stress-induced secondary metabolism was activated [44,45]. Because stoichiometric coefficients of undecylprodigiosin (RED) and actinorhodin (ACT) in biomass formation objective function were set based on experimental measurement, the accuracy in predicting the growth rate, R-squared value = 0.95, could reflect the accuracy of predicting secondary metabolite production (**Figure 2.1A**). This technique could account for the metabolic switch, as the studies by Alam et al., 2010 [44] and Amara et al., 2018 [45] demonstrated significant correlations between gene expression levels and predicted metabolic fluxes for most genes in secondary metabolic pathways. However, anti-correlations of predicted fluxes and gene expression levels for some genes in calcium dependent antibiotics (CDA) and undecylprodigiosin (RED) biosynthesis indicated that the failure of including regulatory constraints was an important source of error [44,45]. From the perspective of modeling technique, in this case, FBA still optimizes the growth rate and makes the production of secondary metabolites growth-associated, which contradicts with the biological fact that the biosynthesis of most secondary metabolites is growth-unassociated. Also, this modeling technique manually fixes the ratio between secondary metabolite production and biomass formation (i.e., product yield per unit biomass) with experimental measurements, which restricts

its applicability for cases where the yield of the secondary metabolite is varied or unknown. Due to the lack of a solid biological basis, this approach can only model flux distributions if the microbial cell behaves according to the artificially-adjusted objective function. Though such an approach can, to some extent, reflect statuses of primary and secondary metabolisms, it can hardly be used directly for design purposes, because it cannot predict secondary metabolite production for different design settings of interest (e.g., different growth media, different strains).

As an alternative to the manipulation of the biomass formation while still adopting the classic objective function that maximizes biomass growth, switching to a new objective function can avoid introducing inappropriate growth association. The commonly used candidates of alternative objective functions are ATP yield, ATP yield per unit flux, biomass formation per unit flux and target secondary metabolite production. Toro et al., 2018 compared four different objective functions' performances on both primary metabolism and clavulanic acid production of *Streptomyces clavuligerus* [46], and the study found that maximization of ATP yield could predict primary metabolism and clavulanic acid production flux with best accuracy (**Figure 2.1B**). ATP yield maximization can be assumed to be the objective for microorganisms in both exponential growth and stationary phases, as the cell is always in need of energy to maintain its activity.

A more direct objective function is to maximize the target secondary metabolite production. In the simulation of riboflavin overproduction by *Ashbya gossypii*, Ledesma-Amaro et al., 2014 assumed that *Ashbya gossypii* in the stationary phase switched from maximizing biomass formation entirely to maximizing riboflavin production [47]. The result showed that the simulated value of riboflavin production rate, 0.0156 mmol/gDW/h [47], was close to the experimental value, 0.0126 mmol/gDW/h [48]. Similar to Ledesma-Amaro et al., 2014 [47], Shen et al. 2013 used epsilon-constraint method to maximize NADPH production in response to the oxidative stress by paraquat, while the biomass formation was set as an inequality constraint [42]. This technique could assist the analysis of the trade-off between the biomass formation

and the generation of reductive power, but it did not provide a mathematical formulation of the relationship between cellular growth and stress resistance, which is dependent on the level of oxidative stress. Also, the quantitative accuracy of this technique was not examined with experimental data.

Using flux variability analysis (FVA) to maximize the target secondary metabolite production from the flux solutions that maximize biomass growth is another approach, which has been used in modeling the production of flavor metabolites, such as acetoin, in lactic acid bacteria [35], but this approach has not been quantitatively examined with experimental data. The mathematical formulation of secondary metabolism-associated objective function in place of biomass formation has shown its ability to predict the production of some secondary metabolites, independently of condition-specific experimental data. However, changing the objective function may be too simplistic to account for metabolic activities that keep the balance of growth and stress resistance as discussed in **section 2.1**.

Since the FBA solution can be inaccurate with an unsuitable objective function, a modeling technique that is independent of objective functions has been adopted to resolve the issue. In the simulation using the Sco-GEM model, a consensus smGSMM for *Streptomyces coelicolor*, the metabolic fluxes were strictly constrained with the enzyme capacity, $v = k_{cat}[E]$. k_{cat} is the turnover rate and $[E]$ is the concentration of the enzyme. Fluxes were approximated using unbiased random sampling in the constrained solution space instead of optimizing a defined objective function [49]. The metabolic switch from glycolysis, fatty acid and nucleotide biosynthesis to ACT, RED, CDA and coelimycin P1 (Cpk) biosynthesis was captured by this technique [49], but no comparison of predicted and observed reaction fluxes was performed for quantitative validation. Because random sampling without a highly constrained space will output multiple flux states with wide ranges, the accuracy of the predicted fluxes relies on the strict constraints imposed by quantified enzyme protein concentrations and enzyme kinetic parameters, which are strain-specific and condition-specific. Therefore, this modeling technique

cannot compute metabolic fluxes through secondary metabolic pathways without proteomic data measured under a given setting. The cost of proteomics quantification makes it hard to predict secondary metabolite production with different growth conditions.

With respect to using the integration of gene expression data into FBA to model secondary metabolism, there exists a transcriptomics based strain optimization tool for secondary metabolite production that uses iMAT to predict the secondary metabolism of *Streptomyces coelicolor* [50]. In the comparison of different gene expression integrated FBA modeling techniques, iMAT was found to be the only technique that can capture the onset of secondary metabolism in *Streptomyces coelicolor* [50]. Nonetheless, the discrepancy between predicted and observed production fluxes of actinorhodin (ACT) and undecylprodigiosin (RED) indicates that iMAT only performs well in a qualitative manner (**Figure 2.1C**). Just like Montezano et al. 2015 [41], using iMAT to model stress-induced secondary metabolism in *Streptomyces coelicolor* also relied on condition-specific gene expression data and the simulation was restricted by design settings. In short, to our knowledge, so far there has not been a gene expression integrated modeling technique that predicts secondary metabolite production with good quantitative accuracy.

In a nutshell, existing FBA-based modeling techniques that manipulate the objective function cannot capture two distinct life objectives of a microbial cell concurrently, i.e., optimizing growth and stress resistance [24]. Besides, how to formulate a mathematical model that keeps the balance between the two objectives still remains an open question. Furthermore, no mechanistic representation of gene expression regulation upstream to metabolism has been built and integrated in FBA to make better use of omics datasets than simply determining “on” and “off” of reactions for fixed settings. In summary, for existing FBA-based modeling techniques that predict stress-induced secondary metabolism, those without integration of gene expression data are either limited to specific cases or too simplistic that only work under certain

ideal assumptions, while gene expression integrated techniques are scarce and unable to make predictions for different settings without costly data generation.

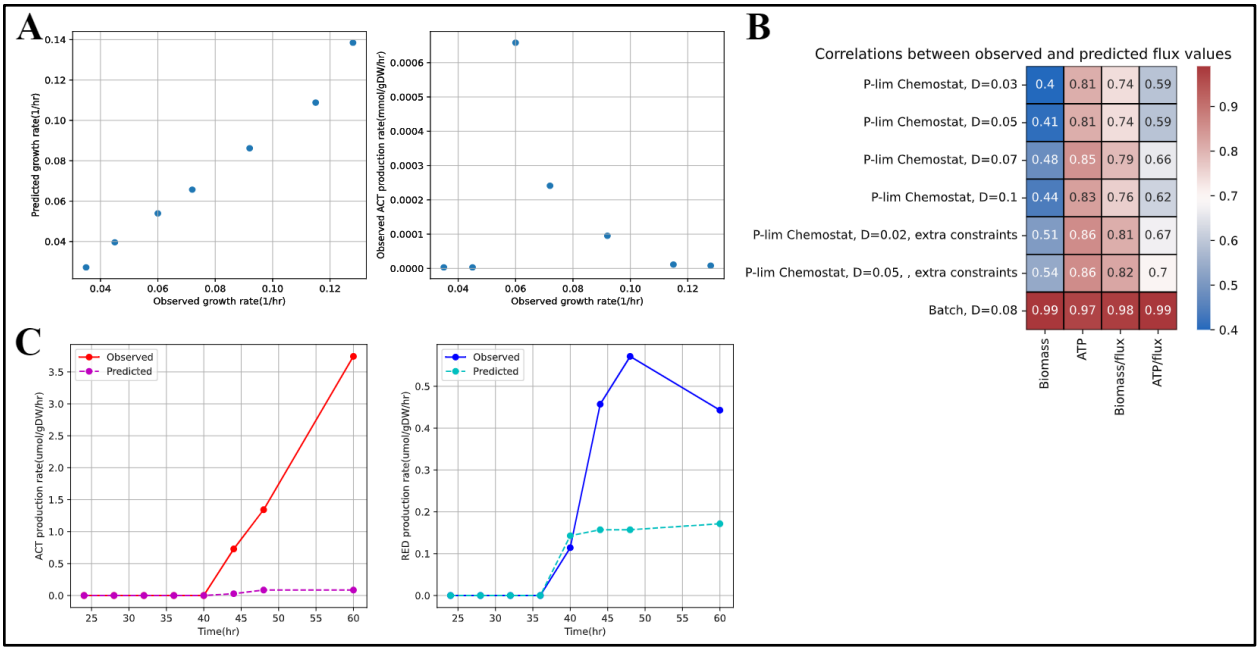

Figure 2.1. Quantitative assessment of existing FBA-based modeling techniques for predicting secondary metabolite production. (A) Comparison of predicted and observed growth rates, and ACT production fluxes at different observed growth rates in Alam et al., 2010 [44] (the first row of Table 2.2). (B) Comparison of 4 different objective functions used in FBA to predict both primary metabolism and clavulanic acid production flux [46] (the second row of Table 2.2). Correlation scores are computed for predicted and observed fluxes. P-lim: limited phosphorus content. (C) Comparison of predicted and observed production fluxes of ACT and RED in Kim et al., 2016 [50] (the last row of Table 2.2).

Table 2.2. Summary of modeling techniques for stress-induced secondary metabolism

Secondary metabolite	Organism	Modeling technique	Predictive power	Reference
ACT, RED	*Streptomyces coelicolor*	Add the secondary metabolite into biomass and maximize	Good accuracy but restricted by settings	[44,45]

		biomass formation		
Clavulanic acid	*Streptomyces clavuligerus*	Use ATP yield as the objective function	Good accuracy and unrestricted by settings	[46]
Riboflavin	*Ashbya gossypii*	Use secondary metabolite yield as the objective function	Good accuracy and unrestricted by settings	[47]
NADPH	*Escherichia coli*	Use secondary metabolite yield as the objective function	Not quantitatively examined	[42]
Acetoin, diacetyl, acetaldehyde, benzaldehyde, 2,3-butanediol and amino acid derived flavor metabolites, 2-methylbutanal, 2-methylpropanoic acid, etc.	*Lactococcus lactis subsp. cremoris, Lactococcus lactis subsp. lactis, Streptococcus thermophilus, Leuconostoc mesenteroides*	Maximize secondary metabolite yield in solution spaces of FVA	Not quantitatively examined	[35]
ACT, RED, CDA, Cpk	*Streptomyces coelicolor*	Random sampling with enzyme constraints	Qualitatively accurate but not quantitatively examined	[49]
ACT, RED	*Streptomyces coelicolor*	Integration of transcriptomics data into FBA	Low quantitative accuracy	[50]

2.5 Potential extensions of FBA towards modeling microbial metabolism under environmental stress

The associated GPR rules, already targeted by some of the existing approaches to model microbial metabolism under stress as summarized in **sections 2.3 and 2.4,** suggests that the ability to determine the metabolic capacity of pathways under environmental stress could be

key to accurately modeling microbial metabolism under stress. Here, this chapter proposes two potential extensions of FBA that may resolve issues of existing modeling techniques discussed in **sections 2.3 and 2.4.**

2.5.1 Constrained proteome allocation for stress-induced secondary metabolism

Proteome allocation, namely the distribution of proteome resources in different pathways, governs cellular metabolism by controlling maximum reaction fluxes [51]. The constrained proteome allocation theory has been proposed to model the competition for proteome resources among different functional sectors such as catabolism, anabolism, transportation, etc., characterizing the coordination of proteome partitions and cellular metabolism under different growth conditions [52–54]. The theoretical model has been applied in various FBA-based models, such as CAFBA [55] or ME-model [56], via converting it into proteomic constraints on reaction fluxes. There are different ways to divide the proteome based on the modeling requirement. For example, the proteome was divided into fermentation, respiration and biomass formation sectors in the prediction of overflow metabolism in *Escherichia coli* [57]. Usually, glycolytic enzymes are clustered as the sector of catabolism or energy, membrane transporter proteins are clustered as the sector of transportation, and a lumped proteome resource for the growth function (biomass formation) is considered the sector of anabolism [55,58,59]. The neglect of other enzymes' proteome costs normally will not significantly affect the simulation, as fluxes through those pathways are far smaller than that through central carbon metabolism [55,58,59].

Though constrained proteome allocation embedded FBA models have achieved good accuracy in case studies such as predicting overflow metabolism in *E. coli* [58] or explaining lactic acid production in lactic acid bacteria [59], to our knowledge it has not been used to

simulate stress-induced secondary metabolism yet. The proteome partitioning has not included a sector defined for secondary metabolism, as we found in various proteome allocation models. To apply proteome allocation to predict secondary metabolite production, one potential approach is integrating Synthetic Chemostat Model (SCM) into FBA [24]. SCM is a differential equation-based model of microbial growth kinetics that divides cellular metabolism into a P-component for growth and a U-component for stress resistance [60]. Being a macroscopic bioreactor-level model, SCM is too coarse to characterize different functional proteome sectors, but the embedded concept can be used to modify the original constrained proteome allocation framework. If a U-sector containing enzymes in biosynthetic pathways of secondary metabolites is included and stress response is introduced into proteome allocation, then the modified model might be able to characterize both growth and stress response via proteome allocation at the metabolic switch that leads to stress-induced secondary metabolism. Below, an illustrative example of pH-induced EPS production in lactic acid bacteria [61,62] is presented for the proposed constrained proteome allocation model for both branches of metabolism (**Figure 2.2**).

$$\phi_Q(\sim 50\%) + \phi_U + \phi_C + \phi_R + \phi_T \leq 1 \; (Eq. 4)$$

$$\frac{\phi_U}{\phi_U + \phi_C + \phi_R + \phi_T} \geq k_1 e^{\frac{-k_2}{(6.5-pH)^2}}, k_1, k_2 > 0 \; (Eq. 5)$$

$$v_i \leq kcat_i[E_i], \sum_i [E_i] \leq \phi_x[P_{TOT}], x = U, C, R, T \; (Eq. 6)$$

The cellular proteome is divided into inflexible housekeeping sector (Q-sector), catabolic sector (C-sector), ribosomal sector for protein translation (R-sector), transportation sector (T-sector) and secondary metabolism sector (U-sector). ϕ_x is the mass fraction of the sector x (Eq. 4, 6). The total amount of proteome resources is conserved (Eq. 4). In response to environmental stress, which is acidity in this case, the fraction of proteome resources allocated to the U-sector, which is EPS biosynthesis, rises up (Eq. 5). The metabolic fluxes are constrained by the total amount of proteome resources allocated to the sector (Eq. 6). $[E_i]$, $kcat_i$ are the concentration

and turnover rate of the enzyme *i* respectively, and $[P_{TOT}]$ is the concentration of total cellular proteins. The proteome allocation of secondary metabolism (Eq. 5) can materialize in different forms in actual implementation, e.g. modeling the consequence of differing proteomic costs caused for different stress factors, such as inhibitors [63] or nutrient limitation [55,58].

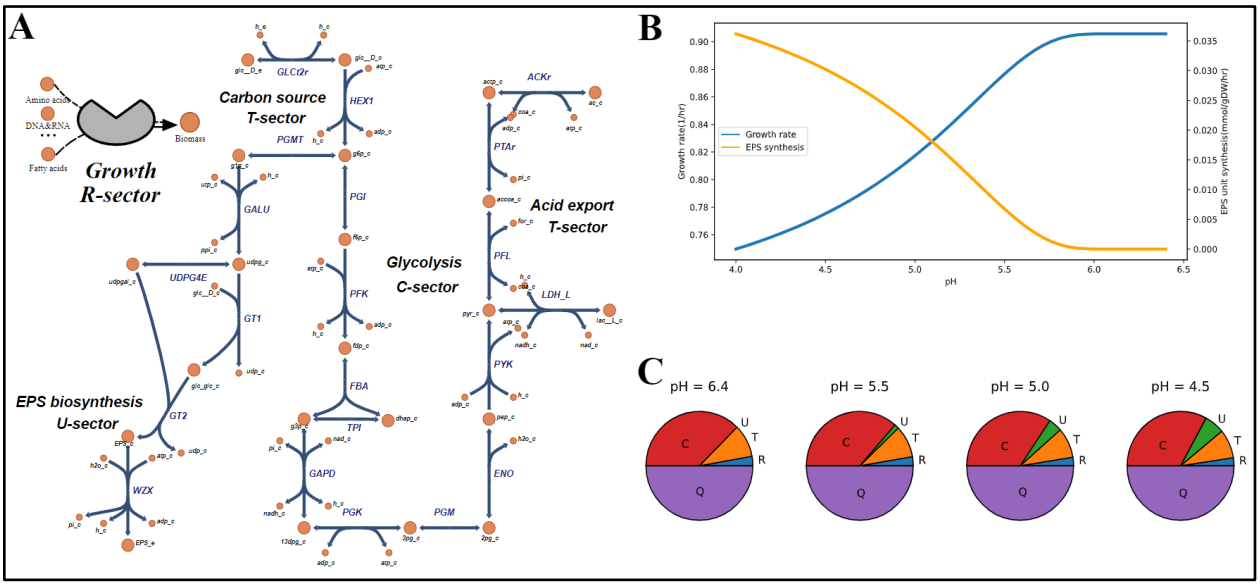

Figure 2.2. An illustrative example of pH induced EPS production in lactic acid bacteria, used to explain the constrained proteome allocation model for both primary and secondary metabolism. (A) The metabolic network of lactic acid bacteria for both primary metabolism (C, R, T sectors) and stress induced secondary metabolism (U sector). (B) Simulated metabolic response to pH: the increase of acidity inhibits the growth rate and induces EPS production. (C) Simulated proteome allocation in response to pH: the increase of acidity activates secondary metabolism, and more proteome resources get allocated to the U sector. **Note: this "toy" model is for illustration only.**

2.5.2 Combine gene regulatory and metabolic networks to capture the environmental stress

Constrained proteome allocation-embedded FBA models can capture the shift in metabolic states via modeling proteome resource distribution among functional sectors [55]. However, they do not characterize regulatory interactions between regulatory factors (RFs) and genes at transcriptional or translational level. On the other hand, existing gene expression integrated FBA modeling techniques, such as iMAT, do not have a mechanistic representation of gene expression regulation, and hence, their applications on modeling microbial metabolism under stress are restricted by condition-specific data availability. Therefore, a regulatory network model needs to be combined with FBA, as illustrated in **Figure 2.3**, to resolve the limitation of direct integration of gene expression into FBA. Though not for modeling the effect of environmental stress on microbial metabolism, several integrated regulatory-metabolic models, also called regulatory FBA (rFBA), have been reported for a while [64–67]. The generic framework of dynamic rFBA is a combination of FBA for intracellular metabolic fluxes, a boolean logic network modeling the regulatory network, and a set of differential equations modeling concentrations of RFs, biomass and extracellular metabolites [68].

Different from gene expression integrated FBA, rFBA can capture different regulated conditions through tailored activation functions of regulator levels, and the boolean logic network determines the expression status of the gene associated with the reaction. The boolean logic network can mechanistically model indirect regulatory interactions, such as the activation of a regulator gene by another regulator resulting in the activation of a metabolic reaction. Then, rFBA constrains metabolic fluxes by only allowing nonzero fluxes through active reactions. Similar regulatory representations are also used in some whole-cell kinetic models to simulate the activation of certain genetic circuits [69,70], but kinetic models are currently difficult to be

implemented on microbial metabolism under environmental stress due to the lack of parameters and complexity of enzyme kinetics, as discussed in the introduction.

Enlightened by the idea of dynamic rFBA, if regulatory interactions related to environmental stress are elucidated in a boolean logic network, dynamic simulation of gene expression regulations can be performed, and thus quantitatively model the change in the metabolic capacity of primary metabolism and the onset of secondary metabolism in response to environmental stress (**Figure 2.3**). The stress response can be appended into rFBA, as the process from stress sensing to regulation can be represented by signal transduction in a network [71], though such approach has not been applied to the modeling of stress responses. In essence, rFBA has the potential to overcome the limitation of conventional FBA by including the mechanistic representation of gene expression regulation, which will enhance the predictive power of FBA on both primary metabolism under stress and stress-induced secondary metabolism.

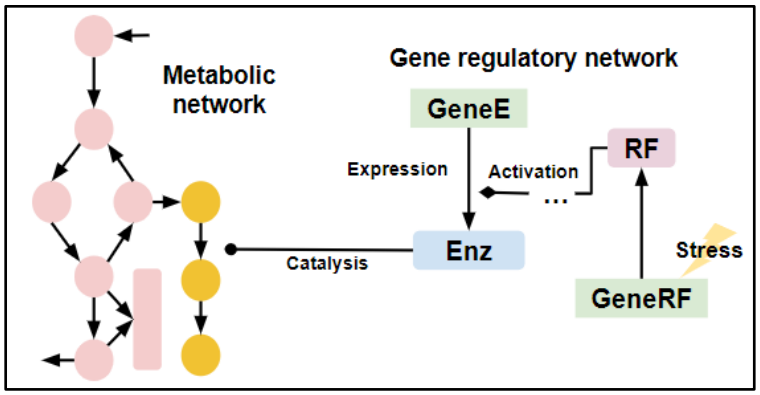

Figure 2.3. Schematic diagram of the combination of metabolic and gene regulatory networks to model stress-induced secondary metabolism. The stress stimulates the expression of regulatory factor (RF), resulting in the activation of enzymes catalyzing reactions for secondary metabolite biosynthesis. **Orange circle:** secondary metabolite; **Pink circle:** primary metabolite; **RF:** regulatory factor; **Enz:** enzyme.

2.6 Conclusions and outlook

This chapter summarized key challenges faced by conventional FBA on modeling microbial metabolism under stress, which mainly derived from the neglect of other important cellular activities apart from cellular growth. In the following sections, existing FBA-based modeling techniques that intend to tackle those challenges have been critically evaluated in this chapter. Those modeling techniques are either too simplistic or have restricted application scopes. None of them are predictive enough to mechanistically capture the effect of environmental stress. Nevertheless, some of them show usefulness in certain cases, such as manipulating the objective function and integrating gene expression data into FBA.

With the aim of building a modeling technique that captures the mechanistic understanding of microbial stress responses, this chapter has suggested the application of two potential extensions of FBA-based modeling techniques. They respectively incorporate constrained proteome allocation that explicitly account for the change of metabolic capacity in response to environmental stress, and connect with the regulatory network model for gene expression states that control reaction activities with high resolution. Detailed implementation approaches in those directions remain to be explored, which we envisage will benefit from further and more mechanistic understanding of how microorganisms mediate environmental stress. Additionally, future curation works of proteomic and gene regulatory information, both in connection with typical and specific stress signaling pathways in microorganisms, will contribute to the advancement of those approaches.

When a modeling technique is developed to capture the mechanism of microbial stress responses, the complex microbial metabolism, subject to the effect of environmental stress, will then become more “white-boxed.” Researchers, particularly those in the field of systems and synthetic biology, will be able to conduct more reliable in-silico analysis of microbial growth and metabolism under changing conditions. This advancement will contribute to the development of

model-based design and pathway engineering approaches aimed at controlling microbial metabolism, such as enhancing the productivity of high-value natural products.

2.7 Rationale for the scope of this PhD study

This chapter explained the limitations of existing FBA-based models on predicting microbial metabolism under environmental stress. Subsequently, constrained proteome allocation and the mechanistic representation of gene regulatory interactions were proposed as potential directions to incorporate the effect of environmental stress into metabolic modeling for the enhancement of prediction accuracy. Consequently, this PhD study focused on using proteome allocation based metabolic modeling to account for various stress responses and analyzing the gene expression regulatory network. The lactic acid bacteria were selected as the research object. For environmental stress, we chose the product inhibition of undissociated lactic acid, the change of temperature, and acidic pH, which were the most important factors in the fermentation by lactic acid bacteria (e.g., yogurt fermentation). First, Project 1 examined the use of regulatory terms on proteome constraints in proteome constrained FBA to model the product inhibition of undissociated lactic acid. Then, Project 2 used deep learning to predict enzyme k_{cat} at different temperatures to resolve the lack of kinetic parameters for most microorganisms, which could impede temperature sensitive metabolic modeling. With predicted k_{cat}, Project 2 experimented temperature-sensitive proteome constrained FBA for two widely used species of lactic acid bacteria. Before the incorporation of secondary metabolism into proteome constrained FBA, Project 3 inferred the transcriptional regulatory network of *L. plantarum* and analyzed the relationship between primary metabolism and EPS biosynthesis mediating acid stress. Next, Project 4 built the regulatory proteome constrained FBA (RPCFBA) model, based on the results of Project 3 and multi-omics analysis, to simulate primary metabolism and EPS production in *L. plantarum* simultaneously. Ultimately, the objective of this

PhD study was to enhance the modeling performance of microbial processes by incorporating the effect of environmental stress into proteome constrained FBA.

References

1. Chen J, Shen J, Ingvar Hellgren L, Ruhdal Jensen P, Solem C. Adaptation of Lactococcus lactis to high growth temperature leads to a dramatic increase in acidification rate. Sci Rep. 2015;5: 14199.

2. Bendig T, Ulmer A, Luzia L, Müller S, Sahle S, Bergmann FT, et al. The pH-dependent lactose metabolism of Lactobacillus delbrueckii subsp. bulgaricus: An integrative view through a mechanistic computational model. J Biotechnol. 2023;374: 90–100.

3. Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Nat Rev Microbiol. 2014;12: 327–340.

4. Seyedsayamdost MR. Toward a global picture of bacterial secondary metabolism. J Ind Microbiol Biotechnol. 2019;46: 301–311.

5. Sanchez S, Guzmán-Trampe S, Ávalos M, Ruiz B, Rodríguez-Sanoja R, Jiménez-Estrada M. Microbial Natural Products. Natural Products in Chemical Biology. 2012. pp. 65–108. doi:10.1002/9781118391815.ch3

6. Korcz E, Varga L. Exopolysaccharides from lactic acid bacteria: Techno-functional application in the food industry. Trends Food Sci Technol. 2021;110: 375–384.

7. Pham JV, Yilma MA, Feliz A, Majid MT, Maffetone N, Walker JR, et al. A Review of the Microbial Production of Bioactive Natural Products and Biologics. Frontiers in Microbiology. 2019. doi:10.3389/fmicb.2019.01404

8. Costa RS, Hartmann A, Vinga S. Kinetic modeling of cell metabolism for microbial production. J Biotechnol. 2016;219: 126–141.

9. Foster CJ, Wang L, Dinh HV, Suthers PF, Maranas CD. Building kinetic models for metabolic engineering. Curr Opin Biotechnol. 2021;67: 35–41.

10. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248.

11. Dochain D, Bastin G. ADAPTIVE CONTROL OF FEDBATCH BIOREACTORS. Chem Eng Commun. 1990;87: 67–85.

12. Bouguettoucha A, Balannec B. Unstructured models for lactic acid fermentation-a review. Food Technol. 2011. Available: https://www.researchgate.net/profile/Abdallah-Bouguettoucha/publication/228355166_Unstructured_Models_for_Lactic_Acid_Fermentation-A_Review/links/02e7e53bedb3eb05fe000000/Unstructured-Models-for-Lactic-Acid-Fermentation-A-Review.pdf

13. Breitling R, Achcar F, Takano E. Modeling challenges in the synthetic biology of secondary metabolism. ACS Synth Biol. 2013;2: 373–378.

14. Nilsson A, Nielsen J, Palsson BO. Metabolic Models of Protein Allocation Call for the Kinetome. Cell Syst. 2017;5: 538–541.

15. Ulusu NN. Evolution of Enzyme Kinetic Mechanisms. J Mol Evol. 2015;80: 251–257.

16. Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Biochem Soc Trans. 2018;46: 931–936.

17. Damiani C, Di Filippo M, Pescini D, Maspero D, Colombo R, Mauri G. popFBA:

tackling intratumour heterogeneity with Flux Balance Analysis. Bioinformatics. 2017;33: i311–i318.

18. Martelli C, De Martino A, Marinari E, Marsili M, Pérez Castillo I. Identifying essential genes in Escherichia coli from a metabolic optimization principle. Proc Natl Acad Sci U S A. 2009;106: 2607–2611.

19. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Brief Bioinform. 2009;10: 435–449.

20. Antoniewicz MR. A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications. Metab Eng. 2021;63: 2–12.

21. Gianchandani EP, Chavali AK, Papin JA. The application of flux balance analysis in systems biology. Wiley Interdiscip Rev Syst Biol Med. 2010;2: 372–382.

22. Sahu A, Blätke M-A, Szymański JJ, Töpfer N. Advances in flux balance analysis by integrating machine learning and mechanism-based models. Comput Struct Biotechnol J. 2021;19: 4626–4640.

23. Mohite OS, Weber T, Kim HU, Lee SY. Genome-Scale Metabolic Reconstruction of Actinomycetes for Antibiotics Production. Biotechnol J. 2019;14: e1800377.

24. Panikov NS. Genome-Scale Reconstruction of Microbial Dynamic Phenotype: Successes and Challenges. Microorganisms. 2021;9. doi:10.3390/microorganisms9112352

25. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols. 2007. pp. 727–738. doi:10.1038/nprot.2007.99

26. Latendresse M, Ong WK, Karp PD. Metabolic Modeling with MetaFlux. Methods

Mol Biol. 2022;2349: 259–289.

27. Grafahrend-Belau E, Klukas C, Junker BH, Schreiber F. FBA-SimVis: interactive visualization of constraint-based metabolic models. Bioinformatics. 2009;25: 2755–2757.

28. Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y. Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. Metabolic Engineering. 2011. pp. 38–48. doi:10.1016/j.ymben.2010.11.004

29. Vijayakumar S, Rahman PKSM, Angione C. A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria. iScience. 2020;23: 101818.

30. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng. 2003;84: 647–657.

31. Zorrilla F, Buric F, Patil KR, Zelezniak A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. Nucleic Acids Res. 2021;49: e126.

32. Toya Y, Shimizu H. Flux analysis and metabolomics for systematic metabolic engineering of microorganisms. Biotechnol Adv. 2013;31: 818–826.

33. Nieselt K, Battke F, Herbig A, Bruheim P, Wentzel A, Jakobsen ØM, et al. The dynamic architecture of the metabolic switch in Streptomyces coelicolor. BMC Genomics. 2010;11: 10.

34. Metris A, George S, Baranyi J. Modelling osmotic stress by Flux Balance Analysis at the genomic scale. Int J Food Microbiol. 2012;152: 123–128.

35. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. Biotechnol Bioeng. 2021;118: 223–237.

36. Vereecken KM, Van Impe JF. Analysis and practical implementation of a model for combined growth and metabolite production of lactic acid bacteria. Int J Food Microbiol. 2002;73: 239–250.

37. Aghababaie M, Khanahmadi M, Beheshti M. Developing a kinetic model for co-culture of yogurt starter bacteria growth in pH controlled batch fermentation. J Food Eng. 2015;166: 72–79.

38. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4: e1000082.

39. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, et al. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009;5: e1000489.

40. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26: 3140–3142.

41. Montezano D, Meek L, Gupta R, Bermudez LE, Bermudez JCM. Flux Balance Analysis with Objective Function Defined by Proteomics Data-Metabolism of Mycobacterium tuberculosis Exposed to Mefloquine. PLoS One. 2015;10: e0134014.

42. Shen T, Rui B, Zhou H, Zhang X, Yi Y, Wen H, et al. Metabolic flux ratio analysis and multi-objective optimization revealed a globally conserved and coordinated metabolic response of E. coli to paraquat-induced oxidative stress. Mol Biosyst. 2013;9: 121–132.

43. García Sánchez CE, Torres Sáez RG. Comparison and analysis of objective functions in flux balance analysis. Biotechnol Prog. 2014;30: 985–991.

44. Alam MT, Merlo ME, STREAM Consortium, Hodgson DA, Wellington EMH, Takano E, et al. Metabolic modeling and analysis of the metabolic switch in Streptomyces coelicolor. BMC Genomics. 2010;11: 202.

45. Amara A, Takano E, Breitling R. Development and validation of an updated computational model of Streptomyces coelicolor primary and secondary metabolism. BMC Genomics. 2018;19: 519.

46. Toro L, Pinilla L, Avignone-Rossa C, Ríos-Estepa R. An enhanced genome-scale metabolic reconstruction of Streptomyces clavuligerus identifies novel strain improvement strategies. Bioprocess Biosyst Eng. 2018;41: 657–669.

47. Ledesma-Amaro R, Kerkhoven EJ, Revuelta JL, Nielsen J. Genome scale metabolic modeling of the riboflavin overproducer Ashbya gossypii. Biotechnol Bioeng. 2014;111: 1191–1199.

48. Stahmann KP, Arst HN Jr, Althöfer H, Revuelta JL, Monschau N, Schlüpen C, et al. Riboflavin, overproduced during sporulation of Ashbya gossypii, protects its hyaline spores against ultraviolet light. Environ Microbiol. 2001;3: 545–550.

49. Sulheim S, Kumelj T, van Dissel D, Salehzadeh-Yazdi A, Du C, van Wezel GP, et al. Enzyme-Constrained Models and Omics Analysis of Streptomyces coelicolor Reveal Metabolic Changes that Enhance Heterologous Production. iScience. 2020;23: 101525.

50. Kim M, Yi JS, Lakshmanan M, Lee D-Y, Kim B-G. Transcriptomics-based strain optimization tool for designing secondary metabolite overproducing strains of Streptomyces coelicolor. Biotechnol Bioeng. 2016;113: 651–660.

51. Sánchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol. 2017;13: 935.

52. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330: 1099–1102.

53. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Mol Syst Biol. 2015;11: 784.

54. Zeng H, Rohani R, Huang WE, Yang A. Understanding and mathematical modelling of cellular resource allocation in microorganisms: a comparative synthesis. BMC Bioinformatics. 2021;22: 467.

55. Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained Allocation Flux Balance Analysis. PLoS Comput Biol. 2016;12: e1004913.

56. Yang L, Yurkovich JT, King ZA, Palsson BO. Modeling the multi-scale mechanisms of macromolecular resource allocation. Curr Opin Microbiol. 2018;45: 8–15.

57. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, et al. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 2015;528: 99–104.

58. Zeng H, Yang A. Modelling overflow metabolism in Escherichia coli with flux balance analysis incorporating differential proteomic efficiencies of energy pathways. BMC Syst Biol. 2019;13: 3.

59. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM, Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

60. Panikov NS. Mechanistic mathematical models of microbial growth in bioreactors and in natural soils: Explanation of complex phenomena. Math Comput Simul. 1996;42: 179–186.

61. Torino MI, Taranto MP, Sesma F, de Valdez GF. Heterofermentative pattern and exopolysaccharide production by Lactobacillus helveticus ATCC 15807 in response to environmental pH. J Appl Microbiol. 2001;91: 846–852.

62. Nguyen P-T, Nguyen T-T, Bui D-C, Hong P-T, Hoang Q-K, Nguyen H-T. Exopolysaccharide production by lactic acid bacteria: the manipulation of environmental stresses for industrial applications. AIMS Microbiol. 2020;6: 451–469.

63. Qiu S, Zeng H, Yang Z, Hung W-L, Wang B, Yang A. Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. Biotechnol Bioeng. 2023. doi:10.1002/bit.28492

64. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004;429: 92–96.

65. Herrgård MJ, Lee B-S, Portnoy V, Palsson BØ. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res. 2006;16: 627–635.

66. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessières P, Aymerich S, et al. Reconstruction and analysis of the genetic and metabolic regulatory networks of

the central metabolism of Bacillus subtilis. BMC Syst Biol. 2008;2: 20.

67. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2010;107: 17845–17850.

68. Liu L, Bockmayr A. Regulatory dynamic enzyme-cost flux balance analysis: A unifying framework for constraint-based modeling. J Theor Biol. 2020;501: 110317.

69. Sechkar K, Steel H, Perrino G, Stan G-B. A coarse-grained bacterial cell model for resource-aware analysis and design of synthetic gene circuits. Nat Commun. 2024;15: 1981.

70. Choudhury S, Narayanan B, Moret M, Hatzimanikatis V, Miskovic L. Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. Nat Catal. 2024. doi:10.1038/s41929-024-01220-6

71. Gonçalves E, Bucher J, Ryll A, Niklas J, Mauch K, Klamt S, et al. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. Mol Biosyst. 2013;9: 1576–1583.

Chapter 3 Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community

This chapter is based on (with minor edits) the following published paper: Qiu S, Zeng H, Yang Z, Hung W-L, Wang B, Yang A. Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. Biotechnol Bioeng. 2023;120: 2186–2198.

Sizhe Qiu, as the first author, contributed to the conceptualization of the research, developed the methodology, conducted data analysis and model construction, and produced the first draft of the paper. The experimental work on metagenome sequencing, growth measurement, and metabolomics was carried out by Hong Zeng and Zhijie Yang. Other listed authors contributed to the revision of the paper.

Summary

Genome-scale metabolic models and flux balance analysis (FBA) have been extensively used for modeling and designing bacterial fermentation. However, FBA-based metabolic models that can accurately simulate the dynamics of co-culture are still rare, especially for lactic acid bacteria used in yogurt fermentation. To investigate the fermentation kinetics and metabolic interactions of the yogurt bacterial community (*Streptococcus thermophilus* and *Lactobacillus delbrueckii subsp. bulgaricus*), this study built a dynamic metagenome-scale metabolic model integrated with constrained proteome allocation and incorporated the product inhibition of undissociated lactic acid into the proteome constraint. The model validation with reference experimental data demonstrated that the model could accurately predict bacterial growth, consumption of lactose and production of lactic acid, under the stress of accumulating

undissociated lactic acid. Then, the model was used to predict the impact of different initial bacterial inoculation ratios on acidification. The dynamic simulation demonstrated the mutual dependence of *S. thermophilus* and *L. d. bulgaricus* during the yogurt fermentation process. As the first dynamic metabolic model of the yogurt bacterial community, it provided a foundation for the computer-aided process design and control of the production of fermented dairy products.

Reporting the first research project in this PhD study, this chapter presents a computational model for the growth-coupled primary metabolism of lactic acid bacteria under the stress of accumulating undissociated lactic acid, and lays the groundwork of mathematical formulation of constrained proteome allocation specialized for lactic acid bacteria, which are used in Chapters 4 and 6.

3.1 Introduction

Yogurt is an important fermented dairy product, traditionally made by a starter culture composed of lactic acid bacteria (LAB), such as *Streptococcus thermophilus*, *Lactobacillus delbrueckii subsp. bulgaricus*, and *Lactobacillus acidophilus* [1]. In industrial yogurt production, fermentation process control, in terms of acidification and production of flavor and probiotic compounds, largely depends on the composition of the starter culture. The interactions of LABs affect the fermentation kinetics and thus influence the properties of the yogurt. For example, it was found that the co-culture of *Streptococcus thermophilus* and *Lactobacillus delbrueckii subsp. bulgaricus* could result in a higher productivity of lactic acid than that of *Streptococcus thermophilus* and *Lactobacillus acidophilus* [2]. Therefore, designing an optimal yogurt starter culture for desired yogurt properties is one of the primary engineering objectives for yogurt manufacturers.

To investigate and rationally engineer the yogurt starter culture in a more efficient and low-cost manner, a computational model is needed to simulate key variables, such as LAB

biomass levels and concentrations of critical compounds, in the fermentation process. There are mainly two types of models for simulating microbial growth and metabolism: differential-equation-based model and flux balance analysis (FBA)-based model. So far, there have been many attempts to use differential-equation-based models to simulate growth, substrate consumption and lactic acid production of LABs [3]. However, some of these models are too simplistic that only consist of Monod or extended Monod equations that empirically link microbial growth and substrate utilization [4–6], leaving the whole metabolic network as a “black box”. There also exist differential-equation-based “white box” models that capture the metabolic activity via a series of enzyme kinetic equations [7]. These models are typically costly to construct due to various enzyme kinetic mechanisms [8] and would require a large number of enzyme kinetic parameters that are difficult to obtain [9].

Alternatively, FBA-based metabolic models can avoid major shortcomings of differential-equation-based models. Firstly, genome scale metabolic models (GSMMs) can be easily reconstructed when annotated genomes are available [10]. Secondly, FBA does not require information on enzyme kinetic mechanisms and kinetic parameters (e.g. k_{cat}) [11]. Finally, the gene-protein-reaction relations in GSMMs allow the integration of multi-omics data, such as quantified proteomics [12]. Currently, several GSMMs for dairy-origin LABs have already been reconstructed [13–16], and a dynamic co-culture metabolic model for cheese starter culture involving those GSMMs has been built [17]. However, there is still a lack of metagenome-scale metabolic models [18] that can simulate the growth and metabolism of LAB co-cultures used in real industrial scenarios (as opposed to assumed ones). Furthermore, existing FBA models of LAB cultures cannot simulate unique inter-species interactions in yogurt fermentation.

With the aim to quantitatively model the fermentation kinetics and metabolic interactions of LABs in yogurt fermentation, this study built the first dynamic metagenome-scale metabolic model of major species identified in the yogurt starter culture, i.e., *Streptococcus thermophilus* (ST) and *Lactobacillus delbrueckii subsp. bulgaricus* (LB). In addition, constrained proteome

allocation, for the first time, was integrated into dynamic community-level FBA. Subsequently, we showed how the model can simulate the growth and metabolism of the ST/LB co-culture during yogurt fermentation with good accuracy. Finally, we explored the potential of the developed model in supporting the design and optimization of the yogurt fermentation process via simulating the impact of differential ST/LB inoculation ratio on the overall fermentation behavior.

3.2 Materials and methods

3.2.1 Sample preparation and fermentation conditions

To conduct the experiment, a commercial yogurt fermentation starter (YoFlex Premium 1.0, CHR Hansen, Denmark) of pack size 250 U was used. A growth medium was prepared by adding Fonterra whole milk powder to 1 liter of water and stirring for 1 hour to achieve a 12% (w/w) concentration. The growth medium was then heated to 95℃ for 5 minutes and cooled to 42℃ for the inoculation of the fermentation starter. The fermentation process was carried out at a temperature of 43℃ for 5 hours until the system reached the endpoint at pH = ~4.5.

3.2.2 Cell counting and pH measurement

To quantify the growth kinetics of bacteria during fermentation, viable cell counting was used as a common practice [19–21]. Lactic acid bacteria were isolated from samples taken at different time points (0 min, 30 min, 60 min, 90 min, 105 min, 120 min, 150 min, 180 min, 240 min, 300 min) using the dilution plate method (1:10). The isolated bacteria were then cultured on MRS agar (AOBOX, China) and M17 agar (AOBOX, China) under 37℃ anaerobic conditions for 48 hours to count the colony-forming units (CFU). However, it should be noted that due to the inactivity of bacterial cells under low temperatures in storage, viable cell counting reflected an

additive curve of cell activation and cell growth at the lagging and exponential phase, resulting in inaccurate quantification of growth kinetics. Hence, the validation of dynamic simulation in section 3.3 used reference experimental data from the work of [2] (**see section 3.3.3 and Appendix A, section A.1.2**).

The pH change during fermentation was measured using a fermentation monitor (iCinac, AMS, France) with an Inlab Smart pro-ISM detection electrode (Mettler Toledo, Switzerland). Datapoints were collected at a frequency of three times per minute. Prior to measurement, the electrode was calibrated with a standard pH calibration solution.

3.2.3 Metabolomics quantification

The concentrations of lactic acid and lactose were measured at different time points during fermentation using two different HPLC systems: HPLC-DAD (Agilent 1260, USA) with a C18 column (250 mm ×4.6 mm, Agilent, USA) and HPLC-RI (Waters 2695-2414, USA) with a Zorbax NH2 column (250 mm × 4.6 mm, Agilent, USA). For lactic acid, samples were centrifuged at 8000×g for 15 min in 4 ℃, and the supernatant, after filtered with 0.22 μm pore size, was measured by HPLC. The injection volume was 10 μL, and the flow rate of the mobile phase (A= 0.1% phosphoric acid, B= acetonitrile) was 0.5 mL/min. The column temperature was set at 30 ℃ and the detection wavelength was 220 nm. The gradient elution procedure was as follows: 0-4 min, 90%-60% elution A; 4-6 min, 60%-50% elution A; 6-9 min, 50%-80% elution A; 9-15 min, 80%-90% elution A. For lactose, samples were first mixed with 12% TCA and centrifuged at 8000×g for 15 min in 4 ℃. The supernatant, after filtered with 0.22 μm pore size, was measured by HPLC. The injection volume was 10 μL, and the flow rate of the mobile phase was 1 mL/min, with the ratio of acetonitrile to water fixed at 70:30 (v/v). The column and detector temperature were set at 40 ℃. Lactic acid and lactose concentrations measured in this study were not used to valid the dynamic simulation, as explained in section 2.2. They were only

used to validate predicted lactic acid yield ratios from lactose consumed (**section 3.3.2**) and measured lactic acid concentrations were used to fit a linear function for milk pH (**section 3.2.4.3**).

The concentrations of free amino acids in 12%(w/w) reconstituted milk were measured using UPLC (I-Class, Waters, USA) equipped with a triple quadrupole mass spectrometer (Xevo TQ-S micro, Waters, USA). To prepare the samples, 50% ethanol was added to the milk and then shaken. Next, 10 µL of the shaken samples were mixed with 10 µL of deionized water, 5 µL of D-Norleucine (internal standard), and 40 µL of 0.1% formic acid in isopropanol. The samples were centrifuged for 10 min at 10000×g in 4 ℃. The resulting supernatant was then added to boric acid buffer and AccQ Tag solution (Kairos, USA) for derivatization, and filtered using a 0.22 µm pore size filter. The column was UPLC HSS T3 (1.7 µm, 2.1 mm × 150 mm, Waters, USA). The injection volume was 5 µL, and the flow rate of the mobile phase (A= 0.1% formic acid, B= acetonitrile) was 0.5 mL/min. The column temperature was set at 50 ℃. The mass spectrometer was operated in electrospray ionization (ESI) mode with an ionization energy of 1.5 kV. The cone voltage was set to 20 V, the desolvation temperature was maintained at 600 ℃, the desolvation gas flow rate was set to 1000 L/h, and the cone gas flow rate was set to 10 L/h. The gradient elution procedure was as follows: 0-2.5 min, 96%-90% elution A; 2.5-5 min, 90%-72% elution A; 5-6 min, 72%-5% elution A; 6-7 min, 5%-5% elution A; 7-9 min, 5%-96% elution A.

3.2.4 Dynamic metabolic model from the metagenome of the yogurt starter culture

Building the dynamic metagenome-scale metabolic model of the yogurt starter culture comprised two steps: 1. from metagenome to annotated protein coding genes of major species; 2. from coding genes to GSMMs (**Figure 3.1**). With the resulting GSMMs, dynamic flux balance

analysis (dynamic FBA, or dFBA) was implemented to simulate bacterial growth and metabolism and predict the change on fermentation behavior by perturbation to initial co-culture composition.

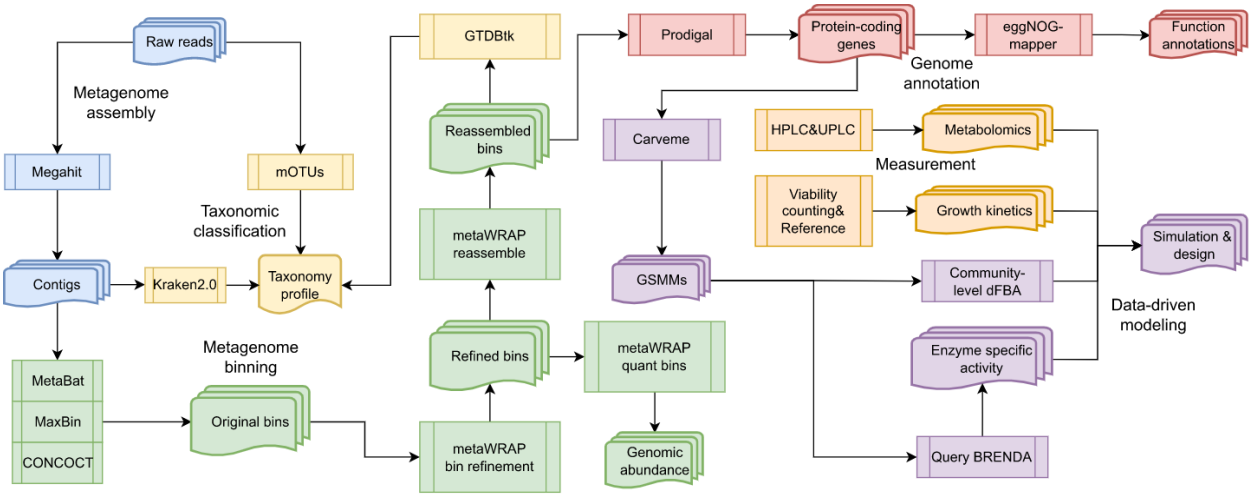

Figure 3.1. The workflow diagram of constructing a dynamic metabolic model for the yogurt starter culture based on metagenomic analysis.

3.2.4.1 Metagenome assembly, binning, and annotation

DNA extraction of the yogurt starter culture followed the procedure in the metagenomic study of cereal vinegar microbiota by [22]. Three parallel DNA samples were sequenced by Illumina PE150 platform, and the raw data was filtered for high quality reads by removing adapter overlaps, reads with quality value lower than 38 and length lower than 350 bp. After the quality control step, the cleaned data was assembled by MEGAHIT [23] with default parameters.

MetaBAT2 [24], MaxBin2 [25] and CONCOCT [26] were used for species-level metagenome binning. Then, bins from three different tools were refined to produce “best bins” based on default contamination and completion cutoffs using the refinement module in metaWRAP [27]. The reassembly module of metaWRAP mapped metagenomic reads to refined bins and reassembled reads into reassembled bins. The final outputs are metagenome assembled genomes (MAGs) of high quality, assessed by CheckM [28]. Those bins’ genomic

abundances in each sample were then computed by the Quant_bin module in metaWRAP. Qaunt_bin uses Salmon [29] to align and count reads from each sample to estimate the abundance of each contig in each sample.

The taxonomy profile was computed with three different tools. Using reference genomes, mOTUs identified species and computed relative abundances with unassembled DNA reads [30]. When the metagenome was assembled, KRAKEN 2.0 [31] assigned taxonomic labels to DNA reads and generated taxonomic profiles for all samples. Once high-quality bins were generated, GTDB-Tk [32] classified the taxonomy of each bin. Species identified consistently by all taxonomic classification tools were taken as dominant species in the yogurt starter culture.

Protein-coding genes were predicted for each MAG by Prodigal [33]. The functional annotation of protein-coding genes using eggNOG-mapper [34] was carried out with KEGG [35] and CAZy databases [36], providing information on cellular pathways and carbohydrate active enzymes .

3.2.4.2 Reconstruction of genome-scale metabolic models

Non-redundant protein sequences were filtered from concatenated protein sequences of three parallel samples, and used as the input for automatic GSMM reconstruction by CarveMe [37]. In addition to protein sequences, other inputs were chemically defined media (CDM) for gap filling and the universal bacterial template. The CDM was adopted from the co-culture metabolic model of cheese starter in [17]. The universal bacterial template used for reconstruction was specialized for gram-positive bacteria.

For refinement of GSMMs, the growth function (biomass synthesis, v_{Growth}) should be set as species-specific (**see Appendix A, section A.1.1**). The stoichiometric coefficients in ST's growth function were adjusted based on measured biomass composition [14], and the growth associated ATP requirement was adopted from the GSMM of *Streptococcus thermophilus* CH8 [38]. Due to data limitation, the biomass composition in LB's growth function was set according

to the default of gram positive bacteria [39], and the growth associated ATP requirement was adopted from the GSMM of Lactobacillus plantarum [40]. Based on functional annotation of protein-coding genes (**section 3.2.4.1**), reactions included erroneously were removed and missing reactions were added (**see Appendix A, section A.1.1**). In addition, to characterize the proteolysis activity in the co-culture system, a self-defined reaction named "**PROLYSIS**" that utilizes the casein peptide was added to the GSMM:

$$PROLYSIS: Casein\ peptide(C141H270N35O73S1) \rightarrow \alpha\ alanine + \beta\ arginine + \ldots + \varepsilon\ valine$$

Stoichiometric coefficients ($\alpha, \beta, \ldots$) in the reaction were approximated from fractions of amino acid in the casein protein of cow milk [41] (**see Appendix A, section A.1.1**). The boundary of the flux through casein peptide utilization in each GSMM was set based on the proteolytic activity of each species (**see Appendix A, section A.1.3 and Figure A.3B**).

3.2.4.3 Dynamic flux balance analysis and proteome allocation constraint

To simulate the growth of bacteria and the production of target metabolites in time, dynamic FBA (dFBA) was adopted, as a combination of FBA and differential-equation-based dynamic system modeling [42]. The code of model implementation can be found in https://github.com/SizheQiu/MetaStLbCom. The intracellular metabolic fluxes of a species j were computed by parsimonious FBA, maximizing the growth rate ($v_{j,growth}$) while minimizing the total sum of individual fluxes ($v_{j,i}$ for reaction *i* in species *j*), based on the assumption that the cell minimizes the use of enzyme catalyzed reactions due to the limited cellular recourse (Eqs. 3.1, 3.2), subject to mass conservation (Eq. 3.3), in which S_j represented the stoichiometric matrix. The concentration change of metabolites and biomass in the extracellular space was modeled by differential equations to account for biomass accumulation (Eq. 3.4) and exchange fluxes of the metabolite e from major species (Eq. 3.5). The extracellular metabolites in Eq. 3.5 involved in the dynamic simulation were lactose, formic acid, lactic acid, acetic acid,

and 20 essential amino acids. The community-level dynamic FBA was performed on separate species-level GSMMs of dominant species identified in metagenomic analysis (**section 3.3.1**).

$$Maximize\ v_{j,Growth}\ \forall\ j\ \in\ major\ species\ (Eq.3.1)$$

$$Minimize\ \sum_{i} v_{j,i}\ (Eq.3.2)$$

$$S_j v_j = 0\ (Eq.3.3)$$

$$\frac{d[Biomass_j]}{dt} = v_{j,Growth}[Biomass_j]\ (Eq.3.4)$$

$$\frac{d[M\_e]}{dt} = \sum_{j} v_{j,EX\_M\_e}[Biomass_j]\ (Eq.3.5)$$

Furthermore, proteome allocation was implemented to constrain reaction fluxes of the central carbon metabolism [43,44]. Proteome was divided into sectors of inflexible housekeeping (Q), anabolism (A), transportation (T), catabolism (C), and the free sector. The upper bound of all flexible sectors combined was assumed to be 50% of the total proteome (Eqs. 3.6, 3.7). The proteome cost on each reaction was computed as the ratio of the flux to the multiplicative product of enzyme activity, a_i,and saturation degree, σ_i (Eq. 3.8). The enzyme saturation degrees (σ_i), except for that of the lactose transporter, are unknown and were set to 0.5, assuming they are similar to glycolytic enzymes in *E. coli* which were reported to be mostly half-saturated [45]. For the milk environment, the saturation degree of the lactose transporter (σ_{LT}) was assumed to be 1 as it was considered to be fully saturated by abundant lactose; when the concentration of lactose was low, σ_{LT} was assumed to follow michaelis-menten kinetics, $\sigma_{LT} = \frac{[Lactose]}{[Lactose]+K_m}$. The activities of the ribosome for the anabolism sector ($a_{ribosome}$) and acid exportation for lactic and acetic acids (a_{acid_T}) were collected from the literature [44,46]. Other enzyme activity values were obtained from BRENDA Enzyme Database [47] (**see Appendix A, Table A.5**). For the uptake of amino acids, the flux was constrained by michaelis-menten equation ($\frac{v_{max}[S]}{[S]+K_m}$). The v_{max} of amino acid uptake in ST was obtained from average amino acid

uptake upper bounds in [17], but for LB, the v_{max} was estimated based on a reported growth rate value on a defined medium (**see Appendix A, section A.1.3 and Figure A.3A**). The michaelis-menten constant K_m was set based on average parameter values obtained from BRENDA and SABIO-RK [48] (**see Appendix A, Table A.5**).

$$\frac{P_{total}}{1\ gDW} = p_Q(50\%\frac{P_{total}}{1\ gDW}) + p_C + p_A + p_T + p_{Free}\ (Eq.3.6)$$

$$0 < p_C + p_A + p_T \leq 50\%\frac{P_{total}}{1\ gDW}\ (Eq.3.7)$$

$$p_C = \sum_i \frac{v_i}{\sigma_i a_i};\ p_A = \frac{v_{Growth}}{a_{ribosome}}; p_T = \frac{v_{LT}}{\sigma_{LT} a_{LT}} + \frac{v_{acid\_T}}{\sigma_{acid\_T} a_{acid\_T}}\ (Eq.3.8)$$

The activity of lactose uptake incorporated inhibition by undissociated lactate (LacH), the product of glycolysis under anaerobic conditions. The exponential decay equation (Eq. 3.9) to model the inhibition of lactose transporter activity was adopted from Vereecken & Van Impe, 2002 [49] and Aghababaie et al., 2015 [50]. The minimal activity of lactose transport a_{LT}^{min} was set to maintain the growth rate at the stationary phase when pH is around 4.5 (**see Appendix A, section A.1.3 and Figure A.3C**). The concentration of undissociated lactate was computed using Henderson-Hasselbalch equation (Eq. 3.10), $pK_a = 3.86$, and pH was approximated as a linear function of lactic acid concentration, $pH = C_1[Lac] + C_2$ (**see Appendix A, Figure A.1**), with measured lactic acid concentrations and pH (**sections 3.2.2 & 3.2.3**). Such inhibition coefficient, $e^{(-k_{LacH}[LacH])}$, was also applied to amino acid uptake rate and casein peptide utilization rate, and similar to a_{LT}^{min}, the minimal rate, v_{min}, was set (**see Appendix A, section A.1.3**).

$$a_{LT} = max(\ a_{LT}^0 e^{(-k_{LacH}[LacH])}, a_{LT}^{min}\ )\ (Eq.3.9)$$

$$[LacH] = \frac{[Lac]}{10^{pH-pK_a}}\ \text{(Eq. 3.10)}$$

3.3 Results

3.3.1 Metagenomic analysis of the yogurt starter culture

Taxonomic classification by different tools on assembled metagenomes of three samples of the yogurt starter showed that *Streptococcus thermophilus* (ST) and *Lactobacillus delbrueckii subsp. bulgaricus* (LB) were two major species that contributed to more than 95% of the overall taxonomy abundance. The genomic abundance ratio of ST and LB in the microbial community was approximately 100:1 (**Figure 3.2A**). Detailed taxonomic profiles computed by mOTUs and KRAKEN-2.0 can be found in **Appendix A, Table A.3**. The finalized metagenome assembled genomes (MAGs), obtained through binning, bin refinement and reassembly, showed high completeness, low contamination and good continuity (**see Appendix A, Figure A.2**). In MAGs of ST and LB from three parallel samples, 2499 and 1801 non-redundant protein-coding genes were identified (**Figure 3.2B**), and functionally annotated. KEGG pathway annotation indicated that ST contained a higher number of genes for amino acid metabolism compared to LB (**Figure 3.2C**). In the carbohydrate-active enzyme annotation, more carbohydrate-active enzymes were identified in ST's genome, and a significant enrichment of CBM41(Carbohydrate-Binding Module Family 41) was observed (**Figure 3.2D**), which was a module of approximately 100 residues found primarily in bacterial pullulanases [51].

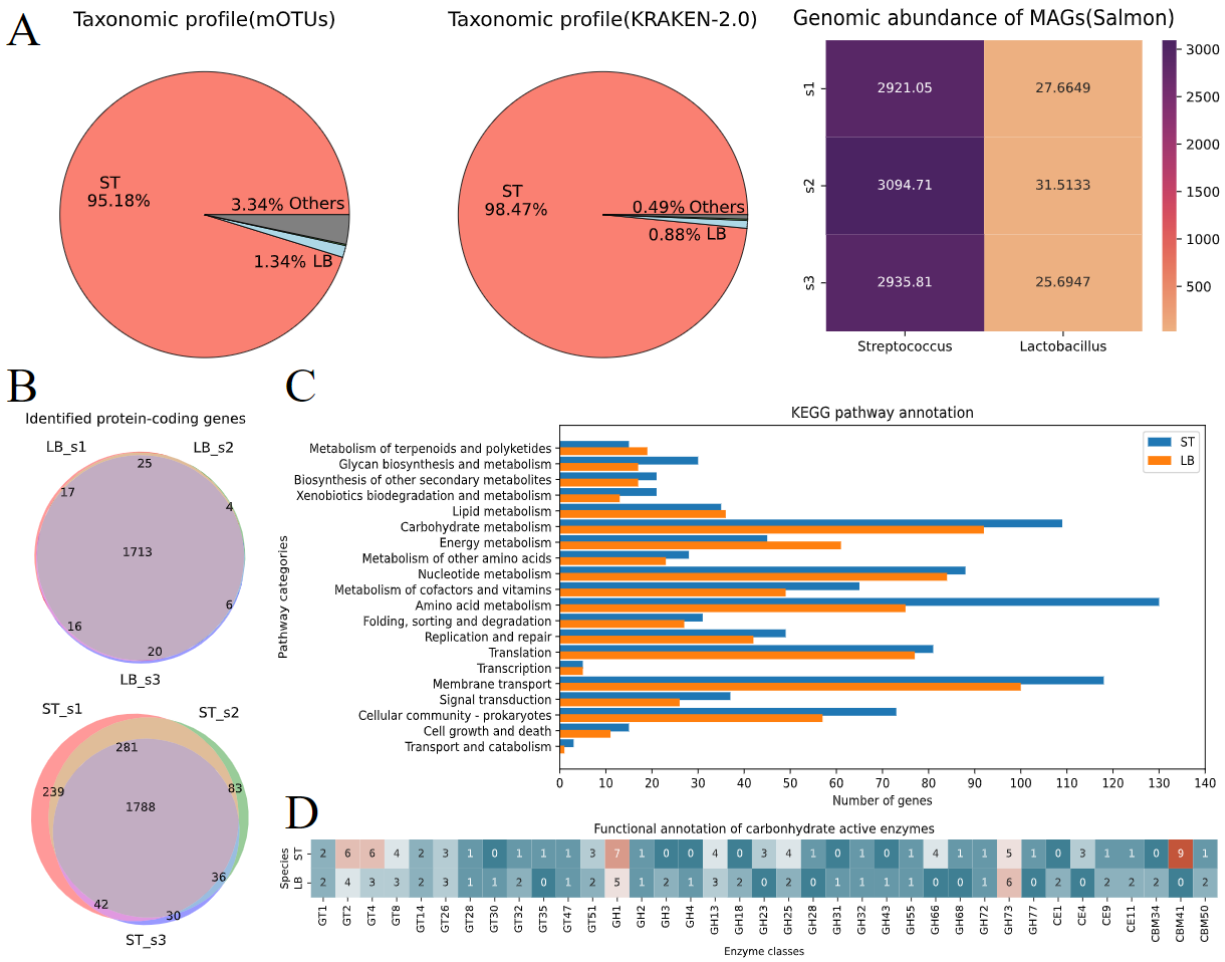

Figure 3.2. Results of metagenomic analysis. (A)Taxonomic classification and genomic abundance, s1-s3: yogurt starter samples 1 to 3. (B) Pan-genome of identified protein-coding genes from three parallel samples. (C) KEGG pathway and (D) carbohydrate active enzyme annotations for both species. ST: *Streptococcus thermophilus*; LB: *Lactobacillus delbrueckii subsp. bulgaricus*.

3.3.2 Reconstruction of genome-scale metabolic models with proteome allocation constraints for the yogurt starter culture

Two individual GSMMs for the dominant species in the yogurt starter culture, i.e., ST and LB, were reconstructed (**see Section 3.2.4.2**). In reconstructed GSMMs, reactions assigned

with gene-protein-reaction relations were around 60% in both models. About 25% gap-filled reactions, without genomic basis, were added into the model for a complete metabolic network and the rest were boundary reactions for metabolite exchange (**Figure 3.3A**). Classical FBA without additional constraints can only predict an untightened solution space of metabolic fluxes, but fails to account for metabolic capacities of enzymes [52] and global regulation of proteome sectors [43] as described in section 2.4.3. Therefore, in this study, proteome allocation constraints were integrated in classical FBA to explain the preference of lactic acid production by LABs that is energetically less favorable than acetic acid production. To implement proteome allocation constraints, enzyme activities of reactions in central carbon metabolism were mapped to the reconstructed metabolic network (full names of reactions can be found in **Appendix A, Table A.4**), and proteome costs for producing 1 unit of flux of lactic acid and acetic acid from pyruvate could be computed (**see Appendix A, Table A.6**). The proteome cost of lactic acid production per unit flux is $0.0071\ mg\ E/gDW$, same for ST and LB. For acetic acid, the proteome cost is $0.1218\ mg\ E/gDW$ in ST, but gets much larger in LB due to the lack of pyruvate formate lyase (PFL), $0.6843\ mg\ E/gDW$. In short, the proteome cost of lactic acid production in LABs is much smaller than that of acetic acid production, though the latter pathway has a higher energy (ATP) yield.

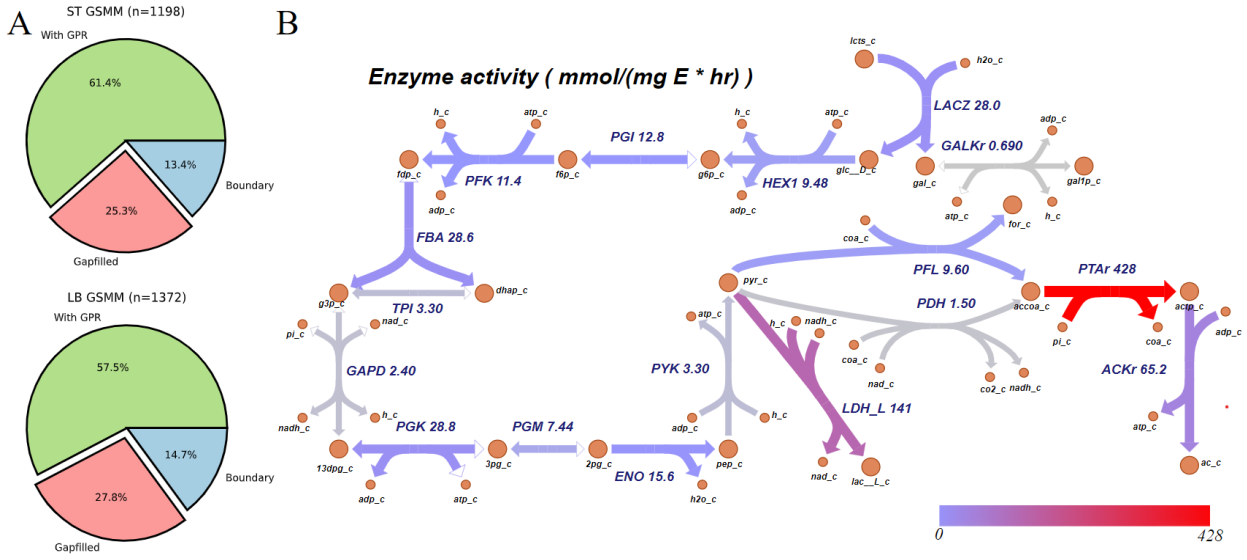

Figure 3.3. Properties of reconstructed GSMMs of the yogurt starter culture. (A) Status of metabolic reaction curation, Green: reactions with gene-protein-reaction rules; Red: reactions that were gap-filled; Blue: exchange reactions at boundary. (B) Enzyme activities mapped to the central carbon metabolic network (full names of reactions can be found in SI), only ST has pyruvate formate lyase (PFL) reaction. Big nodes: primary metabolites; small nodes: cofactors. ST: *Streptococcus thermophilus*; LB: *Lactobacillus delbrueckii subsp. bulgaricus*.

To check the validity of reconstructed GSMMs, FBA with proteome allocation constraints was performed for ST and LB on complete CDM [38] and MPL medium [53] (**Figure 3.4AB**). For ST, the growth rate reported by Rau et al., 2022 on complete CDM was 0.98 hr^{-1} [38], and proteome-constrained FBA, with a fixed upper bound of amino acid uptake rate from [17], predicted a close value, 1.2 hr^{-1}. Due to lack of similar amino acid uptake rate data measured or estimated for LB, the upper bound was estimated in this work by fixing LB's growth rate on MPL medium, reported to be 0.7 hr^{-1} [53] (**see Appendix A, section A.1.3 and Figure A.3A**). The predicted lactic acid yield ratio from lactose consumed (mol lactate/mol lactose) were 1.6919 for ST and 1.9305 for LB, which was consistent with the experimental measurement (**see Appendix A, Table A.1**) as well as with previous studies [17,38,54]. Due to the lack of pyruvate formate lyase (PFL) (**Figure 3.3B**), LB was predicted to have no ability to produce formic acid as well as a much smaller yield of acetic acid compared with ST. Apart from predicting the formation of lactic, acetic and formic acids, proteome-constrained FBA also predicted the secretion fluxes for various flavor compounds, including 4-hydroxy-benzyl alcohol (4hba) and succinic acid (succ) for ST, and 2-methylbutanoic acid (2mba), 2-methylpropanoic acid (2mpa), 3-methylbutanoic acid (3mba) and 2-oxobutanoate (2obut) for LB (**Figure 3.4AB**).

The predicted responses to the change in the presence of methionine and formic acid by ST and LB show their essential nutrient requirement for cellular growth and potential metabolic interactions (**Figure 3.4CD**). The growth rate of ST increases with the increasing concentrations

of methionine (**Figure 3.4C**), whereas other amino acids have little impact (**see Appendix A, Figure A.4**), suggesting that the GSMM of ST predicts that ST in the starter culture is auxotrophic for methionine and prototrophic for all other 19 essential amino acids. On the other hand, the GSMM of LB reveals its auxotrophy for numerous amino acids (**see Appendix A, Figure A.5**), which is consistent with the finding in previous studies that Lactobacillus species developed proteolytic ability to compensate for their amino acid auxotrophy [55]. The estimated amino acid auxotrophy in ST and LB by varying amino acid supply is consistent with the auxotrophy calculated by ReFramed (https://github.com/cdanielmachado/reframed) [56]. In the *in-silico* milk environment that has no purine (adenine, guanine and xanthine) to mimic the nutrient composition of actual reconstituted milk used for yogurt fermentation, LB's growth is promoted by the increase of the concentration of formic acid (**Figure 3.4D**). In contrast, when purines are supplied, formic acid has little influence on LB's growth rate, which validates the previous finding that LB requires formic acid to synthesize DNA/RNA materials in environments with low levels of purines [57]. Succinctly, the potential metabolic interactions of ST/LB community in the yogurt starter culture inferred by the GSMMs reconstructed in this study can be summarized as the following: ST provides formic acid for LB to synthesize DNA/RNA materials, and LB utilizes casein proteins to supply methionine to ST; they both consume lactose and produce lactic acid, which in turn inhibits their growth [49] (**Figure 3.4E**).

The addition of proteome constraints allows the model to account for the metabolic switch from acetic acid production to lactic acid production in ST when its growth rate increases with the increase of carbon source concentration, which has been demonstrated in Regueira et al., 2021 with a generic LAB-GSMM that can respond to the change of glucose concentration [44] (**Figure 3.4FG**). The proteome costs allocated to acetate and lactate productions are computed as the proteome cost of pyruvate formate lyase (PFL), pyruvate dehydrogenase (PDH), phosphotransacetylase (PTAr) and acetate kinase (ACKr), and that of lactate dehydrogenase (LDH) (**Figure 3.4H**). In general, only two moles of lactic acid can be produced

from the fermentation of one mole of lactose. The predicted ratio of acetic acid produced to lactose consumed sometimes exceeds 2, which results from the metabolism of other nutrients in the *in-silico* culture medium (**see Appendix A, Figure A.6**). With LB's GSMM, proteome constrained FBA could not demonstrate growth rate triggered metabolic switch from acetic acid production to lactic acid production due to the much larger protein requirement of converting pyruvate to acetyl-CoA by solely PDH in LB (**Figure 3.3B**) and the biosynthetic requirement of fatty acids (**see Appendix A, Figure A.7**). When the uptake of lactose provides enough carbon flux that exceeds the biosynthetic requirement of fatty acid, a surplus flux will be predicted to go through acetic acid production (**see Appendix A, Figure A.7**).

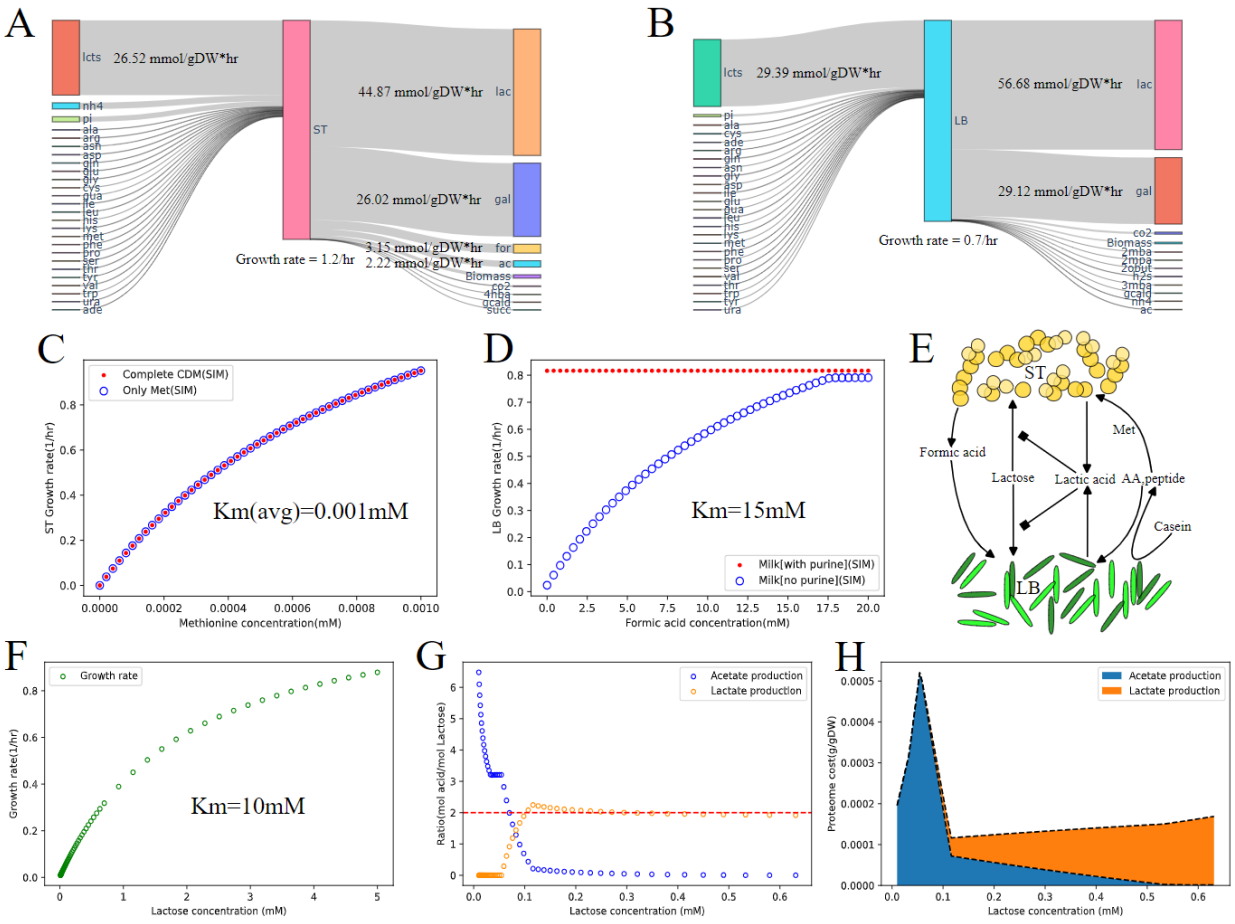

Figure 3.4. Model predictions of the metabolic fluxes on the chemically defined medium. (A, B) Predicted uptake and secretion fluxes of (A) ST on complete CDM and (B) LB on MPL medium.

(C, D) Assessment of (C) the requirement of methionine by ST ($v_{max}(average) = 0.2\ mmol/gDW * hr$ [58], $K_m(average) = 0.001mM$ [17]) and (D) the requirement of formic acid by LB ($v_{max} = 0.3\ mmol/gDW * hr, K_m = 15mM$ [59]). SIM: simulation shown in dotted lines. (E) Potential metabolic interactions of ST/LB co-culture. (F-H) The response of ST to the increase of lactose concentration ($K_m = 10mM$ [60]): (F) predicted growth rates, (G) fluxes through acetic/lactic acid productions, and (H) proteome costs allocated to biosynthetic pathways of acetic and lactic acids.

3.3.3 Dynamic simulation of growth kinetics and metabolism of ST/LB co-culture

The dynamic simulation of yogurt fermentation was performed using the initial conditions of the co-culture experiment in Oliveira et al., 2012 [2]. The simulation was validated by comparison with reference experimental data of bacterial biomass, lactose and lactic acid concentrations, also from Oliveira et al., 2012 [2] (**see Appendix A, section A.1.2**). The accuracy of the simulation was assessed and demonstrated by R-squared values, which were all around 0.8 (**Figure 3.5A-D**).

Overall, the model, without parameter re-calibration using the reference experimental data, adequately captured the initiation of the exponential growth phase and the transition to the stationary phase for both ST and LB, as well as the protocooperation between two species (**Figure 3.5AB**). The accumulation of formic acid (produced by ST) initiates the exponential growth of LB by activating its synthesis of purines, whose natural concentration is too low to support the growth of LB in the milk environment. In return, the activation of growth and metabolism of LB with strong proteolytic activity enhances the growth rate of ST by supplementing methionine, which is also limited in the milk environment (**Figure 3.5EF**). With

the accumulation of lactic acid, the growth rates of ST and LB are both progressively inhibited. Finally, their growths are halted, as shown by the stationary phase (**Figure 3.5A-C**).

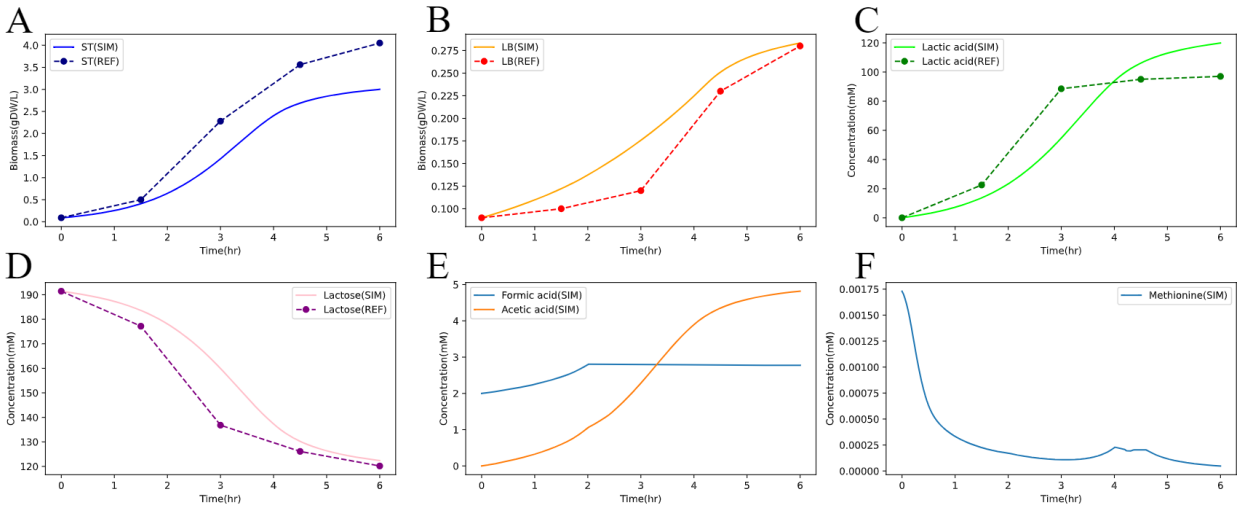

Figure 3.5. Dynamic simulation results of ST/LB co-culture ($I.C.[Biomass_{ST}] = [Biomass_{LB}] = 0.09\ gDW/L$). Comparison between simulated and experimental growth kinetics of ST, R2=0.79 (A), and LB, R2=0.86 (B). Comparison between simulated and experimental concentration profiles of lactic acid, R2=0.78 (C), and lactose, R2=0.85 (D). Simulated concentration profiles of formic acid, acetic acid (E), and methionine (F). SIM: simulation shown in solid lines, REF: reference data shown in dotted lines.

3.3.4 Prediction of the impact of different initial ST/LB inoculation ratios on the fermentation behavior

Perturbations on the initial inoculation ratio of ST and LB of the yogurt starter culture was conducted to investigate its impact on the fermentation behavior. The total initial biomass concentration of the co-culture was fixed at 0.18 gDW/L, same as the initial condition (I.C.) setting in section 3.3. The simulated growth curves of ST and LB from different initial ST/LB inoculation ratios demonstrate the mutual dependence of ST and LB in the co-culture (**Figure 3.6A**). When the initial ST:LB inoculation ratio is modulated to increase from 1 to 10, the

proteolytic activity of the microbial community becomes weaker due to the decrease of the initial biomass concentration of LB, and correspondingly, the growth of ST is also reduced due to lowered supply of methionine; When the initial ST:LB inoculation ratio is set to decrease from 1 to 0.1, the lowered productivity of formic acid from ST makes LB enter the exponential phase slower than that of the case with initial ST:LB = 1 (**Figure 3.6A**). In addition, the simulation shows that, given a certain range of initial ST:LB ratios (in this case, 0.1~10), ST will eventually become the dominant species (**Figure 3.6B**), which agrees with the previous study on the rods (LB) to cocci (ST) ratio in cheese fermentation [61].

The predicted average acidification rates (lactic acid concentration/total fermentation time) in 5 hours by different starter culture compositions rank as follows: ST:LB = 1 > ST:LB = 2 > ST:LB = 5 > ST:LB = 10 > ST:LB = 0.1 > ST:LB = 100 > ST:LB = 0.01 (**Figure 3.6CD**). At the early stage of yogurt fermentation (0-3hr), starter cultures with the initial ST:LB ratio larger than 1 have more than twice lactic acid production rates, in contrast to starter cultures with the initial ST:LB ratio lower than 1; Later (3-5hr), the lactic acid production rate of the starter culture with initial ST:LB = 0.1 catches up (**Figure 3.6D**). To sum up, initial ST:LB = 1 is predicted to be optimal for lactic acid production in 5 hours if the initial total biomass is fixed at 0.18 gDW/L, and the acidification will generally be faster initially when ST is the dominant species in the co-culture. Based on predicted acidification kinetics, the starter culture composition can be designed for targeted acidification patterns, for example, ST:LB ratio = 0.1 is suitable for slow acidification first but fast acidification later.

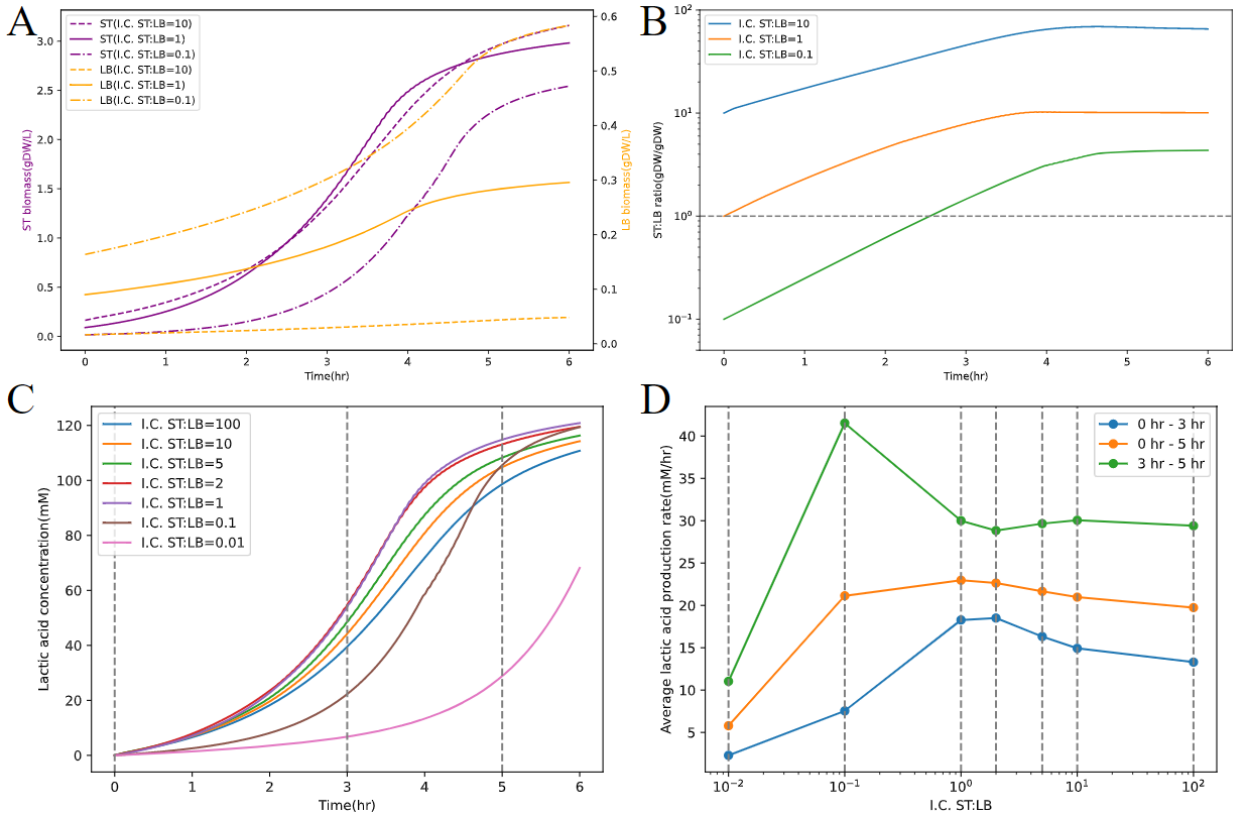

Figure 3.6. Predicted growth kinetics of ST/LB co-culture and lactic acid production with different initial ST:LB inoculation ratios. (A) Biomass concentration profiles of ST and LB. (B) Bacterial community composition dynamics: the change of ST/LB ratio in time. Purple: ST; orange: LB. Dashed line: I.C. ST:LB = 10; solid line: I.C. ST:LB = 1; dot dash line: I.C. ST:LB = 0.1. (C) Lactic acid concentration profiles for different initial ST/LB inoculation ratios. (D) The assessment of acidification rate for different initial ST/LB inoculation ratios with lactic acid levels at early stage (T=3hr), and later stage (T=5hr).

3.4 Conclusions and discussion

Overall, this work presented the reconstruction of GSMMs for the yogurt starter culture (i.e., the co-culture of ST and LB) based on metagenomic analysis and provided a dynamic metagenome-scale metabolic modeling approach for simulating the microbial growth and

metabolism during the yogurt fermentation process. Although community-level FBA has already been used in a few scenarios to simulate growth and metabolism [18,62,63], the model in this study, for the first time, integrated constrained proteome allocation into dynamic community-level FBA. Different from direct integration of gene expression data into FBA, such as the work of Blasche et al., 2021 [64], the proteome allocation constraint aims to capture the process of redistribution of proteome resources to each cellular pathway at different growth stages; it does not directly incorporate gene expression data into the model but tightens the solution space of fluxes by setting a global constraint on the metabolic capacity. Also, this model introduced feed-back inhibition function on the enzyme activity, previously used in modeling growth kinetics of LABs [49], and effectively simulated the product inhibition by undissociated lactic acid on the growth of LABs (**section 3.3.3**).

The proposed dynamic model quantitatively demonstrated the metabolic mutual dependence of ST and LB in the milk environment [2,62], and provided confirmation of such ecological interaction with data-driven modeling. As shown in single species FBA (**Figure 3.4CD**), the uptake fluxes of formic acid and methionine were essential for the growth of LB and ST, respectively. Subsequently, the dynamic simulation (**Figure 3.5**) and the perturbation on ST/LB initial ratios (**Figure 3.6**) further elucidated the mutual dependence between those two species in the milk environment through cross-feeding of formic acid and methionine.

Despite the model showed good predictive accuracy in bacterial growth and metabolism (**sections 3.3.2 & 3.3.3**), it had limitations in mainly three aspects: (1) MAG-derived GSMMs of ST and LB lacked accurate strain-specific biomass compositions and growth associated ATP requirements; (2) the model could not yet predict fluxes through the biosynthesis of important secondary metabolites, such as methyl ketones, and the probiotic exopolysaccharides that are often produced in yogurt fermentation; (3) no mechanistic representation of regulatory activities was included in the current model. For instance, acidification by ST was previously found to be stimulated by formic acid, casitone, pyruvic acid, folic acid, and polysorbate 20 [65]. Mono-

culture of dominant ST and LB strains separated from the starter culture will be needed to approximate the growth associated ATP requirement from the carbon source utilized [40], and gas chromatography/mass spectrometry or HPLC can be used to quantify major components in cellular biomass, i.e. protein, DNA, RNA, lipids and glycogen [66,67]. To resolve the other two limitations, the proposed next step is to further refine the established GSMMs by manually adding the biosynthetic pathways of secondary metabolites of interest and implement dynamic regulatory FBA (rFBA) [68] with gene expression regulatory information obtained from meta-transcriptome profiling. For the current model, the failure to predict active fluxes towards the formation of diacetyl and acetoin was not caused by the lack of pathway reconstruction, but the lack of a constraint to divert the downstream metabolic flux to those biomass-independent products.

Despite that several limitations remain to be overcome, the genome-scale metabolic reconstructions and the dynamic community-level FBA model for the classical yogurt starter culture of ST and LB presented in this work have been shown to have the potential to offer an efficient tool to guide engineering decisions in fermentation processes, which could be used to address issues such as the optimal initial biomass ratio of ST and LB to maximize the rate of acidification or possibly other process targets (e.g., flavor formation). Furthermore, the two-species model can be expanded by including GSMMs of other probiotic bacteria (e.g., bifidobacterium [69]), to simulate the fermentation dynamics of more complex starter cultures.

References

1. Mohammadi R, Sohrabvandi S, Mohammad Mortazavian A. The starter culture characteristics of probiotic microorganisms in fermented milks. Eng Life Sci. 2012;12: 399–409.

2. Oliveira RP de S, Torres BR, Perego P, Oliveira MN de, Converti A. Co-metabolic

models of Streptococcus thermophilus in co-culture with Lactobacillus bulgaricus or Lactobacillus acidophilus. Biochem Eng J. 2012;62: 62–69.

3. Bouguettoucha A, Balannec B, Amrane A. Unstructured models for lactic acid fermentation-a review. Food Technol Biotechnol. 2011;49: 3.

4. Bâati L, Roux G, Dahhou B, Uribelarrea J-L. Unstructured modelling growth of Lactobacillus acidophilus as a function of the temperature. Math Comput Simul. 2004;65: 137–145.

5. Youssef CB, Ben Youssef C, Goma G, Olmos-Dichara A. Kinetic Modelling of Lactobacillus casei ssp. rhamnosus Growth and Lactic Acid Production in Batch Cultures Under Various Medium Conditions. Biotechnology Letters. 2005. pp. 1785–1789. doi:10.1007/s10529-005-3557-0

6. Vázquez JA, Murado MA. Unstructured mathematical model for biomass, lactic acid and bacteriocin production by lactic acid bacteria in batch fermentation. Journal of Chemical Technology &. 2008. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jctb.1789

7. Foster CJ, Wang L, Dinh HV, Suthers PF, Maranas CD. Building kinetic models for metabolic engineering. Curr Opin Biotechnol. 2021;67: 35–41.

8. Ulusu NN. Evolution of Enzyme Kinetic Mechanisms. J Mol Evol. 2015;80: 251–257.

9. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011;50: 4402–4410.

10. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biol. 2019;20: 158.

11. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248.

12. Bakker BM, van Eunen K, Jeneson JAL, van Riel NAW, Bruggeman FJ, Teusink B. Systems biology from micro-organisms to human metabolic diseases: the role of detailed kinetic models. Biochem Soc Trans. 2010;38: 1294–1301.

13. Oliveira AP, Nielsen J, Förster J. Modeling Lactococcus lactis using a genome-scale flux model. BMC Microbiol. 2005;5: 39.

14. Pastink MI, Teusink B, Hols P, Visser S, de Vos WM, Hugenholtz J. Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria. Appl Environ Microbiol. 2009;75: 3627–3633.

15. Flahaut NAL, Wiersma A, van de Bunt B, Martens DE, Schaap PJ, Sijtsma L, et al. Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. Appl Microbiol Biotechnol. 2013;97: 8729–8739.

16. Özcan E, Selvi SS, Nikerel E, Teusink B, Toksoy Öner E, Çakır T. A genome-scale metabolic network of the aroma bacterium Leuconostoc mesenteroides subsp. cremoris. Appl Microbiol Biotechnol. 2019;103: 3153–3165.

17. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. Biotechnol Bioeng. 2021;118: 223–237.

18. Branco dos Santos F, de Vos WM, Teusink B. Towards metagenome-scale models for industrial applications—the case of Lactic Acid Bacteria. Curr Opin Biotechnol. 2013;24: 200–206.

19. Jia Z, Huang L, Wei Z, Yao Y, Fang T, Li C. Dynamic kinetic analysis of growth of Listeria monocytogenes in pasteurized cow milk. J Dairy Sci. 2021;104: 2654–2667.

20. Chelladhurai K, Warakaulle S, Ali SN, Turner MS, Ayyash M, Kamal-Eldin A. Differences in the growth, acidification, and proteolytic activities of Lactobacillus helveticus, Lactobacillus delbrueckii subsp. bulgaricus, and Streptococcus thermophilus in camel and cow milk fermentation. Int Dairy J. 2025;160: 106075.

21. Yamamoto E, Watanabe R, Ichimura T, Ishida T, Kimura K. Effect of lactose hydrolysis on the milk-fermenting properties of Lactobacillus delbrueckii ssp. bulgaricus 2038 and Streptococcus thermophilus 1131. J Dairy Sci. 2021;104: 1454–1464.

22. Wu L-H, Lu Z-M, Zhang X-J, Wang Z-M, Yu Y-J, Shi J-S, et al. Metagenomics reveals flavour metabolic network of cereal vinegar microbiota. Food Microbiol. 2017;62: 23–31.

23. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102: 3–11.

24. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. 2019. doi:10.7717/peerj.7359

25. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32: 605–607.

26. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. CONCOCT: Clustering cONtigs on COverage and ComposiTion. arXiv [q-bio.GN]. 2013.

Available: http://arxiv.org/abs/1312.4038

27. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6: 1–13.

28. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055.

29. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

30. Ruscheweyh H-J, Milanese A, Paoli L, Sintsova A, Mende DR, Zeller G, et al. mOTUs: Profiling Taxonomic Composition, Transcriptional Activity and Strain Populations of Microbial Communities. Curr Protoc. 2021;1: e218.

31. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20: 257.

32. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz848

33. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119.

34. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38: 5825–5829.

35. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–62.

36. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37: D233–8.

37. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018;46: 7542–7553.

38. Rau MH, Gaspar P, Jensen ML, Geppel A, Neves AR, Zeidan AA. Genome-Scale Metabolic Modeling Combined with Transcriptome Profiling Provides Mechanistic Understanding of Streptococcus thermophilus CH8 Metabolism. Appl Environ Microbiol. 2022;88: e0078022.

39. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol. 2017;35: 81–89.

40. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem. 2006;281: 40041–40048.

41. Landi N, Ragucci S, Di Maro A. Amino Acid Composition of Milk from Cow, Sheep and Goat Raised in Ailano and Valle Agricola, Two Localities of “Alto Casertano” (Campania Region). Foods. 2021;10. doi:10.3390/foods10102431

42. Henson MA, Hanly TJ. Dynamic flux balance analysis for synthetic microbial communities. IET Syst Biol. 2014;8: 214–229.

43. Zeng H, Yang A. Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. Sci Rep. 2020;10: 4283.

44. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM, Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

45. Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat Chem Biol. 2009;5: 593–599.

46. Schumacher R. Metabolic trade-offs arising from increased free energy conservation in Saccharomyces cerevisiae. Delft University of Technology. 2018. doi:10.4233/UUID:177E9F4C-F847-436D-9FD4-9ED97BA709D9

47. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021;49: D498–D508.

48. Wittig U, Rey M, Weidemann A, Kania R, Müller W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res. 2018;46: D656–D660.

49. Vereecken KM, Van Impe JF. Analysis and practical implementation of a model for combined growth and metabolite production of lactic acid bacteria. Int J Food Microbiol. 2002;73: 239–250.

50. Aghababaie M, Khanahmadi M, Beheshti M. Developing a kinetic model for co-

culture of yogurt starter bacteria growth in pH controlled batch fermentation. J Food Eng. 2015;166: 72–79.

51. Lammerts van Bueren A, Finn R, Ausió J, Boraston AB. Alpha-glucan recognition by a new family of carbohydrate-binding modules found primarily in bacterial pathogens. Biochemistry. 2004;43: 15633–15642.

52. Sánchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol. 2017;13: 935.

53. Chervaux C, Ehrlich SD, Maguin E. Physiological study of Lactobacillus delbrueckii subsp. bulgaricus strains in a novel chemically defined medium. Appl Environ Microbiol. 2000;66: 5306–5311.

54. Ghasemi M, Najafpour G, Rahimnejad M, Beigi PA, Sedighi M, Hashemiyeh B. Effect of different media on production of lactic acid from whey by Lactobacillus bulgaricus. Afr J Biotechnol. 2009;8. doi:10.4314/ajb.v8i1.59741

55. Raveschot C, Cudennec B, Coutte F, Flahaut C, Fremont M, Drider D, et al. Production of Bioactive Peptides by Lactobacillus Species: From Gene to Application. Front Microbiol. 2018;9: 2354.

56. Machado D. A benchmark of optimization solvers for genome-scale metabolic modeling. bioRxiv. 2023. p. 2023.04.11.536343. doi:10.1101/2023.04.11.536343

57. Suzuki I, Kato S, Kitada T, Yano N, Morichi T. Growth of Lactobacillus bulgaricus in Milk. 1. Cell Elongation and the Role of Formic Acid in Boiled Milk. J Dairy Sci. 1986;69: 311–320.

58. Kaiser JC, Omer S, Sheldon JR, Welch I, Heinrichs DE. Role of BrnQ1 and BrnQ2 in branched-chain amino acid transport and virulence in Staphylococcus aureus. Infect Immun. 2015;83: 1019–1029.

59. Schmidt JDR, Beitz E. Mutational widening of constrictions in a formate–nitrite/H+ transporter enables aquaporin-like water permeability and proton conductance. J Biol Chem. 2022;298. doi:10.1016/j.jbc.2021.101513

60. Poolman B, Knol J, Lolkema JS. Kinetic Analysis of Lactose and Proton Coupling in Glu379 Mutants of the Lactose Transport Protein of Streptococcus thermophilus∗. J Biol Chem. 1995;270: 12995–13003.

61. Yun JJ, Joseph Yun J, Barbano DM, Joseph Kiely L, Kindstedt PS. Mozzarella Cheese: Impact of Rod:Coccus Ratio on Composition, Proteolysis, and Functional Properties. Journal of Dairy Science. 1995. pp. 751–760. doi:10.3168/jds.s0022-0302(95)76686-3

62. Sieuwerts S. Analysis of molecular interactions between yoghurt bacteria by an integrated genomics approach. 2009. Available: https://search.proquest.com/openview/db5cc2850745b4df8f5f1d51cc0140be/1?pq-origsite=gscholar&cbl=2026366&diss=y

63. Khandelwal RA, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ. Community flux balance analysis for microbial consortia at balanced growth. PLoS One. 2013;8: e64567.

64. Blasche S, Kim Y, Mars RAT, Machado D, Maansson M, Kafkia E, et al. Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. Nat Microbiol. 2021;6: 196–208.

65. Sieuwerts S, Molenaar D, van Hijum SAFT, Beerthuyzen M, Stevens MJA, Janssen PWM, et al. Mixed-culture transcriptome analysis reveals the molecular basis of mixed-culture growth in Streptococcus thermophilus and Lactobacillus bulgaricus. Appl Environ Microbiol. 2010;76: 7775–7784.

66. Long CP, Antoniewicz MR. Quantifying biomass composition by gas chromatography/mass spectrometry. Anal Chem. 2014;86: 9423–9427.

67. Simensen V, Schulz C, Karlsen E, Bråtelund S, Burgos I, Thorfinnsdottir LB, et al. Experimental determination of Escherichia coli biomass composition for constraint-based metabolic modeling. PLoS One. 2022;17: e0262450.

68. Liu L, Bockmayr A. Regulatory dynamic enzyme-cost flux balance analysis: A unifying framework for constraint-based modeling. J Theor Biol. 2020;501: 110317.

69. Lamoureux L, Roy D, Gauthier SF. Production of oligosaccharides in yogurt containing bifidobacteria and yogurt cultures. J Dairy Sci. 2002;85: 1058–1069.

Chapter 4 Deep learning-based prediction of temperature dependent enzyme k_{cat} values

This chapter is based on (with minor edits) the following published paper: Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. Brief Bioinform. 2023;25. doi:10.1093/bib/bbad506

Sizhe Qiu, as the first author, constructed and analyzed the deep learning model, performed case studies, and produced the first draft of the paper. Other listed authors contributed to the revision of the paper.

Summary

The enzyme k_{cat} quantifies enzyme kinetics by indicating the maximum efficiency of enzyme catalysis. Despite its importance, k_{cat} values remain scarce in databases for most organisms, primarily due to the cost of experimental measurements. To predict k_{cat} and account for its strong temperature dependence, DLTKcat was developed in this study and demonstrated superior performance (log10-scale RMSE = 0.88, R2 = 0.66) than previously published models. The prediction of k_{cat} of wild-type and mutated *Pyrococcus furiosus* Ornithine Carbamoyltransferases showed that DLTKcat could predict the effects of amino acid substitutions. Subsequently, this study built temperature sensitive proteome constrained flux balance analysis models with predicted k_{cat} at varying temperatures for two widely used lactic acid bacteria, *Lactococcus lactis* MG1363 and *Streptococcus thermophilus* LMG18311, and captured the metabolic response to temperature changes. Although its accuracy is not high enough yet to quantitatively model temperature dependent cellular metabolism, DLTKcat has

the potential to eventually become a computational tool to describe the temperature dependence of biological systems.

Reporting the second research project in this PhD study, this chapter presents a deep learning driven modeling approach for the growth-coupled primary metabolism of lactic acid bacteria under temperature stress, and DLTKcat was used in Chapter 6 to estimate enzyme k_{cat} for parameter gap-filling.

4.1 Introduction

In the age of synthetic biology, more and more chemical processes are being catalyzed by enzymes [1,2], and therefore, the quantitative study of enzyme kinetics becomes an important topic. The enzyme turnover number, k_{cat}, is one of the most important parameters in describing enzyme kinetics, which quantifies the maximum efficiency of an enzyme in catalyzing a specific reaction [3]. In spite of its importance, there currently exists a huge gap of measured k_{cat} for most organisms in commonly used enzyme databases [4], i.e., BRENDA [5] and SABIO-RK [6]. Also, measuring k_{cat} values via enzyme assays is expensive and labor intensive [4], which means that it's hard to obtain k_{cat} values in a high-throughput manner. The limited availability of k_{cat} in databases and the indispensable requirement for k_{cat} in the study of enzyme kinetics and other fields, such as metabolic modeling [7], fuels the impetus behind the development of computational methods to predict k_{cat} values.

There are two main methods to predict k_{cat} values: 1. estimating k_{cat} based on apparent catalytic rate (k_{app}) with proteomic and fluxomic profiling; 2. predicting k_{cat} using the compound protein interaction (CPI) deep learning model. The first method obtains the k_{cat} value by dividing the measured reaction flux by the quantified protein abundance [8,9]. Although this method has been proved successful in resource allocation models of various microorganisms

[10–13], fluxomics and proteomics are costly to measure, making this method difficult to implement.

CPI deep learning models have already been developed to predict biological parameters such as binding affinities (K_d) [14], michaelis-menten constants (K_m) [15] and enzyme turnover numbers (k_{cat}) [16]. The inputs are usually SMILES (simplified molecular-input line-entry system) strings of compounds and subsequences of proteins. Compound and protein features are extracted by graph neural network (GNN), recurrent neural network (RNN) or convolutional neural network (CNN), and then concatenated for the regression of the target value, such as k_{cat} or K_m [17]. For better performance, attention layers are added to capture the interaction between compound and protein features [18,19]. DLKcat [16], the first CPI deep learning model for k_{cat} prediction, can predict $log_{10}(k_{cat})$ with the RMSE (root mean squared error) score below 1 and Pearson's r = 0.71 for the test dataset. However, one limitation of DLKcat and most other CPI models is that they do not account for experimental conditions like temperature, pH or ionic strength. As k_{cat} has a strong dependence on temperature [20] and temperature is widely available in databases, developing a deep learning model that takes compound, protein and temperature features together as inputs is both necessary and approachable.

TurNuP [21], a CPI model for k_{cat} with enhanced performance than DLKcat, included temperature as a feature in a case study to predict k_{cat} for *E. coli*, but it was not a general predictive model for temperature dependent k_{cat}. EF-PreKcat and Revised PreKcat [22] were developed to predict temperature dependent k_{cat} values. They considered k_{cat} values at different temperatures and include the temperature value as a feature. However, the feature importance of temperature in those two models was not assessed, and no case studies were conducted to show the model's ability to predict the effect of temperature on k_{cat} values. Also, the R-squared (R2) scores of predictions by those two models were reported to be below 0.5.

With the aim to construct a deep learning model on k_{cat} prediction that is more accurate than previously published models, this study developed DLTKcat. DLTKcat is a bi-directional attention CPI model with molecular graphs converted from SMILES strings, 3-mer subsequences of proteins and temperature features as inputs. It showed superior performance (log10-scale RMSE = 0.88, R2 = 0.66) than previously published models (e.g., EF-PreKcat), and demonstrated the feature importance of temperature. Then, DLTKcat exhibited its potential application in enzyme sequence design by predicting the effect of amino acid substitutions on k_{cat} at different temperatures. Finally, we incorporated temperature dependent proteome constraints in bacterial metabolic modeling with predicted k_{cat} at different temperatures, to explore the possibility of using DLTKcat to make metabolic modeling sensitive to temperature changes.

4.2 Methods

4.2.1 Dataset preparation

The dataset used to construct the deep learning model was extracted from the BRENDA and SABIO-RK databases. EC number, substrate name, organism name, protein identifier (UniProt ID), enzyme type, temperature and k_{cat} values were queried from SABIO-RK via application programming interface (API). The data in BRENDA was fetched using BRENDApyrser [23]. The canonical SMILES string [24] of the substrate, that describes the molecular structure of chemical species, was obtained by querying the PubChem compound database [25] via API. The amino acid sequence of each enzyme protein was queried from the UniProt database [26] based on the UniProt ID also via API. The sequences of wild type (WT) enzymes were mapped directly. For mutants caused by amino acid substitutions, amino acids at mutated locations were changed based on mutation information from BRENDA and SABIO-RK.

Entries with other types of mutations were removed. All API codes can be found at https://github.com/SizheQiu/DLTKcat.

After SMILE strings and amino acid sequences were obtained, the dataset filtered out all redundant entries with the same SMILE string, amino acid sequence, temperature and k_{cat} value. For entries with the same SMILE string, amino acid sequence, temperature but different k_{cat} values, only the entry with the largest k_{cat} value was kept, as done in Li et al., 2022 [16]. Finally, 4383 entries from SABIO-RK and 11866 entries from BRENDA remained. 10556 entries' enzymes were WTs and 5693 entries' enzymes were mutants (**see Appendix B, Figure B.1**). k_{cat} values of 87 enzyme classes (EC numbers) were found to have significant correlations with temperature, which covered 2430 entries (**see Appendix B, Figure B.2**). Considering the uneven distribution of temperature values in the dataset, oversampling was performed to append two times of entries at low (T<20℃) and high (T>40℃) temperature ranges by randomly duplicating existing entries at those temperature ranges (**see Appendix B, Figure B.3**). Because previously published CPI deep learning models have shown that additional features, such as enzyme molar mass or the octanol–water partition coefficient of substrate, could not improve model performance [15,21], the finalized dataset of this study only contained SMILES strings of substrates, amino acid sequences of enzyme proteins and temperature values.

4.2.2 Construction of the deep learning model

Similar to other CPI deep learning models, DLTKcat uses Graph Attention Network (GAT) and Convolutional Neural Network (CNN) to extract features from the substrate molecular graph and enzyme protein sequence, respectively (**Figure 4.1**). The use of bi-directional attention, adopted from BACPI by Li et al., 2022 [27], and integration of temperature and inverse temperature values capture the temperature dependent interactions between atoms of the

compound and residues of the protein. Finally, the concatenated features of compound, protein and temperature are fed into several dense layers (fully connected layers) to predict the $log_{10}(k_{cat})$ value.

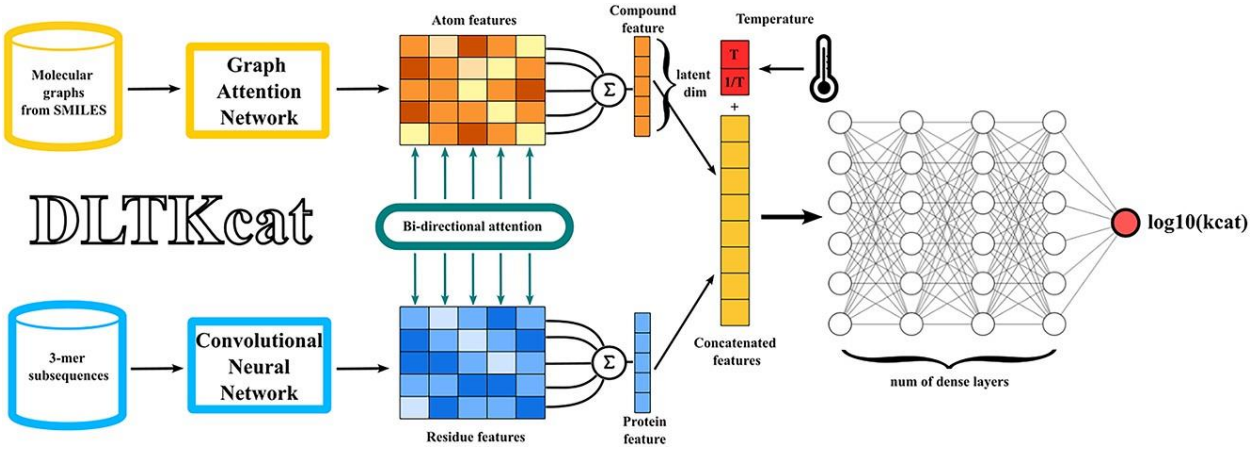

Figure 4.1. The overview of DLTKcat. With a pair of substrate and enzyme as the input, a GAT and a CNN learn the representations of the atom and residue from the compound molecular graph and protein sequence. Next, atom and residue representations are fed into the bi-directional attention neural network to integrate the representations and capture the important regions of compounds and proteins. Then, temperature (T) and inverse temperature ($\frac{1}{T}$) are integrated into the concatenated features. Finally, the concatenated features are used to predict the $log_{10}(k_{cat})$ value.

4.2.2.1 Compound representation

RDKit [28] converts the SMILES string into the molecular graph of the substrate with atoms as vertices, and chemical bonds as edges. The graph, along with the initial embeddings of its vertices, is fed into the graph attentional layer of GAT. A linear learnable transformation converts the embeddings ($v_i^{init} \in R^{H_c}$, H_c=80) into higher-level features of the compound ($v_i^{`} \in R^{H_c^{`}}$, $H_c^{`}$=50). The multi-head attention mechanism in GAT concatenates output features from 3 independent graph attentional layers to increase the stability of the self-attention learning

process. Finally, a single-layer neural network transforms concatenated features into the compound space. The final output features are atom features ($v_i \in R^{H_c}$) (**Figure 4.1**). Extended Connectivity Fingerprints (ECFPs) [29] of length 1024, computed by RDKit, are also used to represent the compound. A multi-layer neural network transforms ECFPs into the compound space ($f \in R^{H_c}$).

4.2.2.2 Protein representation

To capture diverse protein residue patterns, the protein sequence is split into overlapping 3-mer subsequences. 3-mer subsequences are then translated to randomly initialized embeddings ($r_i^{init} \in R^{H_p}$, H_p=80). Through 4 convolutional layers with leaky ReLU [30] as the activation function, embeddings are transformed to higher-level features of the protein sequence that can capture the complex relationships of residues. The final output features are residue features ($r_i \in R^{H_p}$) (**Figure 4.1**).

4.2.2.3 Bi-directional attention and integration of temperature

The bi-directional attention mechanism is used to represent the interactions between atoms of the compound and residues of the protein. Residue, atom features and fingerprints are transformed into vectors ($c_i \in R^d$, $p_i \in R^d$, $hf \in R^d$), and a soft alignment matrix ($A \in R^{N_v \times N_r}$) indicates the interaction strengths. d is the unified latent dimension (d=40, 64). The weighted information is extracted from the soft alignment matrix, and attention weights are computed in both atom-to-residue ($\alpha_{a2r} \in R^{N_v}$) and residue-to-atom ($\alpha_{r2a} \in R^{N_r}$) directions. The outputs are compound ($h_c \in R^d$) and protein ($h_p \in R^d$) features (**Figure 4.1**). To improve learning stability and representation capacity, a multi-head attention model (number of heads=3) is used to capture diverse aspects of compound-protein interactions ($h_c^{final} \in R^d$, $h_p^{final} \in R^d$).

Inspired by the Arrhenius equation ($k_{cat} = Ae^{-\frac{E_a}{RT}}$) [20], temperature ($T$) and inverse temperature ($\frac{1}{T}$) are first normalized ($\frac{x-x_{min}}{x_{max}-x_{min}}$), and then concatenated with compound and protein features output by the bi-directional attention process. The inverse of temperature best represents the linear relationship between $\frac{1}{T}$ and $log_{10}(k_{cat})$. The concatenated features ($h_c^{final}||h_f||h_p^{final}||[T, \frac{1}{T}]$, || is the concatenation operation) are then fed into several dense layers (layer number = 3, 4, 5, 6), with leaky ReLU as the activation function, for the regression of the $log_{10}(k_{cat})$ value.

4.2.2.4 Model training

Due to the large size of the dataset, batch training was used with a batch size of 32. Adam optimization algorithm [31] was used to update neural network weights iteratively. The loss function was mean squared error (MSE). The initial learning rate was 0.001, and the learning rate decayed by 50% for every 10 epochs to prevent overfitting. For details of software and hardware, please see the section S1.1 of the supplementary information.

4.2.3 Interpretation of attention weights on protein residues

The bi-direction attention mechanism in section 4.2.2.3 assigns attention weights to protein subsequences and atoms of the substrate. A higher attention weight of one residue means that residue is more important for the enzyme kinetics towards a certain substrate. The residue attention weights (α_{r2a}) can be computed based on the intermediate output in the deep learning model.

$$c_i = LeakyReLU(W_v v_i)\ (Eq. 4.1)$$

$$p_i = LeakyReLU(W_r r_i)\ (Eq. 4.2)$$

$$I_p = A^T tanh(CW_{a2r})\ (Eq. 4.3)$$

$$\alpha_{r2a} = softmax([PW_p \,||\, I_p]a_{r2a})\ (Eq. 4.4)$$

v_i and r_i are atom and residue feature vectors (**sections 4.2.2.1 and 4.2.2.2**), $W_v \in R^{d \times H_c}$ and $W_r \in R^{d \times H_p}$ transform v_i and r_i to c_i and p_i, respectively (Eqs. 4.1 & 4.2). d is the latent dimension in the bi-directional attention mechanism. $C = [c_1, c_2, \ldots, c_{N_v}]$, $P = [p_1, p_2, \ldots, p_{N_r}]$, $W_{a2r} \in R^{d \times d}$, $A = tanh(CUP^T) \in R^{N_v \times N_r}$ is a pairwise interaction matrix for atoms and residues, and $I_p \in R^{N_v \times d}$ represents information from atoms to residues (Eq. 4.3). $W_p \in R^{d \times d}$, || is the concatenation operation, $a_{r2a} \in R^{2d}$, and the vector of residue attention weights (α_{r2a}) is protein attention weights normalized by the softmax function (Eq. 4.4).

4.2.4 Proteome constrained flux balance analysis with predicted k_{cat}

Flux balance analysis (FBA) has been used to estimate metabolic fluxes and cellular growth rates for decades [32]. The basic required inputs are the stoichiometric matrix (S) from the genome-scale metabolic model (GSMM) [32] and growth medium parameters that set upper bounds for nutrient uptake rates. FBA computes metabolic fluxes (v_i) by maximizing an objective function (Eq. 4.5), which is usually the growth function (v_{growth}, biomass formation rate normalized to 1 gram dry weight (gDW) of biomass), via linear optimization in a constrained solution space of mass conservation (Eq. 4.6) and lower/upper bounds (v_{lb}, v_{ub}) of reaction fluxes (Eq. 4.7). FBA was conducted using COBRApy [33] in this study.

$$Max\ v_{growth}\ (Eq. 4.5)$$

$$S * v = 0\ (Eq. 4.6)$$

$$v_{lb} \leq v_i \leq v_{ub}\ (Eq. 4.7)$$

Proteome constrained FBA tightens the solution space by integrating proteome constraints of reactions into conventional FBA [34]. The reaction flux (v_i, $\frac{mmol}{hr * gDW}$) is constrained by the enzyme capacity ($k_i[E_i]$ or $a_i(MW_i * [E_i])$) (Eq. 4.8). k_i is the k_{cat} of reaction i and $[E_i]$ is the enzyme molar concentration ($\frac{mmol}{gDW}$). a_i ($\frac{\mu mol}{min * mg\ E}$) is the enzyme specific activity, defined as

the micro moles of products formed by an enzyme in a given amount of time per milligram of the enzyme protein. MW_i is enzyme molar mass ($\frac{g}{mol}$). Proteome was divided into sectors of inflexible housekeeping (Q), anabolism (A), transportation (T) and catabolism (C). The upper bound of all flexible sectors (i.e., C, A, T) combined was assumed to be 50% of the total proteome (Eq. 4.9) [35–37].

$$v_i \leq k_i[E_i] \;\; or \; v_i \leq a_i(MW_i * [E_i]) \; (Eq. 4.8)$$

$$\phi_Q(50\%) + \phi_C + \phi_A + \phi_T \leq 100\% \; (Eq. 4.9)$$

$$\phi_A * P_{TOT} = MW_{ribosome} * [E_{ribosome}] = \frac{v_{growth}}{a_{ribosome}} \; (Eq. 4.10)$$

$$\phi_C * P_{TOT} = \sum_i MW_i * [E_i] = \sum_i \frac{v_i * MW_i}{k_i} \; (Eq. 4.11)$$

ϕ_x is the mass fraction of sector x for $x = \; A, C, T$. P_{TOT} is the total mass of the proteome normalized to 1 gDW of biomass ($\frac{g}{gDW}$). The enzyme activity of the ribosome for the anabolism sector ($a_{ribosome}$) was set as $107.4 \; \frac{mmol}{hr * g\,E}$ (Eq. 4.10) [36,38]. k_{cat} values were predicted for the catabolic sector (sector C), by DTLKcat (Eq. 4.11).

In this study, proteome constrained FBA was performed for *Lactococcus lactis MG1363* (**LL**) and *Streptococcus thermophilus LMG18311* (**ST**). The GSMMs used were obtained from the work of Flahaut et al., 2013 [39] and Pastink et al., 2009 [40]. Experimental data of LL and ST's growth rates at different temperatures were obtained from Chen *et al.*, 2015 [41] and Vaningelgem *et al.*, 2004 [42]. The carbon sources of LL and ST, in experiments, were glucose and lactose, respectively. Therefore, the enzyme activities (a_{CT}, CT stands for carbon source transportation) of glucose transport via phosphotransferase system and lactose: galactose antiporter were set as $361.14 \; \frac{mmol}{hr * gE}$ [43], and $540 \; \frac{mmol}{hr * gE}$ [44] (Eq. 4.11). Because both lactose and glucose were sufficient in the growth medium [41,42], no michaelis-menten kinetics was needed for transporter proteins. Lactic and acetic acids were two major products of the central

carbon metabolism of lactic acid bacteria, and the enzyme activity of acid exportation (a_{AT}) was set as $6360\ \frac{mmol}{hr * gE}$ (Eq. 4.12) [36,38].

$$\phi_T * P_{TOT} = MW_{AT} * [E_{AT}] + MW_{CT} * [E_{CT}] = \frac{v_{AT}}{a_{AT}} + \frac{v_{CT}}{a_{CT}}\ (Eq. 4.12)$$

Temperature dependent k_{cat} values were predicted for enzymes in two bacteria's central carbon metabolism (**see Appendix B, Tables B.1 & B.2**). The SMILES strings of substrates were queried from PubChem with metabolite names in GSMMs, and protein sequences were queried from UniProt with gene locus tags in genome assemblies of LL, GCF_000009425.1 [45], and ST, GCF_000011825.1 [46]. The predicted k_{cat} for the primary substrate of each reaction was selected as the k_{cat} of the reaction. For isozymes that catalyze the same metabolic reaction, the largest k_{cat} was selected. Both ST and LL are important and widely used lactic acid bacteria, but their enzyme k_{cat} values are quite limited in databases. For example, there are only 11 entries for ST in SABIO-RK, most were contributed by Simon and Hofer, 1981 [47]. Therefore, this study used DLTKcat to fill the gap and examined DLTKcat's performance in predicting metabolic responses to temperature changes.

4.3 Results

4.3.1 DLTKcat has good performance on temperature dependent k_{cat} prediction

With optimal hyperparameters (**see Appendix B, section B.1.3 and Figure B.4**), the model training process reduced RMSE (**see Appendix B, Eq. B.2**) scores of predicted $log_{10}(k_{cat})$ of the test dataset from 1.33 to 0.88, and enhanced R2 (**see Appendix B, Eq. B.1**) scores from 0.25 to 0.66 after 20 epochs (**Figure 4.2A**). The R2 scores of previously published deep learning models on temperature dependent k_{cat} were all reported to be below 0.5 [22], and DLTKcat has outperformed them by reaching a R2 score of 0.66 on the randomly selected

test dataset (**Figure 4.2B**). In addition, DLTKcat showed good prediction accuracy with low RMSE and mean absolute error (MAE, **see Appendix B, Eq. B.3**) scores for sub-datasets with experimental $log_{10}(k_{cat})$ values at the lower 25%, middle 50% and upper 25% ranges (**Figure 4.2C**). In a nutshell, DLTKcat demonstrated superior performance in comparison to previously published deep learning models for temperature dependent k_{cat}, and a robust accuracy for target values (experimental $log_{10}(k_{cat})$ values) at different ranges.

To explore the predictive power of DLTKcat across different metabolic contexts, the prediction accuracy of $log_{10}(k_{cat})$ values for enzymes in four different pathways, categorized based on enzyme modules in KEGG database [48], were assessed, and R2, RMSE and MAE scores were all around 0.8, 0.6 and 0.35 (**Figure 4.2D**). After the validation of DLTKcat's good accuracy across different metabolic contexts, the model showed its ability to discriminate enzymes in primary metabolism - catabolism/energy (primary-CE) and other pathways, with higher predicted $log_{10}(k_{cat})$ values in primary-CE (p-value<0.001) (**Figure 4.2E**). In short, DLTKcat could well characterize enzymes from different metabolic contexts.

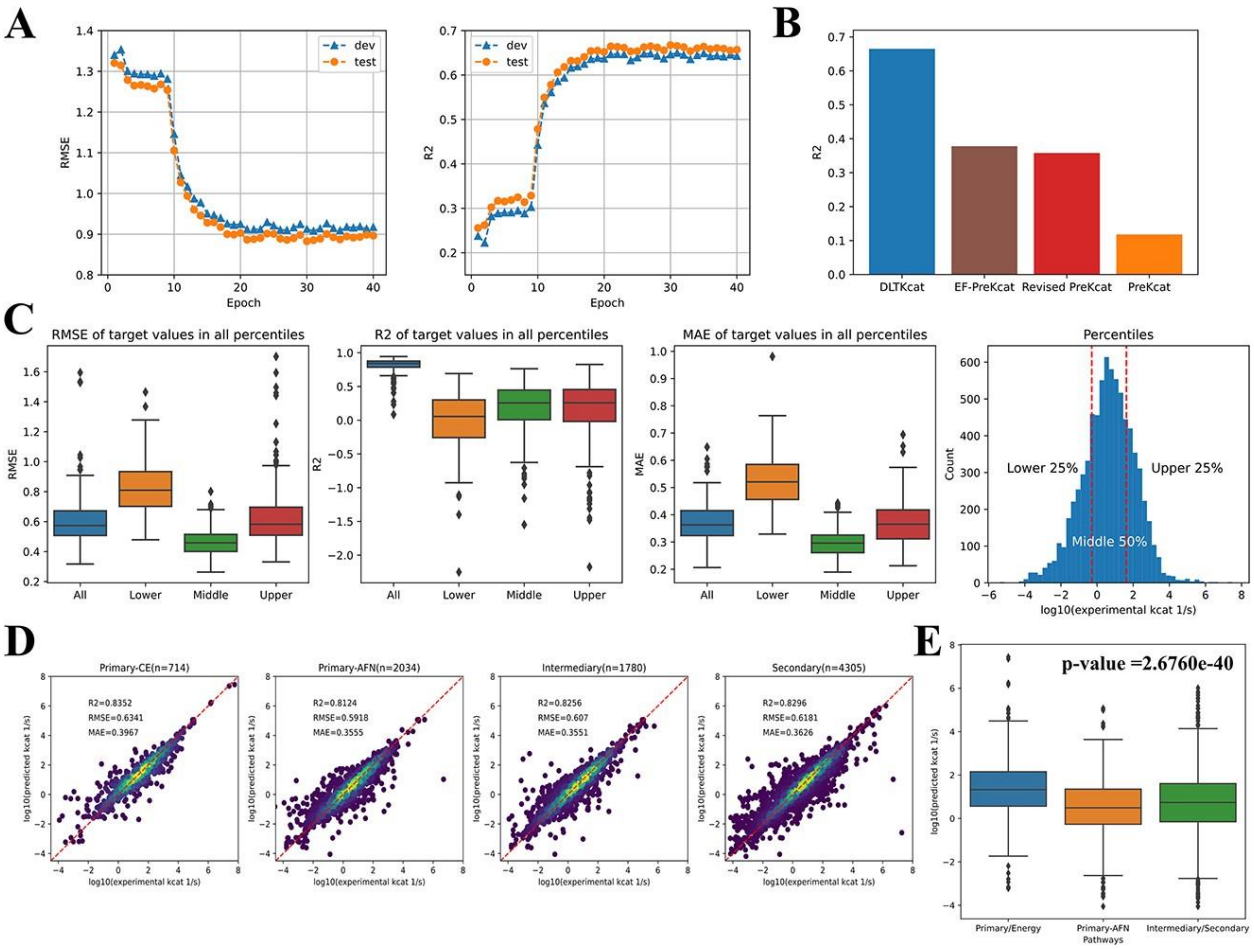

Figure 4.2. Assessment of the model performance. (A) The RMSE and R2 scores of $log_{10}(k_{cat})$ prediction during the training process. test: the test set; dev: the validation set. The RMSE and R2 of the test set at the end of training are 0.88 and 0.66. (B) Comparison of reported R2 scores of DLTKcat, EF-PreKcat, Revised PreKcat and PreKcat on $log_{10}(k_{cat})$ prediction with temperature values. (C) The distributions of RMSE, R2 and MAE scores of $log_{10}(k_{cat})$ prediction for target values at lower 25%, middle 50% and upper 25% percentiles. (D) R2, RMSE and MAE scores of $log_{10}(k_{cat})$ prediction for enzymes in primary-CE, primary-AFN, intermediary and secondary metabolism. (E) The comparison of distributions of predicted $log_{10}(k_{cat})$ values in primary-CE and other metabolic pathways (p-value < 0.001). Primary-CE: primary metabolism - catabolism/energy; Primary-AFN: primary metabolism - amino acid/fatty acid/nucleotide.

4.3.2 Interpretation of k_{cat} prediction of mutated enzymes

First, the accuracy of DLTKcat for both WT and mutated enzymes was examined, and R2, RMSE and MAE scores were around 0.8, 0.6 and 0.4, respectively (**see Appendix B, Figure B.5**). After the prediction accuracy was ensured, this study selected 3 enzyme-substrate pairs with more than 20 mutations in the dataset to investigate how DLTKcat captures amino acid substitutions. The 3 enzyme-substrate pairs were glucose-6-phosphate isomerase and D-glucose 6-phosphate (G6PI+g6p), benzoylformate decarboxylase and benzoylformate (BFDC+bzfor), and ADP-ribose diphosphatase and ADP-ribose (ADPRDP+adprib). The uniprot IDs of 3 enzyme proteins wereP06744, P20906, and Q5SKW5. Amino acid substitutions on protein sequences of 3 enzymes all resulted in the decrease of k_{cat} (**Figure 4.3A**). The prediction accuracy of the selected 3 enzyme-substrate pairs was slightly lower than that of all mutated enzymes, but the prediction error was still around one order of magnitude (**Figure 4.3B, Figure B.5**). Next, the mapping of mutation sites to residue attention weights (**section 4.2.3**) shows that most mutation sites (< 0.1-fold WT k_{cat}) distribute closely to peaks of attention weights (**Figure 4.3C-E**). The overlapping between mutation sites (< 0.1-fold WT k_{cat}) and residues with high attention weights was most noticeable for residue 70, 460 and 464 on BFDC (**Figure 4.3D**). Generally speaking, DLTKcat is a good predictor for mutated enzymes, and residue attention weights can reflect the impact of amino acid substitutions on enzyme kinetics.

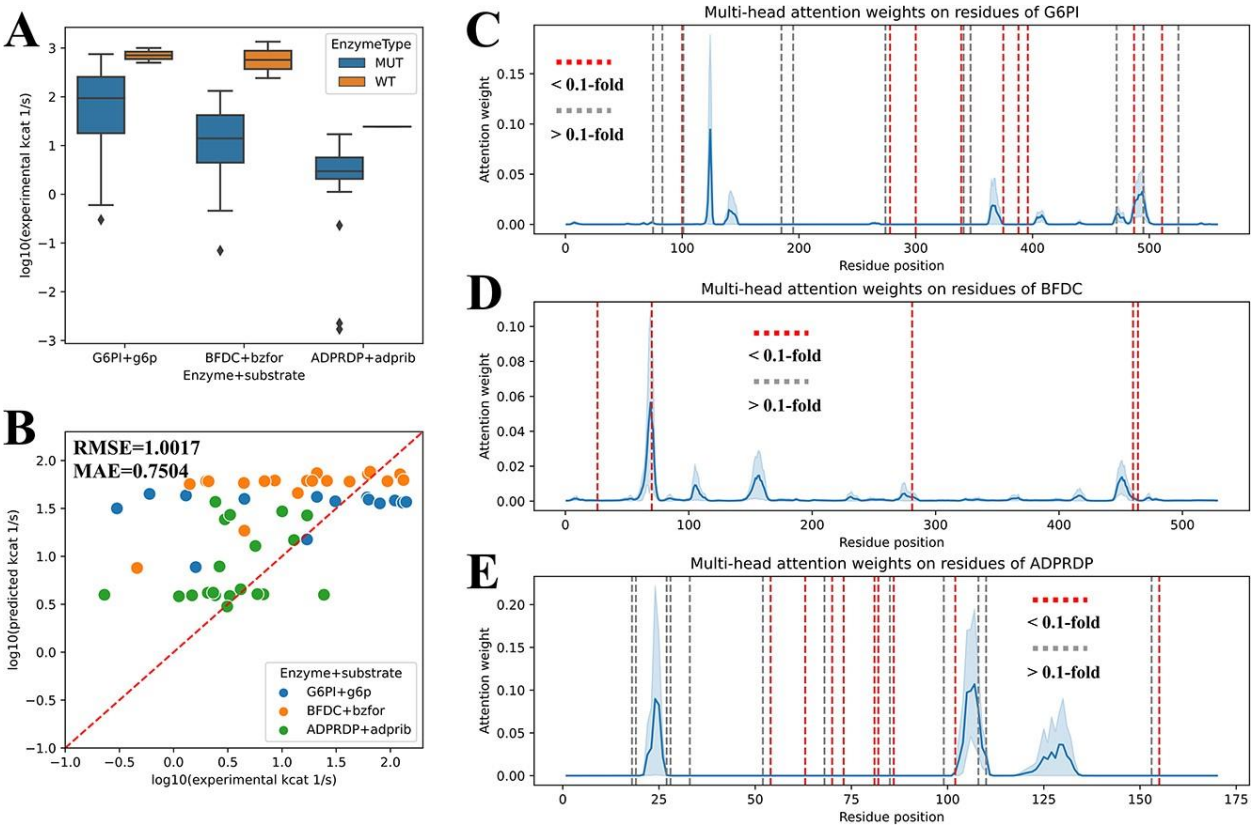

Figure 4.3. DLTKcat for the prediction and interpretation of k_{cat} of mutated enzymes. (A) The comparison between experimental $log_{10}(k_{cat})$ of WT and mutated enzymes for G6PI+g6p, BFDC+bzfor, and ADPRDP+adprib. WT: wild-type; MUT: mutant. (B) RMSE and MAE scores of predicted $log_{10}(k_{cat})$ values for G6PI+g6p, BFDC+bzfor, and ADPRDP+adprib. RMSE=1.0017, MAE=0.7504. (C) Multi-head attention weights on residues of the WT G6PI and mutation sites. (D) Multi-head attention weights on residues of the WT BFDC and mutation sites. (E) Multi-head attention weights on residues of the WT ADPRDP and mutation sites. Red dash line: mutation site (< 0.1-fold WT k_{cat}); Grey dash line: mutation site (> 0.1-fold WT k_{cat}); Blue curve: attention weight. G6PI: glucose-6-phosphate isomerase; g6p: D-glucose 6-phosphate; BFDC: benzoylformate decarboxylase; bzfor: benzoylformate; ADPRDP: ADP-ribose diphosphatase; adprib: ADP-ribose.

4.3.3 The contribution of temperature related features to k_{cat} prediction

Before feature importance analysis, the prediction accuracy was examined for different temperature ranges (below 20℃, above 40℃ and between 20℃ and 40℃). High R2 and low RMSE scores reflected that DLTKcat could accurately predict k_{cat} for low, middle and high temperatures, with an error far below one order of magnitude (**see Appendix B, Figure B.6**). Then, feature shuffling, also known as feature permutation, was performed to show the importance of temperature and inverse temperature values (**see Appendix B, section B.1.4**). The shuffling of temperature features resulted in significantly higher distributions of the prediction error (RMSE and MAE), and lower distributions of R2 than those of predictions with unshuffled temperature features (**Figure 4.4A**). The comparison between predicted and experimental values showed that the RMSE and MAE scores increased by around 0.1 and R2 decreased by around 0.1 when temperature related features were shuffled (**Figure 4.4B**). For high (T>40℃) and low (T<20℃) temperature ranges, the increase of RMSE, MAE and decrease of R2, caused by feature shuffling, became larger (**Figure 4.4CD**). In short, the decrease in prediction accuracy with shuffled temperature related features demonstrated the importance of temperature related features in DLTKcat.

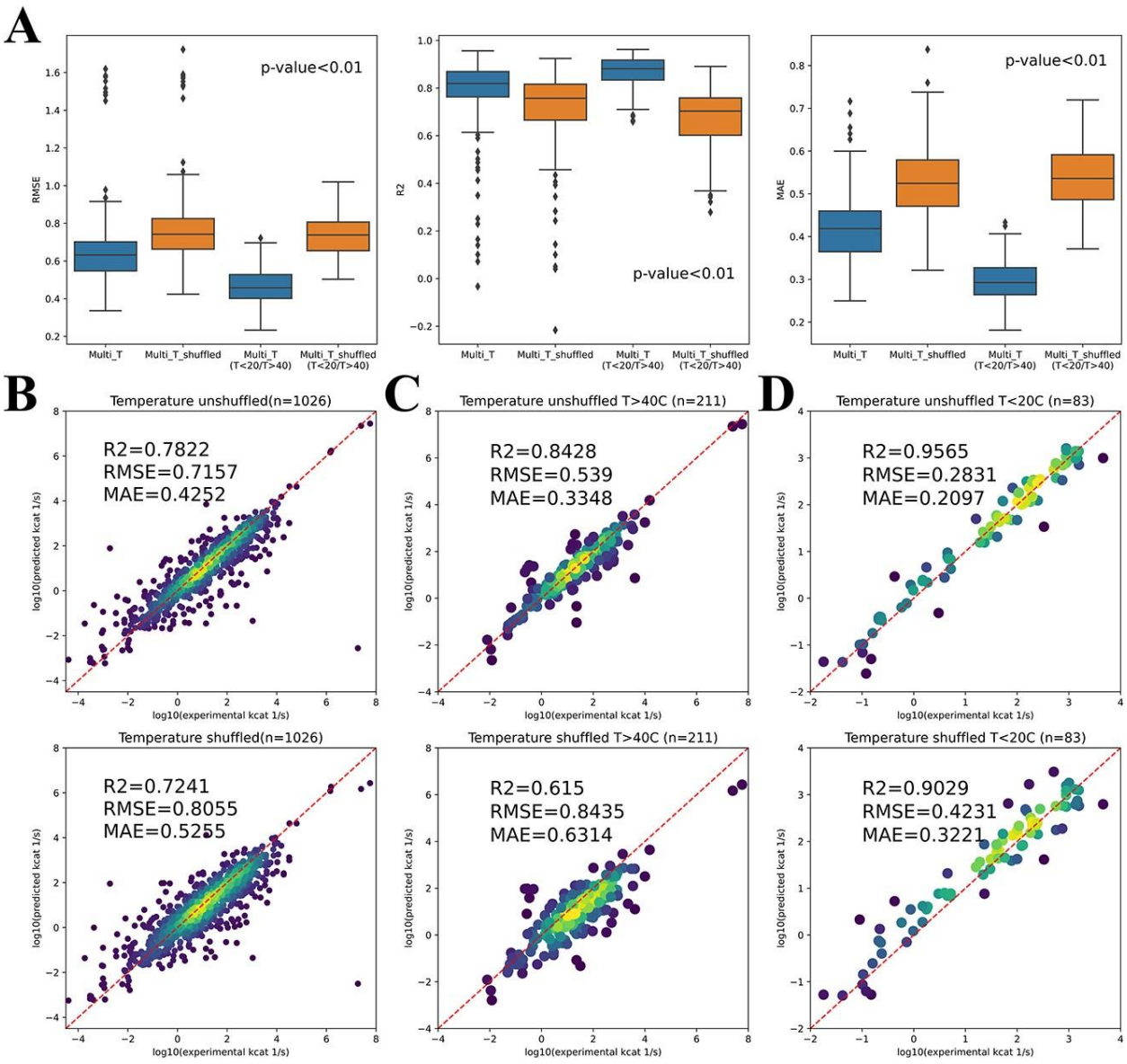

Figure 4.4. The importance of temperature related features in DLTKcat. (A)The distributions of RMSE, R2, MAE scores of predicted $log_{10}(k_{cat})$ values with unshuffled and shuffled temperature related features for the selected dataset with 1026 entries (Multi_T) and for entries of low (T<20℃) and high (T>40℃) temperature. (B) R2, RMSE and MAE scores of predicted $log_{10}(k_{cat})$ values with unshuffled and shuffled temperature related features for the selected dataset with 1026 entries (C) R2, RMSE and MAE scores of predicted $log_{10}(k_{cat})$ values with unshuffled and shuffled temperature related features for entries of high temperature. (D) R2,

RMSE and MAE scores of predicted $log_{10}(k_{cat})$ values with unshuffled and shuffled temperature related features for entries of low temperature.

4.3.4 Use DLTKcat to predict k_{cat} of wild-type and mutated *Pyrococcus furiosus* Ornithine Carbamoyltransferases

The k_{cat} values of WT and mutated *Pyrococcus furiosus* Ornithine Carbamoyltransferases at 30℃ and 55℃ were obtained from Roovers et al., 2001 [49]. Roovers et al., 2001 provided enzyme k_{cat} measured for both WT and mutants, and under different temperatures, allowing the examination of DLTKcat's sensitivity on both point mutations and temperature changes. Also, the data of Roovers et al., 2001 was not included in the dataset used to develop DLTKcat. The protein sequence of *Pyrococcus furiosus* Ornithine Carbamoyltransferase was obtained from Uniprot with the Uniprot ID of Q51742. The prediction achieved high accuracy (RMSE = 0.5, MAE=0.4338) (**Figure 4.5A**). Predicted k_{cat} values at 55 ℃ were higher than those at 30℃ (**Figure 4.5A-C**), which was both consistent with the experimental data and the nature of *Pyrococcus furiosus* being a hyperthermophile favoring high temperature [50].

With respect to the effect of mutations, DLTKcat suggested that amino acid substitutions at 227th, 240th and 277th amino acids could increase the k_{cat} value, consistent with the experimental data, despite that the numerical difference between predicted k_{cat} values of mutants and WT was small (**Figure 4.5BC**; note the difference in scale between upper and lower y-axes). Furthermore, DLTKcat also captured that the combination of two amino acid substitutions, Y227C/E277G and A240D/E277G, could result in greater improvement on the k_{cat} value than the substitution at each single site, though it failed to predict that the k_{cat} of A240D/E277G was higher than that of Y227C/E277G (**Figure 4.5BC**). The mapping of mutation sites to residue attention weights showed that E277G, as the mutation with a higher enhancement of k_{cat} than other two mutations, was also closer to the high peak of attention

weights (**Figure 4.5D**). In addition, residue attention weights indicated other potential mutation sites on *Pyrococcus furiosus* Ornithine Carbamoyltransferase that might have substantial effects on k_{cat} (**Figure 4.5D**).

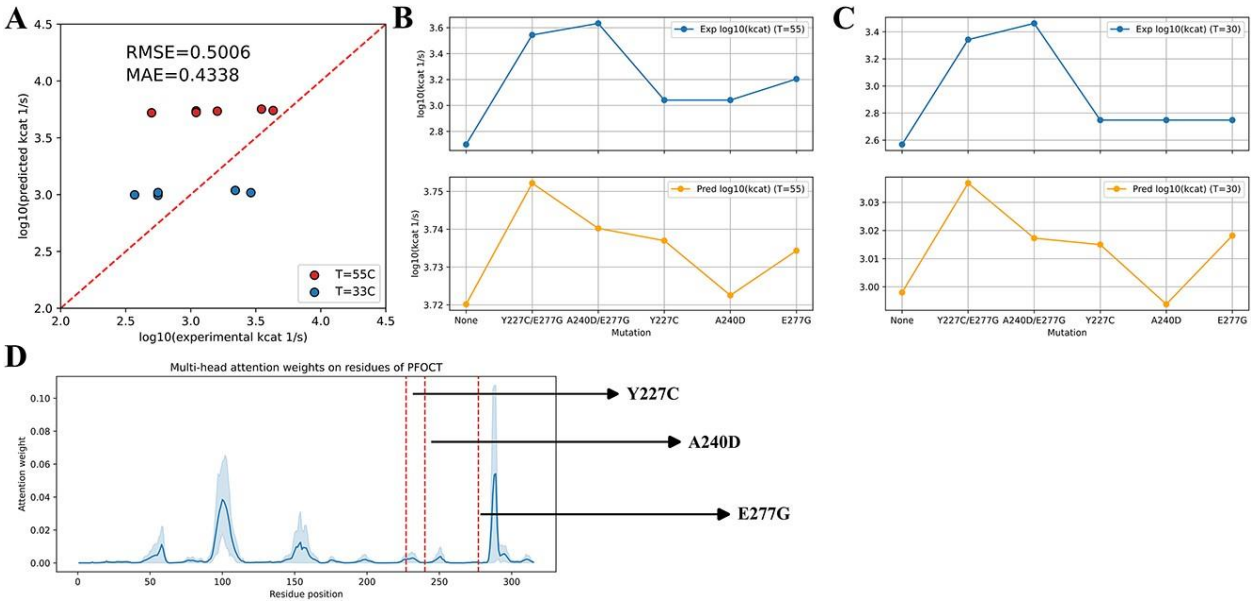

Figure 4.5. Prediction of the effect of amino acid substitutions on k_{cat} values. (A) Comparison between experimental and predicted $log_{10}(k_{cat})$ of *Pyrococcus furiosus* Ornithine Carbamoyltransferase, RMSE = 0.5006, MAE = 0.4338. (B) Experimental (blue line) and predicted (orange line) $log_{10}(k_{cat})$ values of WT and mutants at 55℃. (C) Experimental (blue line) and predicted (orange line) $log_{10}(k_{cat})$ values of WT and mutants at 30℃. Exp: experimental value; Pred: predicted value. (D) Multi-head attention weights on residues of the WT *Pyrococcus furiosus* Ornithine Carbamoyltransferase protein sequence. Red dash-line: mutation site.

4.3.5 Temperature sensitive metabolic modeling with predicted k_{cat}

DLTKcat predicted k_{cat} values for enzymes of *Lactococcus lactis MG1363* (LL) at 30, 32, 34, 36 and 38 ℃, and of *Streptococcus thermophilus LMG18311* (ST) at 25, 32, 37, 42, 46 and 49 ℃, which were temperatures where LL and ST's growth rates were measured in

experimental data [41,42]. DLTKcat predicted that k_{cat} of most catabolic enzymes in LL would decrease when temperature increased from 30 ℃ to 38 ℃, especially for glucose-6-phosphate isomerase (PGI), phosphofructokinase (PFK), phosphoglycerate kinase (PGK), pyruvate kinase (PYK), pyruvate formate lyase (PFL) and phosphotransacetylase (PTAr) (**Figure 4.6A**). The predicted decrease of the activity of catabolism in LL in response to temperature increase is consistent with the experimental observation that LL stopped growing after temperature became larger than 38 ℃ [41]. For catabolic enzymes in ST, DLTKcat predicted that most enzymes' k_{cat} would increase when temperature increased from 25 ℃ to 42 ℃, especially for fructose-bisphosphate aldolase (FBA), Glyceraldehyde-3-phosphate dehydrogenase (GAPD), phosphoglycerate mutase (PGM), enolase (ENO) and pyruvate kinase (PYK) (**Figure 4.6B**). The predicted increase of catabolic activity in ST when temperature increases to 42 ℃ is consistent with both the experimental data [42] and the nature of ST being a thermophile [51]. These results showed that, in general, DLTKcat could qualitatively predict metabolic responses of bacteria to certain temperature changes.

However, the quantitative accuracy of growth rates computed by proteome constrained FBA was low. In proteome constrained FBA for LL, the k_{cat} of fructose-bisphosphate aldolase (FBA) in LL was fixed at $13.9\ \frac{1}{s}$ [52] in sacrifice of temperature sensitivity, because predicted k_{cat} values at different temperatures were unrealistically low ($0.04 \sim 0.065\ \frac{1}{s}$), compared with experiment k_{cat} values in other bacteria [52,53]. The predicted growth rates of LL by proteome constrained FBA captured the decreasing trend in response to the increase of temperature, but the predicted values were deviant from experimental values (**Figure 4.6C**). The proteome constrained FBA predicted the increase of ST's growth rate from 25 ℃ to 42 ℃, but it failed to predict the drop of growth rate from 42 ℃ to 49 ℃ (**Figure 4.6D**). Also, the predicted increase of k_{cat} values from 42 ℃ to 49 ℃ by DLTKcat (**see Appendix B, Figure B.7**) contradicted the experimental finding that 49 ℃ is close to the theoretical maximum temperature for ST to

survive, 47~50 ℃ [51]. To conclude, the log10-scale RMSE score within 1 of DLTKcat is not low enough to enable temperature sensitive proteome constrained FBA to predict bacterial growth and metabolism with good quantitative accuracy.

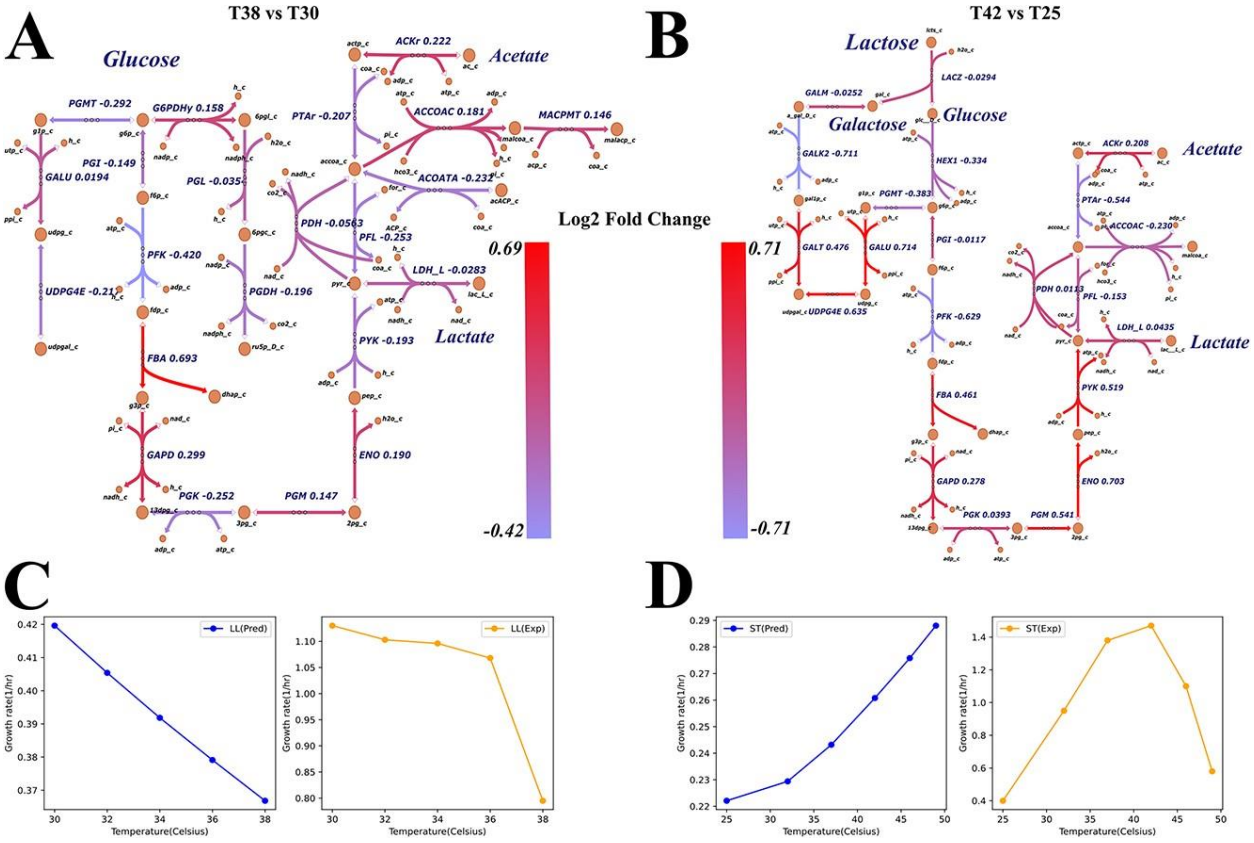

Figure 4.6. Prediction of bacteria metabolism at different temperatures. (A) Log2-fold change of predicted k_{cat} values for LL at 38 ℃ and 30 ℃ (38 ℃ vs 30 ℃). (B) Log2-fold change of predicted k_{cat} values for ST at 42 ℃ and 25 ℃ (42 ℃ vs 25 ℃). (C) Comparison of predicted (blue) and experimental (orange) growth rates of LL at 30, 32, 34, 36 and 38 ℃. (D) Comparison of predicted (blue) and experimental (orange) growth rates of ST at 25, 32, 37, 42, 46 and 49 ℃. LL: *Lactococcus lactis MG1363*; ST:*Streptococcus thermophilus LMG18311*; Exp: experimental value; Pred: predicted value. Reaction information can be found in **Appendix B, Tables B.1 & B.2**.

4.4 Conclusions and discussion

The expensive cost of obtaining enzyme k_{cat} values in wet lab stimulates the need of developing computational models to predict k_{cat}. Nevertheless, predicting temperature dependent k_{cat} is a challenging task, as temperature is not only a variable in the exponential factor of the Arrhenius equation, it also affects the activation energy of the enzyme catalyzed reaction, which is governed by the compound protein interaction [20]. To tackle the challenging task, this study constructed a CPI deep learning model called DLTKcat. DLTKcat used the bi-directional attention mechanism [27] to represent the interactions between compounds and proteins, and attention weights could capture important regions on protein sequences (**section 4.3.2**) and important atoms on substrates. Therefore, DLTKcat could obtain more information from pairs of compounds and proteins, in contrast to CPI models using unidirectional attention [16]. The use of both temperature and inverse temperature values facilitated the learning process of the neural network by representing features in the most biophysical relevant form to k_{cat} [20]. Also, oversampling on entries at low and high temperature ranges compensated for the imbalanced distribution of temperature values in the dataset (**see Appendix B, Figure B.3**). As a result, DLTKcat showed superior performance (log10-scale RMSE = 0.88, R2 = 0.66) than previously published models (e.g., EF-PreKcat) and robust accuracy for k_{cat} predictions for different conditions (e.g., metabolic contexts). In addition, feature shuffling demonstrated the contribution of temperature related features to this deep learning model.

By accurately predicting the effect of protein sequence mutations on the k_{cat} value of *Pyrococcus furiosus* Ornithine Carbamoyltransferase at different temperatures (**section 4.3.4**), DLTKcat exhibited its function in scoring the efficiency of in-silico designed enzyme protein sequences. Imaginably, the combination of DLTKcat and optimization algorithms (e.g., genetic programming) can become a computational tool to design site-specific mutagenesis to optimize enzyme catalysis, which will be more efficient than directed evolution that relies on random

mutagenesis. Nonetheless, the small variations of predicted enzyme k_{cat} values of WT and mutated enzymes in Figure 4.3B and Figure 4.5A suggested that the quantitative accuracy of DLTKcat still remains to be improved to precisely capture the effect of point mutations, although DLTKcat successfully predicted the significant reduction of enzyme k_{cat} caused by mutations (**Figure 4.3A**). Appending data entries of mutated enzymes to the training dataset might be a solution.

The second case study (**section 4.3.5**) of generating temperature dependent proteome constraints for metabolic modeling revealed the limitation of DLTKcat that its prediction error was not low enough to accurately model the response of cellular metabolism to temperature changes. Because all k_{cat} values of catabolic enzymes in ST and LL were predicted by DLTKcat, the propagation of error led to the inaccuracy of proteome constrained FBA. In short, deep learning can gap fill a few missing k_{cat} values in the metabolic network, as done in Li et al., 2022 [16], but the accuracy of proteome constrained FBA will not be high if most proteome constraints are based on predicted k_{cat} values.

To further improve the performance and utility of DLTKcat, including additional experimental conditions like pH, metal ion concentrations might be an approach, but the lack of data restricted existing models from accounting for those factors [22]. The inclusion of parameters such as pH and organismal information will largely reduce the size of the training dataset, because data entries containing those parameters are limited in BRENDA and SABIO-RK. Including the optimal enzyme temperature either from databases or predictions [54] might be able to enhance the temperature sensitivity of DLTKcat. The difference between the experimental temperature and optimal temperature could inform the model whether the temperature feature has a negative or positive effect on the k_{cat} value. However, the success of this approach depends on the accuracy of enzyme optimal temperature prediction, which was reported to have a RMSE around 2 [54,55].

After the publication of DLTKcat, some CPI models of enzyme k_{cat} were developed to further enhance the prediction accuracy. For example, DeepEnzyme (R2=0.6) used attention graph convolutional networks to extract features from both proteins and compounds[56]; CatPred (R2=0.608) included 3D protein structures as additional features [57]. However, the improvement of those models was not significant, and most of them could not account for condition parameters (e.g., temperature) except for MPEK [58]. MPEK could achieve R2=0.648 and log10-scale RMSE=0.594 on temperature and pH dependent enzyme k_{cat}, but the model interpretability was weak due to its lack of straightforward feature importance analysis functions (e.g., protein residue attention weight analysis in DLKcat [16] and DLTKcat [59]).

Overall, DLTKcat can provide accurate predictions of k_{cat} and account for the effect of temperature changes. Two case studies (3.3 and 3.4) have revealed potential applications of DLTKcat on protein engineering, bacterial phenotype prediction, etc. Additionally, DLTKcat can be easily modified to predict other temperature dependent CPIs, such as K_m [60,61]. With future improvements of the model framework, DLTKcat, as we envisage, will become a computational tool to quantitatively model the temperature dependence of biological systems, and contribute to the development of bioprocess digital twins.

References

1. Stephanopoulos G. Synthetic biology and metabolic engineering. ACS Synth Biol. 2012;1: 514–525.

2. Madhavan A, Arun KB, Binod P, Sirohi R, Tarafdar A, Reshmy R, et al. Design of novel enzyme biocatalysts for industrial bioprocess: Harnessing the power of protein engineering, high throughput screening and synthetic biology. Bioresour Technol. 2021;325: 124617.

3. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat

measurements. Proc Natl Acad Sci U S A. 2016;113: 3401–3406.

4. Nilsson A, Nielsen J, Palsson BO. Metabolic Models of Protein Allocation Call for the Kinetome. Cell Syst. 2017;5: 538–541.

5. Schomburg I, Jeske L, Ulbrich M, Placzek S, Chang A, Schomburg D. The BRENDA enzyme information system–From a database to an expert system. J Biotechnol. 2017;261: 194–206.

6. Wittig U, Rey M, Weidemann A, Kania R, Müller W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res. 2018;46: D656–D660.

7. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol. 2012;8: e1002575.

8. Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, et al. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proc Natl Acad Sci U S A. 2020;117: 23182–23190.

9. Bulović A, Fischer S, Dinh M, Golib F, Liebermeister W, Poirier C, et al. Automated generation of bacterial resource allocation models. Metab Eng. 2019;55: 12–22.

10. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, et al. Quantitative prediction of genome-wide resource allocation in bacteria. Metab Eng. 2015;32: 232–243.

11. Jahn M, Crang N, Janasch M, Hober A, Forsström B, Kimler K, et al. Protein allocation and utilization in the versatile chemolithoautotroph Cupriavidus necator. Elife. 2021;10. doi:10.7554/eLife.69019

12. Coppens L, Tschirhart T, Leary DH, Colston SM, Compton JR, Hervey WJ IV, et al. Vibrio natriegens genome-scale modeling reveals insights into halophilic adaptations and resource allocation. Mol Syst Biol. 2023;19: e10523.

13. Wendering P, Arend M, Razaghi-Moghadam Z, Nikoloski Z. Data integration across conditions improves turnover number estimates and metabolic predictions. Nat Commun. 2023;14: 1485.

14. Li S, Wan F, Shu H, Jiang T, Zhao D, Zeng J. MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. Cell Syst. 2020;10: 308–322.e11.

15. Kroll A, Engqvist MKM, Heckmann D, Lercher MJ. Deep learning allows genome-scale prediction of Michaelis constants from structural features. PLoS Biol. 2021;19: e3001402.

16. Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nature Catalysis. 2022; 1–11.

17. Lim S, Lu Y, Cho CY, Sung I, Kim J, Kim Y, et al. A review on compound-protein interaction prediction methods: Data, format, representation and model. Comput Struct Biotechnol J. 2021;19: 1541–1556.

18. Shin B, Park S, Kang K, Ho JC. Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, et al., editors. Proceedings of the 4th Machine Learning for Healthcare Conference. PMLR; 09--10 Aug 2019. pp. 230–248.

19. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of

compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019;35: 3329–3338.

20. Arroyo JI, Díez B, Kempes CP, West GB, Marquet PA. A general theory for temperature dependence in biology. Proc Natl Acad Sci U S A. 2022;119: e2119872119.

21. Kroll A, Rousset Y, Hu X-P, Liebrand NA, Lercher MJ. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. Nat Commun. 2023;14: 4139.

22. Yu H, Deng H, He J, Keasling J, Luo X. Highly accurate enzyme turnover number prediction and enzyme engineering with PreKcat. 2023. Available: https://www.researchsquare.com/article/rs-2749688/latest

23. Estévez SR. Robaina/BRENDApyrser: zenodo. 2022. doi:10.5281/zenodo.7026555

24. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28: 31–36.

25. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51: D1373–D1380.

26. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–D531.

27. Li M, Lu Z, Wu Y, Li Y. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. Bioinformatics. 2022;38: 1995–2002.

28. Landrum G. RDKit: Open-source cheminformatics. 2006. Google Scholar. 2006

[cited 5 Aug 2023]. Available: https://cir.nii.ac.jp/crid/1370004237630036224

29. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50: 742–754.

30. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. 2013 [cited 5 Aug 2023]. Available: http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

31. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG]. 2014. Available: http://arxiv.org/abs/1412.6980

32. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248.

33. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7: 74.

34. Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained Allocation Flux Balance Analysis. PLoS Comput Biol. 2016;12: e1004913.

35. Zeng H, Yang A. Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. Sci Rep. 2020;10: 4283.

36. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM, Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

37. Qiu S, Zeng H, Yang Z, Hung W-L, Wang B, Yang A. Dynamic metagenome-

scale metabolic modeling of a yogurt bacterial community. Biotechnol Bioeng. 2023. doi:10.1002/bit.28492

38. Schumacher R. Metabolic trade-offs arising from increased free energy conservation in Saccharomyces cerevisiae. Delft University of Technology. 2018. doi:10.4233/UUID:177E9F4C-F847-436D-9FD4-9ED97BA709D9

39. Flahaut NAL, Wiersma A, van de Bunt B, Martens DE, Schaap PJ, Sijtsma L, et al. Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. Appl Microbiol Biotechnol. 2013;97: 8729–8739.

40. Pastink MI, Teusink B, Hols P, Visser S, de Vos WM, Hugenholtz J. Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria. Appl Environ Microbiol. 2009;75: 3627–3633.

41. Chen J, Shen J, Ingvar Hellgren L, Ruhdal Jensen P, Solem C. Adaptation of Lactococcus lactis to high growth temperature leads to a dramatic increase in acidification rate. Sci Rep. 2015;5: 14199.

42. Vaningelgem F, Zamfir M, Adriany T, De Vuyst L. Fermentation conditions affecting the bacterial growth and exopolysaccharide production by Streptococcus thermophilus ST 111 in milk-based medium. J Appl Microbiol. 2004;97: 1257–1273.

43. Christiansen I, Hengstenberg W. Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system--two highly similar glucose permeases in Staphylococcus carnosus with different glucoside specificity: protein engineering in vivo? Microbiology. 1999;145 (Pt 10): 2881–2889.

44. Geertsma ER, Duurkens RH, Poolman B. The activity of the lactose transporter from Streptococcus thermophilus is increased by phosphorylated IIA and the action of beta-

galactosidase. Biochemistry. 2005;44: 15889–15897.

45. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, et al. Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. J Bacteriol. 2007;189: 3256–3270.

46. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, et al. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat Biotechnol. 2004;22: 1554–1558.

47. Simon WA, Hofer HW. Phosphofructokinases from Lactobacteriaceae. II. Purification and properties of phosphofructokinase from Streptococcus thermophilus. Biochim Biophys Acta. 1981;661: 158–163.

48. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45: D353–D361.

49. Roovers M, Sanchez R, Legrain C, Glansdorff N. Experimental evolution of enzyme temperature activity profile: selection in vivo and characterization of low-temperature-adapted mutants of Pyrococcus furiosus ornithine carbamoyltransferase. J Bacteriol. 2001;183: 1101–1105.

50. Fiala G, Stetter KO. Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100 ℃. Arch Microbiol. 1986;145: 56–61.

51. Harnett J, Davey G, Patrick A, Caddick C, Pearce L. Lactic Acid Bacteria | Streptococcus thermophilus. In: Fuquay JW, editor. Encyclopedia of Dairy Sciences (Second Edition). San Diego: Academic Press; 2011. pp. 143–148.

52. Callens M, Kuntz DA, Opperdoes FR. Kinetic properties of fructose bisphosphate aldolase from Trypanosoma brucei compared to aldolase from rabbit muscle and Staphylococcus aureus. Mol Biochem Parasitol. 1991;47: 1–9.

53. Plater AR, Zgiby SM, Thomson GJ, Qamar S, Wharton CW, Berry A. Conserved residues in the mechanism of the E. coli Class II FBP-aldolase. J Mol Biol. 1999;285: 843–855.

54. Gado JE, Beckham GT, Payne CM. Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. J Chem Inf Model. 2020;60: 4098–4107.

55. Li G, Rabe KS, Nielsen J, Engqvist MKM. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. ACS Synth Biol. 2019;8: 1411–1420.

56. Wang T, Xiang G, He S, Su L, Yan X, Lu H. DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D structures. bioRxiv. 2023. p. 2023.12.09.570923. doi:10.1101/2023.12.09.570923

57. Boorla VS, Maranas CD. CatPred: A comprehensive framework for deep learning in vitro enzyme kinetic parameters kcat, Km and Ki. bioRxiv. 2024. p. 2024.03.10.584340. doi:10.1101/2024.03.10.584340

58. Wang J, Yang Z, Chen C, Yao G, Wan X, Bao S, et al. MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. Brief Bioinform. 2024;25. doi:10.1093/bib/bbae387

59. Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-

dependent enzyme turnover rates. Brief Bioinform. 2023;25. doi:10.1093/bib/bbad506

60. Quinlan AV. The Thermal Sensitivity of Michaelis-Menten Kinetics as a Function of Substrate Concentration. J Franklin Inst. 1980;310: 325–342.

61. Maggi F, M. Tang FH, Riley WJ. The thermodynamic links between substrate, enzyme, and microbial dynamics in Michaelis-Menten-Monod kinetics. Int J Chem Kinet. 2018;50: 343–356.

Chapter 5 Systematic elucidation of independently modulated genes in *Lactiplantibacillus plantarum* reveals a trade-off between secondary and primary metabolism

Published as: Qiu S, Huang Y, Liang S, Zeng H, Yang A. Systematic elucidation of independently modulated genes in *Lactiplantibacillus plantarum* reveals a trade-off between secondary and primary metabolism. Microb Biotechnol. 2024;17: e14425.

Sizhe Qiu, as the first author, conceptualized the workflow, performed and analyzed the independent component analysis of transcriptomic data, and produced the first draft of the paper. Other listed authors contributed to the revision of the paper.

Summary

Lactiplantibacillus plantarum is a probiotic bacterium widely used in food and health industries, but its gene regulatory information is limited in existing databases, which impedes the research of its physiology and applications. To obtain a better understanding of the transcriptional regulatory network of *L. plantarum*, independent component analysis (ICA) of its transcriptomes was used to derive 45 sets of independently modulated genes (iModulons). Over 50% of identified iModulons were annotated for associated transcription factors (TFs) and functional pathways, and stress-active iModulons in response to different growth conditions (e.g., acidic pH, change of the carbon source) were characterized in detail. Eventually, the

analysis of iModulon activities reveals a trade-off between regulatory activities of primary metabolism and stress induced secondary metabolism (i.e., EPS biosynthesis) in *L. plantarum*.

Reporting the third research project in this PhD study, this chapter elucidates a transcriptomic trade-off between growth-coupled primary metabolism and stress-induced secondary metabolism in *L. plantarum*, which provides the theoretical foundation for the research work in Chapter 6 that investigates the acid stress induced proteome allocation.

5.1 Introduction

The transcriptional regulatory network (TRN) of a bacterium consists of all regulatory interactions between its transcription factors (TFs) and genes [1]. TFs, also referred to as sequence-specific DNA-binding factors, sense external signals and then bind to promoter regions of operons to regulate gene expression levels [2]. To identify regulatory interactions between TFs and genes, the most commonly used experimental method is chromatin immunoprecipitation followed by sequencing (CHiP-seq) [3]. In CHiP-seq, antibodies are used to select TF proteins, and then DNA bound to TF proteins will be purified. DNA sequencing for the DNA-TF protein complex will determine the binding site on the genome. A group of genes with binding sites of the same TF are considered as a regulon. However, the drawbacks of CHiP-seq lie in its high cost, time-intensive nature, and challenges in capturing the diverse growth conditions of bacteria [4].

In recent years, many computational methods of *in-silico* reconstruction of TRN have been developed, such as coexpression network analysis [5] or supervised learning-based methods (e.g., GENIE3 [6]). One of the most popular methods to reconstruct TRN is using independent component analysis (ICA) to decompose the gene expression matrix, which consists of transcriptomic data of different samples, into sets of independently modulated genes, called iModulons (IMs) [7]. ICA has been demonstrated as the best TRN inference

method in the comparison of 42 different methods, as unsupervised matrix decomposition can provide an unbiased and global overview of biological datasets [8]. Apart from derived IMs, ICA can also quantify IM activities in different samples. Unlike CHiP-seq being a 'bottom-up' method, ICA follows a 'top-down' approach. ICA has been extensively applied to study and improve the understanding of many bacteria's TRNs. For example, ICA of *Vibrio natriegens* transcriptomes unveils the genetic basis of its natural competency [9]. ICA has also been used to discover therapeutic strategies for *Streptococcus pyogenes* by identifying carbon sources that control the expression of hemolytic toxins [10].

Lactiplantibacillus plantarum is a gram-positive lactic acid bacterium that can be found in diverse ecological niches [11]. It has been widely used in food and health industries. For instance, it is the major bacterium involved in the fermentation of mozzarella cheese [12]; *L. plantarum*-derived exopolysaccharides (EPSs) have various probiotic effects [13] and anticancer properties [14]. Due to the importance of *L. plantarum* in various biological processes, such as dairy product fermentation, its gene expression regulation has received interest in several studies. For example, Jung and Lee identified differentially expressed genes when *L. plantarum* was in the acidic condition [15]. Unlike most studies focusing on single regulatory genes, Wels *et al.* reconstructed the gene regulatory network of *L. plantarum* on the basis of correlations between gene expression levels and conserved regulatory motifs [16]. Nonetheless, the regulon information of *L. plantarum* in RegPrecise [17] only recorded 47 regulons and 210 TF binding sites, in contrast to 624 and 943 TF binding sites recorded for *Bacillus subtilis* and *Escherichia coli*, respectively. The lack of gene regulatory information hinders the study of *L. plantarum*'s physiology and rational engineering of its cellular pathways.

Considering the value of *L. plantarum* in industry and research as well as the limited understanding of its TRN, this study managed to infer undiscovered regulatory interactions using ICA decomposition of the gene expression matrix, and to further investigate how *L. plantarum* respond to different growth conditions (e.g., acid stress). Moreover, this study,

through the analysis of IM activities, explored the growth strategy of *L. plantarum*, in terms of how it balances different biological processes (e.g., energy generation, carbohydrate metabolism, stress responses).

5.2 Methods

5.2.1 Data acquisition and preprocessing

The transcriptomic data used in the study were obtained from 4 independent studies that included various experimental conditions: response to pH decrease from 6.2 to 5.0 [15], treatment with N-3-oxododecanoyl homoserine lactone (a quorum sensing molecule) [18], contrasting habitats (e.g., bee extract) [19], and change of carbon sources [20]. The metadata of sample conditions can be found in **Appendix C, Table C.1**. In the data from the selected 4 studies, genes were all annotated based on the genome assembly of *L. plantarum* WCFS1 (ASM20385v3) [21]. All transcriptomic sequencing reads were normalized as RPKM (Reads Per Kilobase Million). Then, all samples were merged as a compendium of transcriptomic data (100 samples and 3000 genes). Before independent component analysis (ICA) was undertaken, the merged dataset was first log-transformed, and then centered by subtracting the expression levels of the reference condition (i.e., wt_pH6.2 in **Appendix C, Table C.1**). The data quality was demonstrated by the higher Pearson correlation coefficients (PCCs) between replicates than PCCs between non-replicates [22] (**Figure 5.1A**).

5.2.2 Determination of iModulons

ICA decomposition of the merged dataset (i.e., the expression matrix, 100 samples and 3000 genes) was conducted using scripts in precise-db (https://github.com/SBRG/precise-db) [23]. The FastICA algorithm in Scikit-Learn (v0.20.3) [24] was used to calculate independent

components with 100 iterations with a tolerance of 10^{-7}, $log(cosh(x))$ as the contrast function, and parallel search algorithm. The OptICA method was used to determine the optimal number of independent components [25]. The outputs of ICA were the iModulon matrix (M matrix, 3000 genes and 53 IMs) and Activity matrix (A matrix, 53 IMs and 100 samples) (**Figure 5.1B**). The M and A matrices can be found in https://github.com/SizheQiu/LPiModulons/tree/main/data/IMdata.

Gene weights in each column (for the corresponding IM) of the M matrix were used to determine each gene's IM membership. The threshold of gene weight absolute values for each IM was computed based on D'Agostino's K^2 test using the PyModulon package (https://github.com/SBRG/pymodulon) [7]. The default K^2-statistic cutoff of 550 was used. The genes with weight absolute values above the threshold were the member genes of the IM. Before annotation, IMs were labeled as IM-1 to 53.

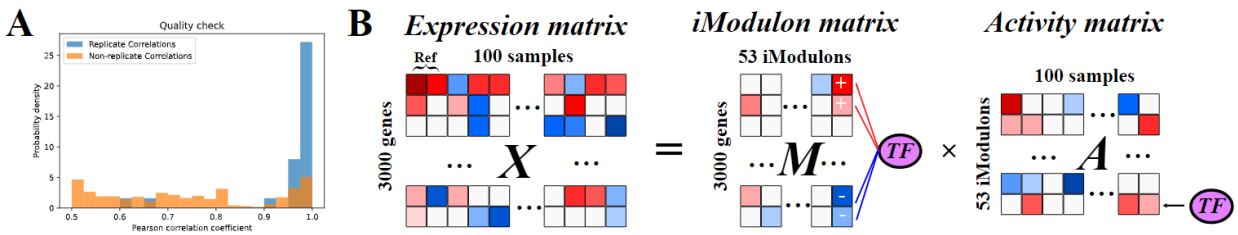

Figure 5.1. ICA decomposes the compendium of transcriptomic data to 45 nonempty iModulons. (A) Quality check of transcriptomic data with PCCs. Blue: replicate correlations; Yellow: non-replicate correlations. (B) Schematic illustration of ICA applied to the gene expression matrix.

5.2.3 Annotation of iModulons via regulon enrichment analysis

Regulons of *Lactiplantibacillus plantarum* WCFS1 were obtained from RegPrecise [17]. IMs that overlap with regulons were annotated via regulon enrichment analysis. The set of genes in each IM was compared to each regulon using the two-sided Fisher's exact test (False Discovery Rate (FDR) < 10^{-5}) [7]. After regulon enrichments were computed for IMs, regulatory

annotations were manually determined based on the Venn diagrams of IMs and regulons (**see Appendix C, Figure C.1**). In addition to IMs associated with only one regulon (e.g., PyrR IM (IM-36)), there were two different annotation expressions for combined regulon enrichments: intersection (+) and union (/). If a specific combinatorial regulation (genes controlled by multiple regulators) was observed in the Venn diagram of the IM and enriched regulons, then the IM was annotated with regulators linked by "+" (e.g., MalR+MdxR IM (IM-47)). Otherwise, "/" was used (e.g., ArgR/MleR IM (IM-26)).

5.2.4 Annotation of iModulons via motif comparison

IMs that do not overlap with known regulons were annotated via motif discovery and motif comparison. The assumption behind such an approach was that most genes in an IM are regulated together by single or multiple TFs, as in *E. coli* and *V. natriegens* [7,9], and thus their upstream non-coding regions contain similar binding sites. If a coding gene's 200 bp upstream region does not overlap with another gene [26] and BDGP Neural Network Promoter Prediction [27] predicted this region to be a possible promoter (probability score > 0.8), then this 200 bp upstream region was used to search for sequence motifs using MEME [28]. Motif comparison by TOMTOM [29] then determined the most possible TF based on the similarity of found motifs and TF binding site motifs in databases (e.g., RegTransBase [30]). The p-value and E-value thresholds set in TOMTOM were 0.05 and 10. To further validate whether genes in the IM are regulated by the found TF, PCCs of the expression levels of the TF gene and IM genes were computed. If the gene had significant correlations (p-value < 0.05) with most genes in the IM, then the TF would be used to annotate the IM.

5.3 Results

5.3.1 Regulatory and functional annotations of identified iModulons

The derived 53 IMs account for 85% explained variance of the gene expression matrix. In each IM, genes with absolute values of weights higher than the threshold are determined as IM member genes (**see Methods, section 5.2.2**). The details of IM member genes can be found in https://github.com/SizheQiu/LPiModulons/tree/main/data/IMdata/ as IM_genes.csv. Among 53 IMs, 45 are nonempty and most IMs' sizes are within 20 (**Figure 5.2A**). Only 17% IM member genes overlap with genes in known regulons (**Figure 5.2B**), and hence only 13 IMs could be annotated via regulon enrichment (**Figure 5.2C**). The details of regulatory annotations can be found in https://github.com/SizheQiu/LPiModulons/blob/main/data/IMdata/IM_annotation.csv.

For the 13 IMs annotated with enriched regulons, most of them have either high recall or high precision (cutoff = 0.6) (**Figure 5.2D**). Venn diagrams showing regulon enrichments in IMs are provided in **Appendix C, Figure C.1**. High recall means that the overlap (of IM and regulon) has high coverage of the regulon, while high precision means that the overlap has high coverage of the IM. IMs with low recall and low precision are considered to be incompletely matched with regulons, but that does not necessarily mean the IM's regulatory annotation is inaccurate. For example, the remaining 3 genes in CopR IM that are not included by the current CopR regulon of *Lactiplantibacillus plantarum* WCFS1 are lp_3055(copA), lp_3057(copper-binding protein) and lp_3058(copper-binding protein), but they are included by the CopR regulon of other closely related lactic acid bacteria (e.g., *Lactococcus lactis subsp. lactis* Il1403) [31].Therefore, the low recall and precision are sometimes resulted by the incompleteness of currently known regulons.

In addition to IMs associated with regulons, there are 11 IMs annotated via motif search and comparison (**Figure 5.2C, Figure C.2**). Two representative examples are NagC IM (IM-8)

and McbR IM (IM-31) (**Figure 5.2E**). Their regulatory annotations are validated by significant correlations between expression levels of TF genes and IM activities (**Figure 5.2F**). The remaining 21 IMs (**Figure 5.2C**) cannot be annotated via motif search and comparison either because the IM does not contain multiple possible promoter sequences for motif search (e.g., IM-19) or TOMTOM (**see Methods, section 5.2.4**) fails to find a TF binding site motif with a high similarity to the found motif (e.g., IM-6).

IMs were also annotated with enriched functional pathways (**see Appendix C, section C.1.1**), and the details of functional annotations can be found in https://github.com/SizheQiu/LPiModulons/blob/main/data/IMdata/IM_annotation.csv. Apart from the uncharacterised group, 3 dominant functions of derived IMs are carbohydrate metabolism, prophage proteins and transcription (**Figure 5.2G**). Fur/LexA IM (IM-1) was functionally annotated as “Stress”, as LexA has already been found as a TF for stress response [32]. IM-19 was annotated as ‘Translation’, because genes in IM-19 were all ribosomal genes (e.g., rplV (lp_1039), large ribosomal subunit protein uL22). 12% (scaled with IM sizes) of IMs are uncharacterized in functional annotation due to the lack of enriched functional pathways.

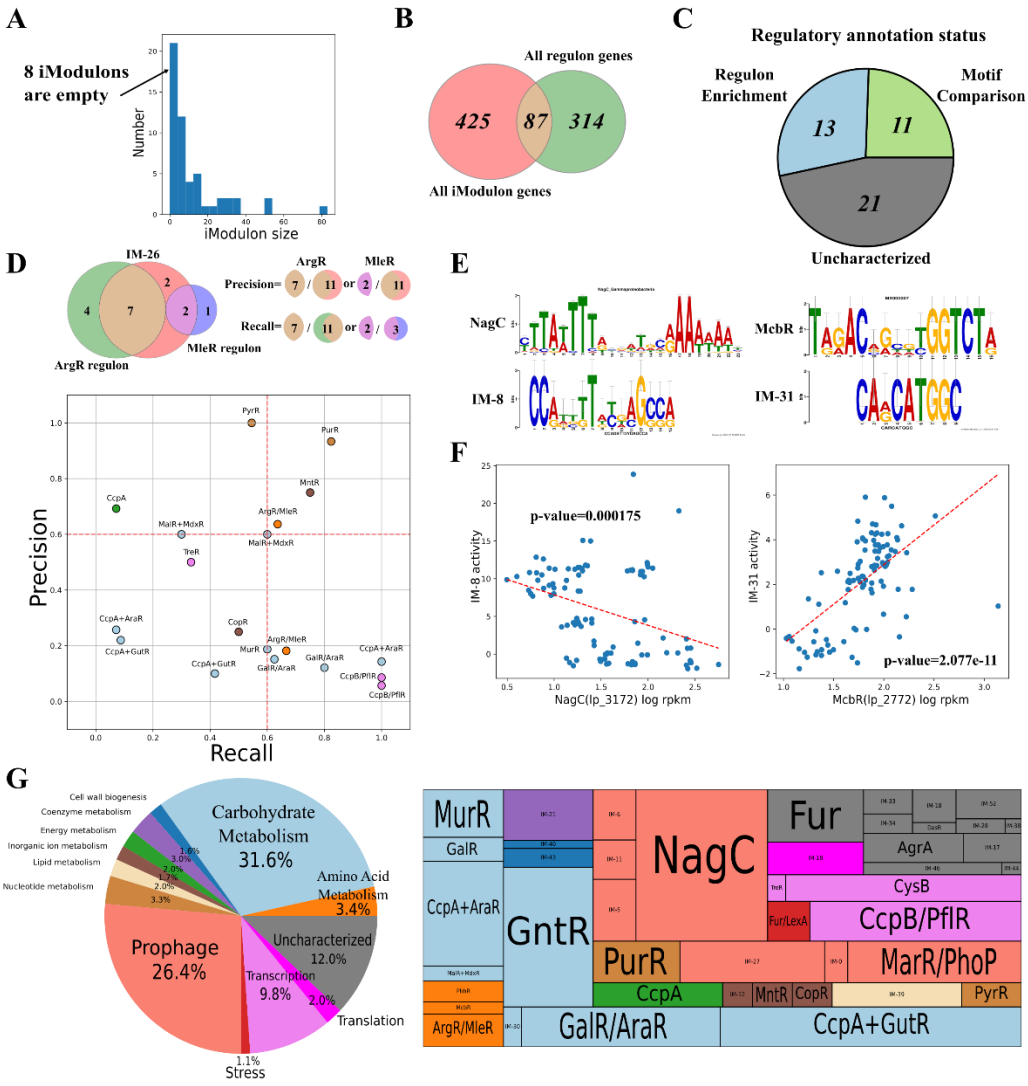

Figure 5.2. Regulatory and functional pathway annotations of IMs. (A) The histogram of IM sizes, 45 out of 53 IMs are nonempty. (B) The Venn diagram of all IM genes and regulon genes. 87 genes in IMs are contained in known regulons. (C) The pie chart of regulatory annotation status. Blue: regulon enrichment; Green: motif comparison; Grey: uncharacterized. (D) Recall and precision of IMs with matched regulons. (E) Motif comparison of IM-8 and IM-31. (F) The significant correlations between IM activities and gene expression levels of associated TFs identified via motif comparison for IM-8 and IM-31 (p-value < 0.05). (G) The pie chart and treemap of functional annotations of IMs, the size of each fraction is scaled with the IM size.

5.3.2 Comparison between iModulons and regulons

The difference between IMs and regulons can provide undiscovered regulatory information. Regulon enrichments of some IMs show combinatorial regulations of multiple TFs, such as MalR+MdxR IM. Based on the genomic organization, 6 genes in the region between 151222bp and 158185bp belong to the same operon (**Figure 5.3A**). While mdxE (lp_0175), mdxG (lp_0177) and lp_0178 are already included by both MalR and MdxR regulons, MalR+MdxR IM also captures the combinatorial regulatory signals for malS (lp_0179) and msmX (lp_0180), which share the same promoter with genes in the overlap of MalR and MdxR regulons. All genes in MalR+MdxR IM are involved in maltose/maltodextrin metabolism, which is the biological process regulated by MalR and MdxR [32,33].

IMs also have the ability to identify genes with strong regulatory interactions with TFs from known regulons. For example, the Pearson correlation coefficients (PCCs) between TF genes and genes in the overlap (of the IM and regulon) exhibit higher distributions compared to those of genes in the regulon for ArgR and CcpA (**Figure 5.3BC**). Nevertheless, the overlap does not always show stronger regulatory interactions. For example, genes in PyrR IM do not have significantly higher PCCs with the PyrR gene than with the genes in PyrR regulon (Figure 3D).

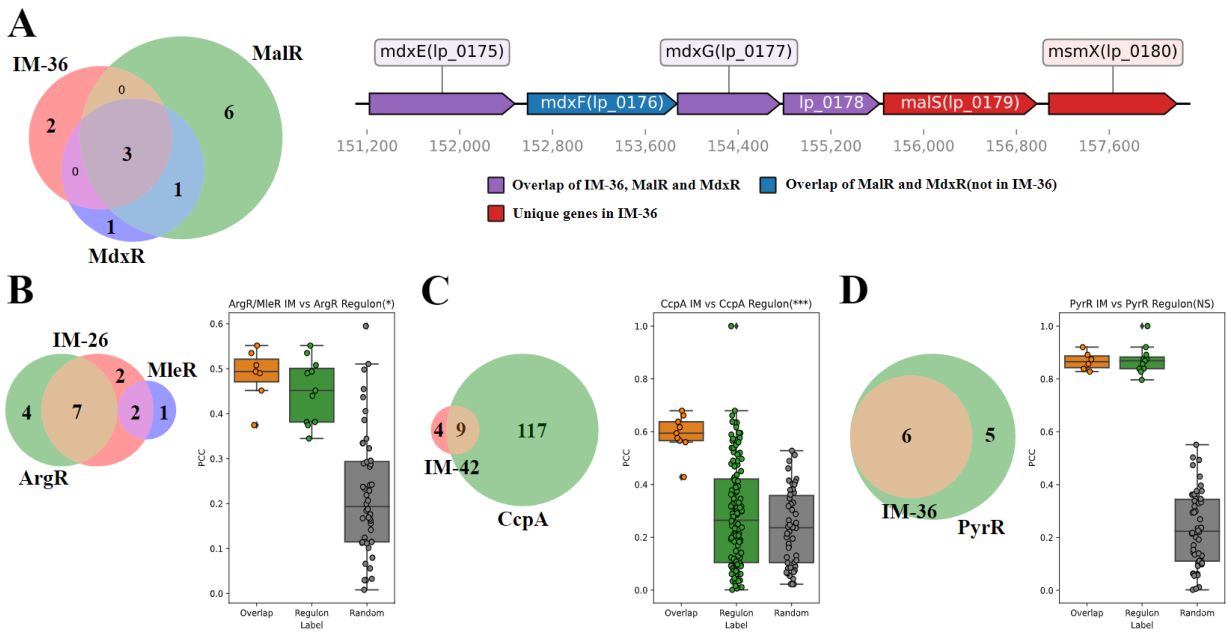

Figure 5.3. Comparison between IMs and regulons of *Lactiplantibacillus plantarum*. (A) Left: The Venn diagram of MalR+MdxR IM (IM-47) and Mdx, MalR regulons; Right: Genomic organization of genes in MalR+MdxR IM. (B-D) Comparison of PCCs of the gene expression levels of TF gene and genes in the overlap of IM and regulon (orange), regulon (green) and randomly sampled genes (gray) for (B) ArgR/MelR, (C) CcpA, and (D) PyrR IMs.

5.3.3 Active iModulons in response to different growth conditions

In addition to the M matrix, the A matrix is another output of ICA decomposition, which reveals IM activities of *L. plantarum* under different growth conditions. In response to acid stress (in terms of pH decrease), 4 active IMs are observed: Fur/LexA IM, CopR IM, McbR IM and PyrR IM (**Figure 5.4A**). IM activities of all 4 active IMs identified increase with the decrease of pH (**Figure 5.4B-E**). The gene expression levels of Fur (lp_3247) and LexA (lp_2063) both decrease with the decrease of pH, though the trends over three pH values are not consistently decreasing (**Figure 5.4F**). Genes in Fur/LexA IM are related to the biosynthesis of exopolysaccharide (EPS), an important secondary metabolite [34], including lp_0302 (extracellular transglycosylase), lp_0304 (extracellular transglycosylase), lp_2809 (extracellular

protein of unknown function), lp_2810 (glycosyl hydrolase, family 25), lp_2845 (extracellular transglycosylase, with LysM peptidoglycan binding domain), lp_3014 (extracellular transglycosylase, with LysM peptidoglycan binding domain) and lp_3050 (extracellular transglycosylase, membrane-bound). Oppositely, the gene expression levels of CopR (lp_3365), McbR (lp_2772) and PyrR (lp_2704) increase with the decrease of pH (**Figure 5.4G-I**). CopR, McbR and PyrR regulate copper homeostasis, amino acid metabolism and pyrimidine metabolism, respectively.

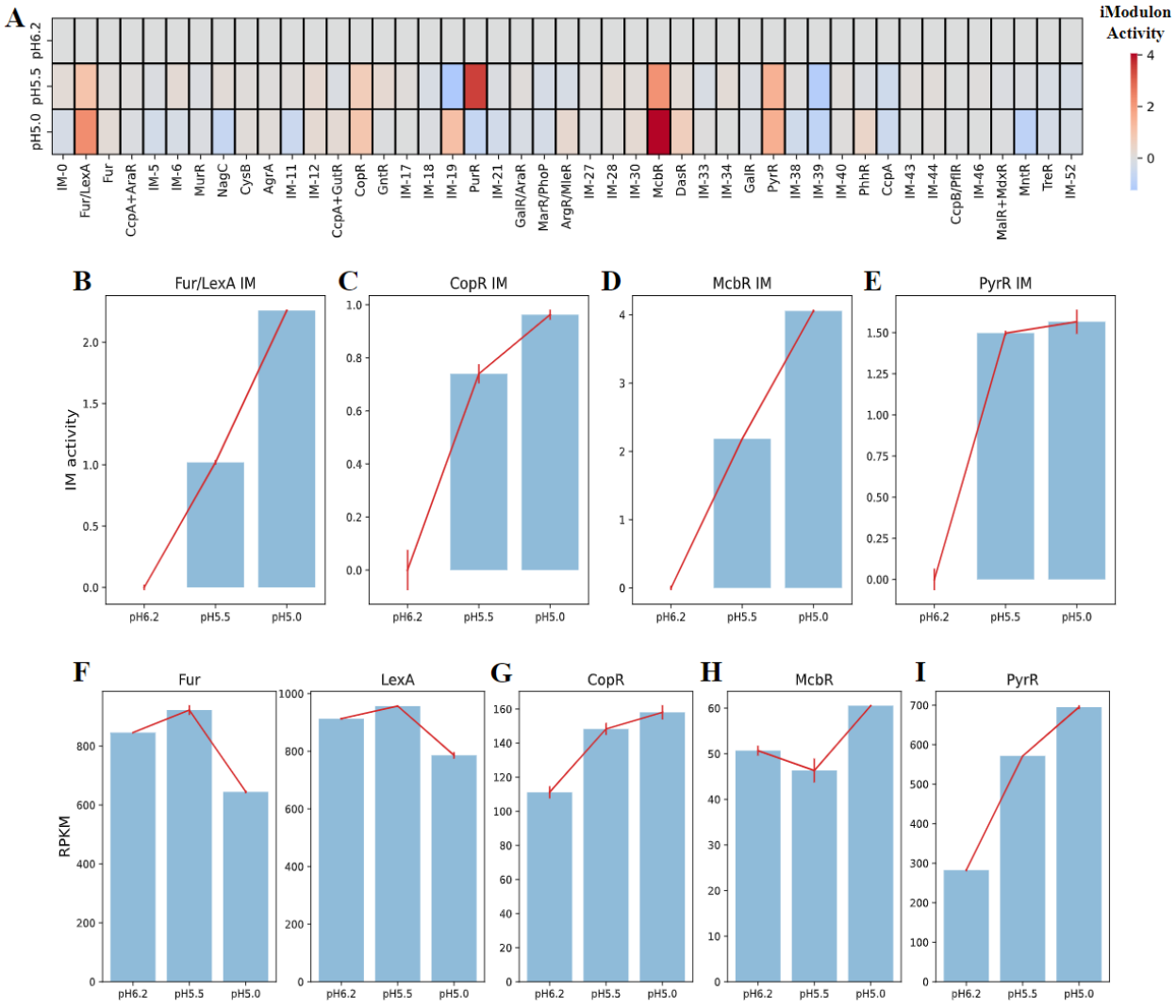

Figure 5.4. Identification of active IMs under the acidic condition. (A) The heatmap of IM activities at pH=6.2, 5.5 and 5.0. (B) IM activities of Fur/LexA IM at different pH values. (C) IM activities of CopR IM at different pH values. (D) IM activities of McbR IM at different pH values.

(E) IM activities of PyrR IM at different pH values. (F) The expression levels of Fur (lp_3247) and LexA (lp_2063) at different pH values. (G) The expression levels of CopR (lp_3365) at different pH values. (H) The expression levels of McbR (lp_2772) at different pH values. (I) The expression levels of PyrR (lp_2704) at different pH values.

To further characterize acid-active IMs, regulatory networks are reconstructed as weighted correlation networks, and genomic organizations of genes in those IMs are further investigated. Fur/LexA IM, based on gene locations and the weighted correlation network, appear to contain two operons regulated by Fur and LexA separately: lp_0302 and lp_0304 regulated by Fur; lp_2809 and lp_2810 regulated by LexA (**Figure 5.5AB**). The correlations between Fur and lp_0302, lp_0304 and lp_3014 are all negative, consistent with the previous finding that Fur is a repressor [35] (**Figure 5.5A**). The correlations between LexA and its regulated genes (i.e., lp_2809, lp_2810 and lp_3050) are positive, indicating that LexA functions as an activator to those genes (**Figure 5.5A**). For CopR, McbR and PyrR IMs, the correlations between TFs and regulated genes are all positive, suggesting that associated TFs all function as activators (**Figure 5.5CEG**). Unlike Fur/LexA IM, member genes of those three IMs are mainly in single operons (**Figure 5.5DFH**).

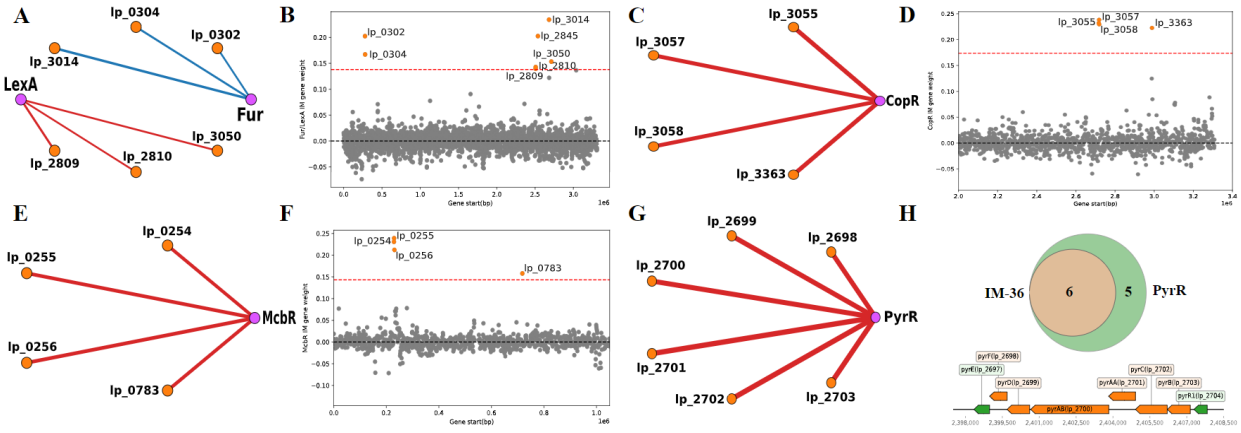

Figure 5.5. Characterization of genes in acidity-active IMs. (A) The weighted correlation network of Fur, LexA and genes in Fur/LexA IM (IM-1). (B) Gene weights and gene locations of Fur/LexA

IM. (C) The weighted correlation network of CopR and genes in CopR IM (IM-15). (D) Gene weights and gene locations of CopR IM. (E) The weighted correlation network of McbR and genes in McbR IM (IM-31). (F) Gene weights and gene locations of McbR IM. (G) The weighted correlation network of PyrR and genes in PyrR IM (IM-36). (H) Genomic organization of genes in PyrR IM (IM-36). Orange: overlap of IM and regulon; Green: genes only in the regulon. Edge weights in weighted correlation networks are scaled to PCCs. Red: positive correlation; Blue: negative correlation; Orange node: the gene in the IM; Purple node: the TF gene.

On the other hand, the change of carbon sources can result in transcriptional regulations of carbohydrate metabolism [36], where GntR IM (IM-16) was found to be the most active IM in this study (**Figure 5.6A**). Genes in GntR IM mainly encode for the utilization of different carbon sources (e.g., pts9C (lp_0576), uptake of mannose; panD (lp_0579), aspartate 1-decarboxylase) and the biosynthesis of capsular polysaccharide (CPS) in the cell wall (e.g., cps1F (lp_1182), CPS biosynthesis protein CpsC). The biosynthesis of CPS is a part of primary metabolism (cellular biomass formation), different from that of EPS, belonging to secondary metabolism [37]. GntR IM is annotated via motif comparison (**see Appendix C, Figure C.2**) due to the lack of regulon information, and hence it is hard to determine which TF in the GntR family regulate genes in this IM. Top 4 GntR TF genes with highest PCCs with activities of GntR IM are lp_2615, lp_2651, lp_3633 and lp_0563 (**Figure 5.6B**). The PCCs between TF genes and genes in GntR IM show that lp_2615 and lp_0563 have significant negative correlations with genes in GntR IM, while lp_2651 and lp_3633 have significant positive correlations with genes in GntR IM (**Figure 5.6C**), which are consistent with the PCCs (**Figure 5.6B**). Possibly, genes in GntR IM are regulated by multiple GntR family TFs.

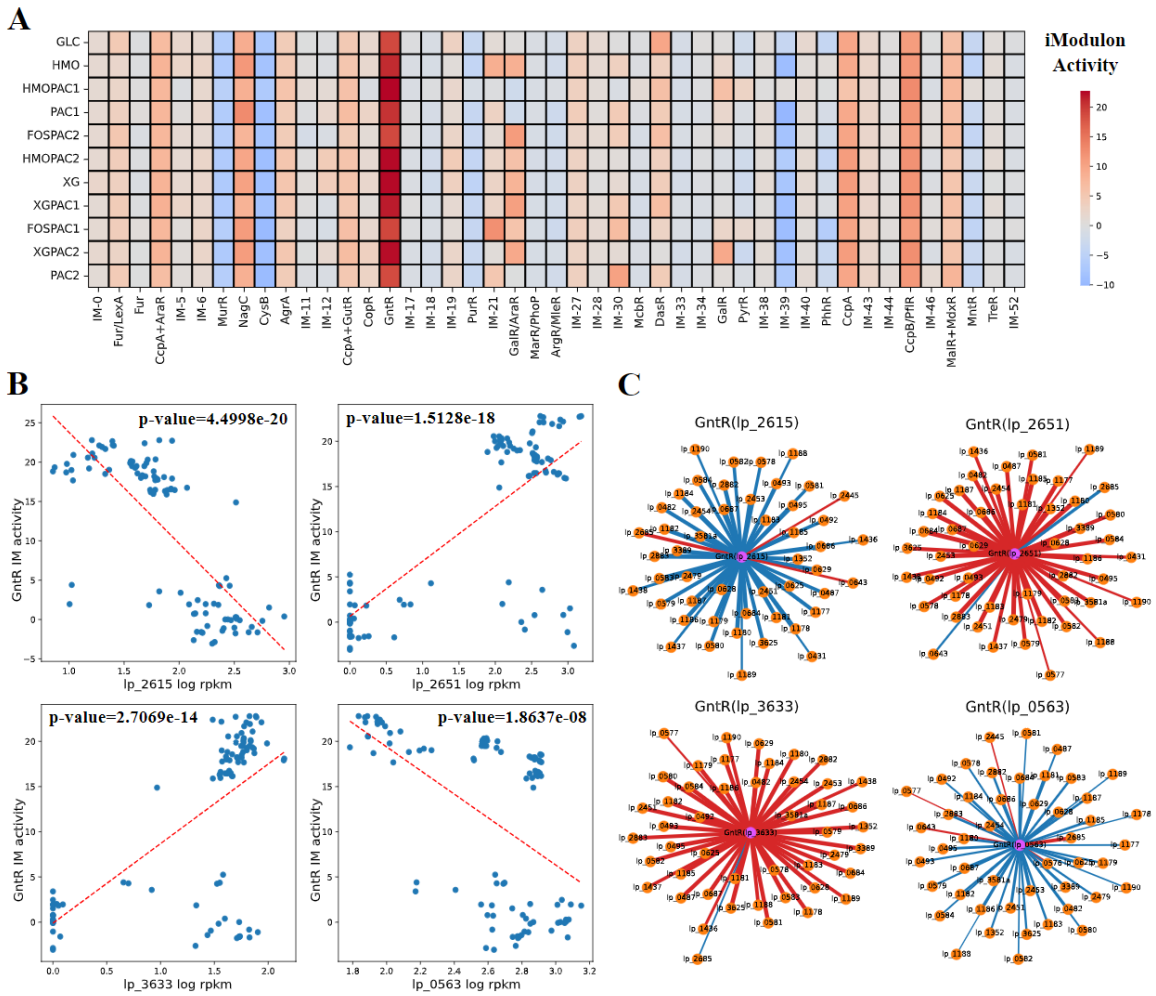

Figure 5.6. Identification of the most active IM in response to different carbon sources: GntR IM (IM-16). (A) The heatmap of IM activities with different carbon sources. GLC: glucose; HMO: human milk oligosaccharides; PAC1: proanthocyanidin fraction 1; PAC2: proanthocyanidin fraction 2; FOS: fructooligosaccharides; XG: xyloglucans. Detailed information can be found in Özcan *et al.*, 2021 [20]. (B) The correlations between expression levels of 4 GntR family TF genes and GntR IM activities (p-value < 0.05). Red dashed line: linear fit. (C) The weighted correlation networks of 4 GntR family TF genes and genes in GntR IM (p-value < 0.05). Edge weights are scaled to PCCs. Red: positive correlation; Blue: negative correlation; Orange node: the gene in the IM; Purple node: the TF gene.

5.3.4 The trade-off between primary and secondary metabolism revealed by iModulon activities

Member genes of IMs derived in this study encode connected reactions in one or several metabolic pathways, and those reactions were visualized as networks (**see Appendix C, section C.1.2**) to investigate the links between IMs and cellular metabolism (**Figure 5.7**). For acid-active IMs identified in section 3.3, genes in McbR IM and PyrR IM encode for the biosynthesis of L-cysteine and uridine monophosphate, respectively (**Figure 5.7AB**). EPS biosynthetic reactions encoded by genes in Fur/LexA IM and copper homeostasis encoded by genes in CopR IM are currently not included by model iBT721.

Next, 4 representative IMs, namely ArgR/MleR IM, CcpA IM, GntR IM, and GalR/AraR IM, functionally annotated for amino acid metabolism, energy metabolism and carbohydrate metabolism are selected to reconstruct metabolic pathways encoded by their member genes (**Figure 5.2G**). ArgR/MleR IM member genes encode for the biosynthesis of N-Acetyl-L-glutamate 5-semialdehyde from L-glutamine (**Figure 5.7C**). CcpA IM, as an IM for energy metabolism, contains a part of glycolysis, the conversion of glycerol to dihydroxyacetone phosphate and phosphorylation of nucleosides (**Figure 5.7D**). GntR IM member genes mainly encode for CPS biosynthesis, from the activation of monosaccharides to the polymerization as explained in section 3.3 (**Figure 5.7E**). Two important carbohydrate metabolic pathways, namely galactose metabolism and pentose phosphate pathway are contained by GalR/AraR IM (**Figure 5.7F**).

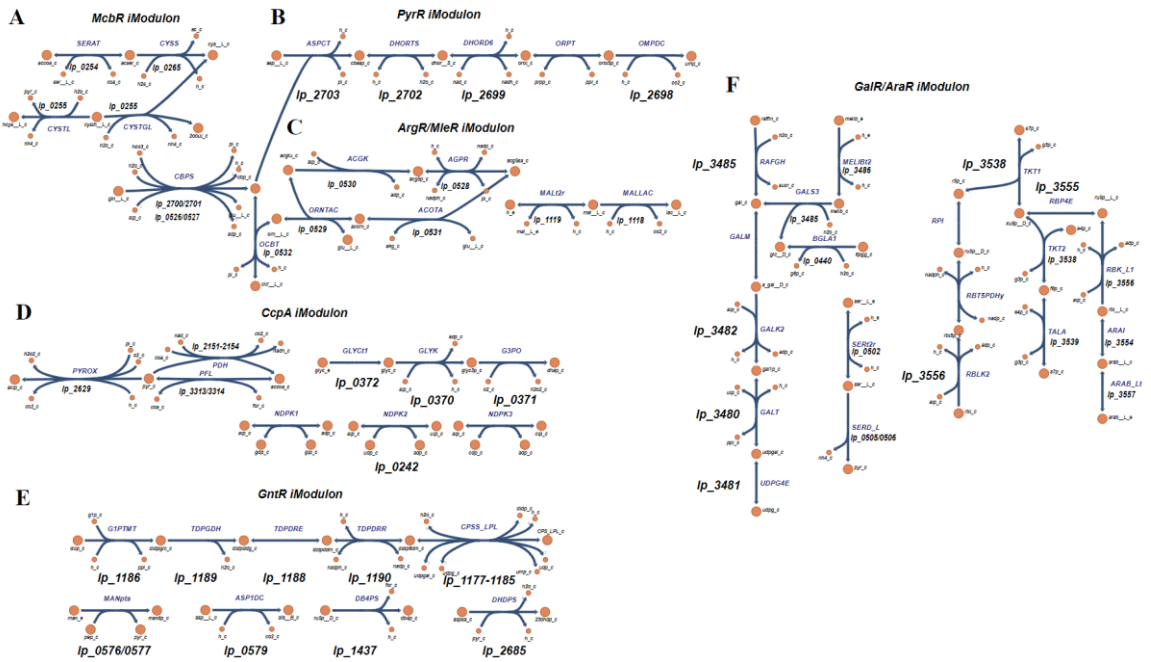

Figure 5.7. Metabolic pathways encoded by IM member genes. Reaction information (names, associated genes and IMs) can be found in **Appendix C, Table C.2**. Reaction abbreviations are adopted from the BIGG database (http://bigg.ucsd.edu/) [38]. (A) McbR IM. (B) PyrR IM. (C) ArgR/MleR IM. (D) CcpA IM. (E) GntR IM. (F) GalR/AraR IM.

In contrast to Fur/LexA IM controlling secondary metabolism (EPS biosynthesis induced by acid stress) as shown in section 3.3, ArgR/MleR, CcpA, GntR and GalR/AraR IMs (metabolic pathways visualized in **Figure 5.7C-F**) regulate primary metabolism. To investigate the relationship between regulatory activities of two branches of cellular metabolism, PCCs were computed for the activities of Fur/LexA IM and 4 IMs for primary metabolism (**Figure 5.8A-D**). Significant inverse correlations between the activity of Fur/LexA IM and activities of ArgR/MleR IM, CcpA IM, GntR IM and GalR/AraR IM can be observed, suggesting a trade-off between the regulatory activities of secondary and primary metabolisms. *Lactiplantibacillus plantarum* in acidic media (e.g., bee extract (pH=4.7), tomato juice (pH=3.5), **see Appendix C, Table C.1**) have higher Fur/LexA IM activities and lower IM activities of the 4 IMs for primary metabolism

than those in relatively neutral media (e.g., fecal extract (pH=5.9), **see Appendix C, Table C.1**). Therefore, the balance between regulations of EPS biosynthesis and primary metabolism in *Lactiplantibacillus plantarum* appears to significantly depend on the acidity of extracellular environments.

To assess whether a trade-off relationship also exists between gene expression levels (in addition to regulatory activities) of secondary and primary metabolism, PCCs were computed between the total expression levels of genes in Fur/LexA IM (EPS biosynthetic genes) and (i) all glycolytic genes (central carbon catabolism) (**Figure 5.8E**), and (ii) genes in Translational IM (IM-19, ribosomal genes) (**Figure 5.8F**). An inverse correlation between gene expression levels of EPS biosynthetic genes and glycolytic genes is also observed (**Figure 5.8E**), though the correlation is not statistically significant. For EPS biosynthetic genes versus ribosomal genes, there is no inverse correlation between them (**Figure 5.8F**).

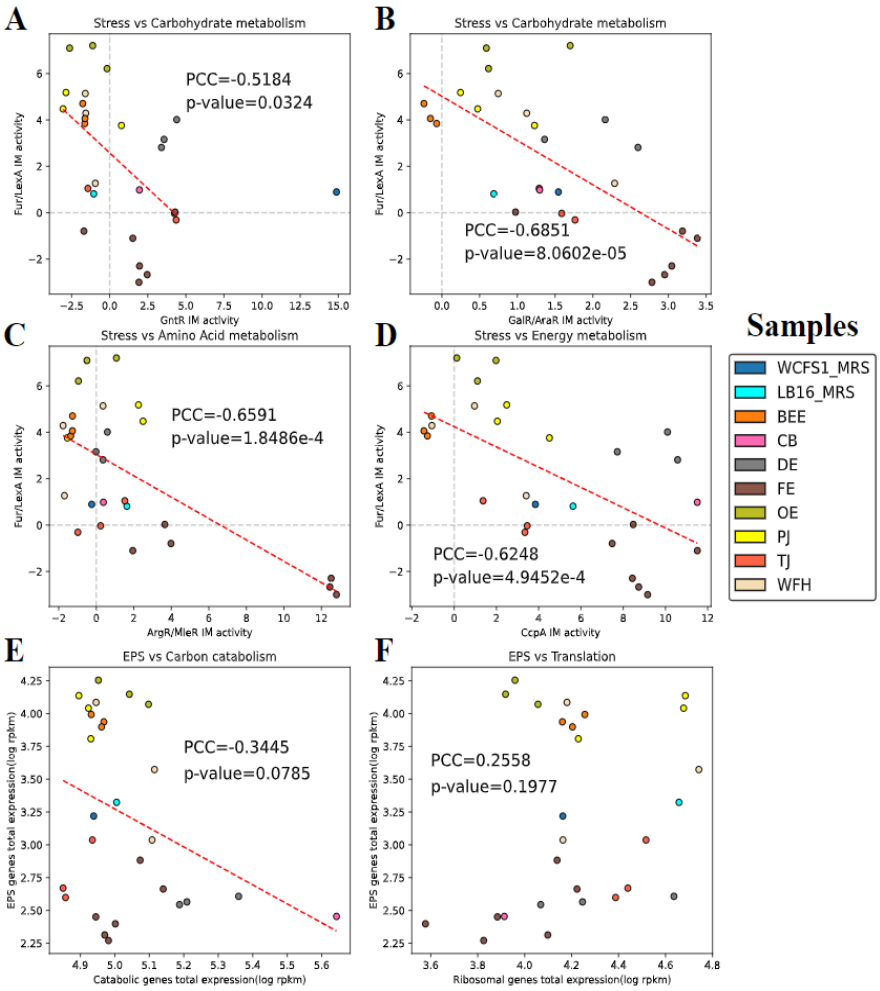

Figure 5.8. The relationships between secondary and primary metabolisms for *L. plantarum* cultivated in different growth conditions. (A) Fur/LexA IM activity versus GntR IM activity. (B) Fur/LexA IM activity versus GalR/AraR IM activity. (C) Fur/LexA IM activity versus ArgR/MleR IM activity. (D) Fur/LexA IM activity versus CcpA IM activity. (E) The total expression levels (log RPKM) of central catabolic genes and EPS biosynthetic genes (genes in Fur/LexA IM). (F) The total expression levels (log RPKM) of ribosomal genes (genes in Translation IM (IM-19)) and EPS biosynthetic genes. WCFS1_MRS: *L. plantarum* WCFS1 in MRS broth; LB16_MRS: *L. plantarum* LB16 in MRS broth; BEE: Bee extract; CB: Cheese broth; DE: Drosophila sp. extract; FE: Fecal extract; OE: Olive extract; PJ: Pineapple juice; TJ: Tomato juice; WFH: Wheat flour hydrolyzate.

5.4 Conclusions and discussion

ICA decomposition of *Lactiplantibacillus plantarum* transcriptomes allowed us to identify 45 nonempty IMs, 53.3% of which were annotated with associated TFs via either regulon enrichment analysis (13 IMs) or motif comparison (11 IMs). Annotated IMs revealed several regulatory interactions that have not been reported by known regulons of *L. plantarum*, e.g., malS (lp_0179) and msmX (lp_0180) captured by MalR+MdxR IM (**section 5.3.2**), which contributed to the reconstruction of a more complete TRN. Furthermore, the Activity matrix (A matrix) output by ICA decomposition showed the change of regulatory activities of TFs in response to different growth conditions (e.g., acid stress, carbon source switch), leading to the identification and characterization of relevant active IMs (**section 5.3.3**). Lastly, the analysis of relationships between IM activities unveiled a trade-off between secondary metabolism (acid stress induced EPS biosynthesis) and primary metabolism in *L. plantarum* (**section 5.3.4**), which might shed light on evolutionarily beneficial growth strategies.

Though IMs derived in this study provided regulatory information for the reconstruction of the TRN of *L. plantarum*, the performance of ICA decomposition was limited by the size of the expression matrix, compared to other ICA-based studies of bacterial transcriptomes (e.g., ICA of *Corynebacterium glutamicum* collected 263 samples from 29 independent projects [39]). Compared to well-studied organisms such as *Escherichia coli*, the amount of existing transcriptomic data of *Lactiplantibacillus plantarum* on NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) [40] is much smaller. Also, due to the lack of operon annotation in *Lactiplantibacillus plantarum*'s genome, motif search for TF binding sites in this study used estimated promoter regions, which lowered the accuracy and might explain why some IMs were uncharacterized. It is also worth noting that the novel regulatory interactions shown by ICA are just "predicted" instead of "confirmed". To obtain a more valid conclusion, CHiP-seq experiments are needed to confirm those findings in future studies.

With regards to the relationship between secondary and primary metabolism, theoretical models such as Grime's competitor-stress-ruderal triangle [41,42], Synthetic Chemostat Model [43] and regulatory proteome allocation model [44] all adopted a resource allocation framework to capture the balance between two branches of cellular metabolism. Through the correlations between the activities of identified IMs, this study provided evidence to the theoretical models for secondary metabolism proposed in previous studies by showing the growth strategy of *L. plantarum* that adjusts regulatory activities for different metabolic pathways to react to external stress signals (**section 5.3.4**). However, the curated data in this study could not support a significant trade-off relationship between gene expression levels of primary and secondary metabolism. More transcriptomic and proteomic profiling for *L. plantarum* under different growth conditions are needed to quantitatively study the balance between stress and cellular growth.

To conclude, this study provided the *in-silico* TRN reconstruction for *L. plantarum* in a top-down manner and unveiled its growth strategy to balance primary and secondary metabolism with IM activities, in spite of the limitations discussed above. With the growing amount of gene expression data of *L. plantarum* as expected, the quality of IMs derived by ICA will be improved, thus enabling researchers to acquire a better understanding of the underlying rationale of its cellular activities.

References

1. Van Hijum SAFT, Medema MH, Kuipers OP. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. Microbiol Mol Biol Rev. 2009;73: 481–509, Table of Contents.

2. Ishihama A. Prokaryotic genome regulation: a revolutionary paradigm. Proc Jpn Acad Ser B Phys Biol Sci. 2012;88: 485–508.

3. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev

Genet. 2009;10: 669–680.

4. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol. 2011;12: 918–922.

5. Lemoine GG, Scott-Boyer M-P, Ambroise B, Périn O, Droit A. GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. BMC Bioinformatics. 2021;22: 267.

6. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5. doi:10.1371/journal.pone.0012776

7. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun. 2019;10: 5536.

8. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018;9: 1090.

9. Shin J, Rychel K, Palsson BO. Systems biology of competency in Vibrio natriegens is revealed by applying novel data analytics to the transcriptome. Cell Rep. 2023;42: 112619.

10. Hirose Y, Poudel S, Sastry AV, Rychel K, Lamoureux CR, Szubin R, et al. Elucidation of independently modulated genes in Streptococcus pyogenes reveals carbon sources that control its expression of hemolytic toxins. mSystems. 2023;8: e0024723.

11. Seddik HA, Bendali F, Gancel F, Fliss I, Spano G, Drider D. Lactobacillus plantarum and Its Probiotic and Food Potentialities. Probiotics Antimicrob Proteins. 2017;9: 111–122.

12. De Angelis M, de Candia S, Calasso MP, Faccia M, Guinee TP, Simonetti MC, et al. Selection and use of autochthonous multiple strain cultures for the manufacture of high-moisture traditional Mozzarella cheese. Int J Food Microbiol. 2008;125: 123–132.

13. Silva LA, Lopes Neto JHP, Cardarelli HR. Exopolysaccharides produced by Lactobacillus plantarum: technological properties, biological activity, and potential application in the food industry. Ann Microbiol. 2019;69: 321–328.

14. Arasu MV, Al-Dhabi NA, Ilavenil S, Choi KC, Srigopalram S. In vitro importance of probiotic Lactobacillus plantarum related to medical field. Saudi J Biol Sci. 2016;23: S6–S10.

15. Jung S, Lee J-H. Characterization of transcriptional response of Lactobacillus plantarum under acidic conditions provides insight into bacterial adaptation in fermentative environments. Sci Rep. 2020;10: 19203.

16. Wels M, Overmars L, Francke C, Kleerebezem M, Siezen RJ. Reconstruction of the regulatory network of Lactobacillus plantarum WCFS1 on basis of correlated gene expression and conserved regulatory motifs. Microb Biotechnol. 2011;4: 333–344.

17. Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, et al. RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics. 2013;14: 1–12.

18. Spangler JR, Dean SN, Leary DH, Walper SA. Response of Lactobacillus plantarum WCFS1 to the Gram-Negative Pathogen-Associated Quorum Sensing Molecule N-3-Oxododecanoyl Homoserine Lactone. Front Microbiol. 2019;10: 715.

19. Filannino P, De Angelis M, Di Cagno R, Gozzi G, Riciputi Y, Gobbetti M. How Lactobacillus plantarum shapes its transcriptome in response to contrasting habitats.

Environ Microbiol. 2018;20: 3700–3716.

20. Özcan E, Rozycki MR, Sela DA. Cranberry Proanthocyanidins and Dietary Oligosaccharides Synergistically Modulate Lactobacillus plantarum Physiology. Microorganisms. 2021;9. doi:10.3390/microorganisms9030656

21. Siezen RJ, Francke C, Renckens B, Boekhorst J, Wels M, Kleerebezem M, et al. Complete resequencing and reannotation of the Lactobacillus plantarum WCFS1 genome. J Bacteriol. 2012;194: 195–196.

22. Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. Nucleic Acids Res. 2021;49: D112–D120.

23. Rychel K, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. Nat Commun. 2020;11: 6338.

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. arXiv [cs.LG]. 2012. pp. 2825–2830. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

25. McConn JL, Lamoureux CR, Poudel S, Palsson BO, Sastry AV. Optimal dimensionality selection for independent component analysis of transcriptomic data. BMC Bioinformatics. 2021;22: 584.

26. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res. 2010;38: e130.

27. Reese MG. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem. 2001;26: 51–56.

28. Bailey TL. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. Department of Computer Science and Engineering, University of California, San Diego; 1994.

29. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8: R24.

30. Cipriano MJ, Novichkov PN, Kazakov AE, Rodionov DA, Arkin AP, Gelfand MS, et al. RegTransBase--a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. BMC Genomics. 2013;14: 213.

31. Magnani D, Barré O, Gerber SD, Solioz M. Characterization of the CopR regulon of Lactococcus lactis IL1403. J Bacteriol. 2008;190: 536–545.

32. Ravcheev DA, Best AA, Sernova NV, Kazanov MD, Novichkov PS, Rodionov DA. Genomic reconstruction of transcriptional regulatory networks in lactic acid bacteria. BMC Genomics. 2013;14: 94.

33. Muscariello L, Vastano V, Siciliano RA, Sacco M, Marasco R. Expression of the Lactobacillus plantarum malE gene is regulated by CcpA and a MalR-like protein. J Microbiol. 2011;49: 950–955.

34. Welman AD, Maddox IS. Exopolysaccharides from lactic acid bacteria: perspectives and challenges. Trends Biotechnol. 2003;21: 269–274.

35. Bagg A, Neilands JB. Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli. Biochemistry. 1987;26: 5471–5477.

36. Deutscher J. The mechanisms of carbon catabolite repression in bacteria. Curr Opin Microbiol. 2008;11: 87–93.

37. Whitfield C, Wear SS, Sande C. Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides. Annu Rev Microbiol. 2020;74: 521–543.

38. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22.

39. Zhao J, Sun X, Mao Z, Zheng Y, Geng Z, Zhang Y, et al. Independent component analysis of Corynebacterium glutamicum transcriptomes reveals its transcriptional regulatory network. Microbiol Res. 2023;276: 127485.

40. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

41. Grime JP. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory. Am Nat. 1977;111: 1169–1194.

42. Bruggeman FJ, Teusink B, Steuer R. Trade-offs between the instantaneous growth rate and long-term fitness: Consequences for microbial physiology and predictive computational models. Bioessays. 2023;45: e2300015.

43. Panikov NS. Genome-Scale Reconstruction of Microbial Dynamic Phenotype: Successes and Challenges. Microorganisms. 2021;9. doi:10.3390/microorganisms9112352

44. Qiu S, Yang A, Zeng H. Flux balance analysis-based metabolic modeling of microbial secondary metabolism: Current status and outlook. PLoS Comput Biol. 2023;19: e1011391.

Chapter 6 Proteome trade-off between primary and secondary metabolism shapes acid stress induced bacterial exopolysaccharide production

Under review. Preprint available as: Qiu S, Yang A, Yang X, Li W, Zeng H, Wang Y. Proteome trade-off between primary and secondary metabolism shapes acid stress induced bacterial exopolysaccharide production. bioRxiv. 2024. p. 2024.04.19.590233. doi:10.1101/2024.04.19.590233

Sizhe Qiu, as the first author, conceptualized the workflow, conducted the multi-omics analysis, formulated the modeling framework, and produced the first draft of the paper. The experimental work on whole genome sequencing, growth measurement, proteomics, and metabolomics was conducted by Xinyu Yang and Hong Zeng. Other listed authors contributed to the revision of the paper.

Summary

The exopolysaccharide (EPS) produced by *Lactiplantibacillus plantarum* is a high-value bioproduct in food and health industries, and its biosynthesis has been found as a secondary metabolic pathway to mediate acid stress. To quantitatively investigate acid stress response in *L. plantarum* and model EPS production, this study measured metabolomics, proteomics and growth data for *L. plantarum* HMX2 cultured at 4 different pH values. The growth and metabolomics data showed that under acid stress, the EPS production flux was evidently enhanced while the glycolysis and cellular growth were inhibited. The following proteomic analysis found that EPS biosynthetic proteins were significantly up-regulated under acid stress

and pinpointed Fur as the most probable transcriptional factor controlling EPS biosynthesis in *L. plantarum*. Furthermore, we identified a proteome trade-off between primary metabolism and EPS biosynthesis, which were then mechanistically depicted by a regulatory proteome constrained flux balance analysis (RPCFBA) model. As the first metabolic model that can simulate secondary metabolism, the RPCFBA model demonstrated good accuracy in predicting growth rates and EPS production fluxes of *L. plantarum* HMX2, validated by experimental data. The *in-silico* perturbation on carbon sources further showed the potential of applying the presented modeling framework to the design and control of microbial secondary metabolism.

Reporting the fourth research project in this PhD study, this chapter presents a computational model for stress-induced secondary metabolism (i.e., EPS biosynthesis in *L. plantarum*), based on the results in Chapter 5 and multi-omics analysis of *L. plantarum* HMX2 under different levels of acidity.

6.1 Introduction

Lactiplantibacillus plantarum (LP) is a gram-positive lactic acid bacterium found in diverse ecological niches [1]. LP has been widely used in food and health industries. For instance, it is the major bacterium involved in the fermentation of mozzarella cheese [2]; it is considered as a safe probiotic that can suppress pathogenic microorganisms [3] as well as has immunomodulatory properties [4]. Many important properties of LP are related to its secreted exopolysaccharide (EPS), which has been extensively studied for its various probiotic effects [5], anticancer properties [6] and natural ability to improve rheological and sensory properties of fermented foods [7]. In contrast to primary metabolites, the LP derived EPS is a typical secondary metabolite [8], whose biosynthesis directly competes with growth coupled central carbon metabolism for the carbon source and energy. However, the mechanism behind the biosynthesis of EPS has not been well studied compared with its structural and functional

properties, thus has long remained murky. The investigation of the relationship between primary metabolism and EPS biosynthesis can provide critical insights and consummate the knowledge of bacterial growth laws [9].

In order to achieve precise control of LP derived EPS fermentation, researchers have to elucidate the driving force behind its biosynthesis. Previous studies on the physiological function of lactic acid bacteria derived EPS have indicated that the EPS layer can protect the bacterial cell from acid, osmotic and oxidative stresses [10]. Furthermore, the significant increase of EPS yield in LP VAL6 strain at low pH found by Nguyen et al. 2021 [11] suggests that the EPS biosynthesis is a response to acid stress. In the transcriptomic profiling of LP at pH 5.0, 5.5, and 6.2, some genes involved in EPS biosynthesis (e.g., tagE1, tagE5) were found to be up-regulated for pH 5.0 vs pH 5.5 and pH 6.2 [12]. However, the up-regulation of EPS biosynthetic genes was not observed for LP CAUH2 strain under oxidative stress [13]. Subsequently, in our previous work, the analysis of independently modulated gene sets (iModulons) with transcriptomic samples of various conditions (e.g., acid stress, different carbon sources), for the first time, found an interesting trade-off relationship between the regulatory activities of primary metabolism and EPS biosynthesis [14]. Nevertheless, such trade-off has not been validated by the proteomic profiling of LP under acid stress yet [15]. In short, various evidence has suggested that the LP derived EPS is a secondary metabolite produced to mediate acid stress, which is beneficial to cell survival as well as renders a long-term population advantage, but the relationship and the regulatory rule between primary metabolism and EPS biosynthesis still remain unclear.

In recent years, computational methods have been increasingly used to assist the design and optimization of the EPS production in lactic acid bacteria. Response surface modeling (RSM), a statistical method, has been widely used to solve the optimal growth medium composition and environmental condition for EPS production [16–18]. Given its broad application, RSM generally treats bacterial metabolism as a “black box”, which makes its results

case-specific, therefore limiting its ability for in-depth analysis. Comparatively, flux balance analysis (FBA) with genome-scale metabolic models (GSMMs) is a commonly used “white-box” method to compute metabolic fluxes [19]. Existing GSMMs of lactic acid bacteria have already included lumped EPS synthetic reactions in their metabolic networks [20,21]. Nonetheless, several limitations of conventional FBA on secondary metabolism makes it still challenging to model EPS biosynthesis and cellular growth simultaneously [22].

To gain an in-depth and quantitative understanding of how acid stress influences primary metabolism and EPS biosynthesis in LP, this study adopted a multi-omics analysis approach to characterize the phenotypic difference of LP HMX2 [23] cultured at different pH values, representing different degrees of acid stress, and constructed a regulatory proteome constrained FBA (RPCFBA) model that is capable of simulating cellular growth, central carbon metabolism and acid stress induced EPS biosynthesis simultaneously. More specifically, measured bacterial growth and metabolomics data at 4 different fixed pH values (pH 4.5, 5.0, 5.5 and 6.5) were compared to show how the acid stress affected primary metabolism and EPS production differently. Then, a quantitative proteomic analysis elucidated the acid stress induced regulation of protein expressions by identifying and analyzing significantly up- and down-regulated proteins. Furthermore, the RPCFBA model was constructed and validated with experimental data. Finally, the model demonstrated its practical usefulness in the design and control of microbial secondary metabolism with simulations of growth rates and EPS production under *in-silico* perturbations on carbon sources.

6.2 Materials and Methods

6.2.1 Strains, media, and culture conditions

Lactiplantibacillus plantarum HMX2 (LP HMX2), an EPS producing strain isolated from Chinese Northeast Sauerkraut [23], was used to investigate the effect of acid stress on the secondary metabolism of lactic acid bacteria. For activation, 3% inoculum of LP HMX2 was introduced into De Man, Rogosa and Sharpe (MRS) broth and incubated anaerobically at 37°C for 12 hours in a sterile laminar flow hood. For culture experiments, 10ml activated bacterial culture was inoculated into a conical flask with 3L MRS broth, sealed with a membrane. The temperature was maintained at 37°C. The LP HMX2 was cultured at 4 different fixed pH values: 6.5, 5.5, 5.0, and 4.5, and each condition had 3 replicates. The pH of each condition was adjusted to and maintained at the fixed value using titration with sodium hydroxide and sulfuric acid, tracked by the fermentation monitor (iCinac, AMS, France) with an Inlab Smart pro-ISM detection electrode (Mettler Toledo, Switzerland) (**see Appendix D, Figure D.1**). The pH 6.5 was selected as the reference condition in metabolomic and proteomic analysis, as it is the upper bound of optimal growth pH for most LP strains [24]. The culture experiment lasted for 14 hours for each condition and samples of 100ml liquid culture media were taken every 2 hours for further analysis. The OD600 was measured using a spectrophotometer at 600 nm for the quantification of growth kinetics, and converted to dry weight biomass concentrations with a conversion factor of 0.35 $\frac{gDW}{OD600}$ [25]. The concentration of produced EPS was estimated by the phenol-sulfuric acid method [26] using glucose as standard, and subtracting the amount determined at zero time. The measurement of metabolites (e.g., glucose, lactic acid, etc.) was elaborated in **section 6.2.3**.

6.2.2 Whole genome sequencing analysis

The genome sequencing of LP HMX2 was performed using the Hiseq/Novaseq sequencing platform. The genomic DNA was extracted using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany), and then fragmented into a size within 500bp using Bioruptor Sonication Device (Covaris S220). End Prep Enzyme Mix was used for the end repair of fragmented DNA sequences. A HiFi sequencing library for de novo genome assembly was prepared using the SMRTbell library prep kit (PacBio, Menlo Park, CA, USA), according to the manufacturer's protocol. Then, the sequencing was carried out using Illumina HiSeq instrument (Illumina, San Diego, CA, USA) following ABySS v. 12.1 assembly method. After reads with lengths smaller than 500bp were filtered out, HGAP4/Falcon [27] was used to assemble the cleaned reads.

The coding genes on the assembled genome of LP HMX2 were identified using Prodigal [28], and annotated with various databases including Non-Redundant Protein Database (NR) [29], Kyoto Encyclopedia of Genes and Genomes (KEGG) [30], Clusters of Orthologous Genes (COG) [31], Carbohydrate-Active enZYmes (CAZy) [32] and Transporter Classification Database (TCDB) [33]. The locus tags of coding genes in LP HMX2 can be found in https://github.com/SizheQiu/LbPtEPS/tree/main/data/Genome_HMX2. The biosynthetic gene cluster (BGC) of EPS was predicted using antiSMASH bacterial version [34].

6.2.3 Quantification of intra- and extra-cellular metabolomics

The concentrations of intra- and extra-cellular metabolites were measured to quantify metabolic exchange fluxes (uptake and secretion) and intracellular metabolic status. For extracellular metabolites, samples were filtered through a 0.22 μm membrane and the supernatant was collected. For intracellular metabolites, samples were centrifuged at 9000 rpm for 10 minutes. Then, the pellet was collected and washed with the PBS buffer. The extracellular

metabolomic data can be found in https://github.com/SizheQiu/LbPtEPS/blob/main/data/Exp_data/Metabolomics_mM.csv, and the intracellular metabolic data can be found in https://github.com/SizheQiu/LbPtEPS/blob/main/data/Exp_data/IntraMetabolomics.csv.

The concentrations of amino acids, glucose, lactic acid and glycolytic intermediates (except glucose-6 phosphate and fructose 6-phosphate) were measured using LC/MS: UPLC (LC-30AD) equipped with a MS-8050 triple quadrupole mass spectrometer. To prepare the samples, 600 μL of acetonitrile was added to 200 μL of test and QC samples separately. Then, each sample was vortexed and centrifuged with 12000 RCF at 4℃ for 10 min, and the supernatant of each sample was diluted at 4-fold, 20-fold and 100-fold with water. 20 μL of internal standard solution was added to each sample. Next, each sample was vortexed with 1000 RCF at 25 ℃ for 10 min, and centrifuged with 4680 RCF at 4℃ for 10 min. The columns used in LC were ACQUITY UPLC HSS T3 and Discovery HS F5 HPLC, and the temperature was 45℃. The injection volume was 1uL, and the flow rate of the mobile phase (A=0.1% formic acid/water; B=0.1% formic acid/acetonitrile) was 0.3 mL/min. The gradient elution procedure was as follows: 0 min, 100% A; 3 min, 100% A; 7 min, 40% A, 60% B; 8 min, 5% A, 95% B; 10 min, 5% A, 95% B; 10.1 min, 100% A. The acquisition mode of MS was MRM. The interface voltage was 4.0 kV. The temperatures of the interface, desolvation line and heater block were 300 ℃, 250 ℃ and 400 ℃, respectively. The flow rates of automatic gas, drying gas and heating gas were 3.0 L/min, 10.0 L/min and 10.0 L/min.

For glucose-6 phosphate and fructose 6-phosphate, the sample preparation and the LC/MS equipment were the same. The column of LC was ACQUITY UPLC BEH Amide with a temperature at 45℃. The injection volume was 1uL, and the flow rate of the mobile phase (A= 20 mM ammonium acetate, 1.2% ammonia, 95% water, 5% acetonitrile; B= acetonitrile) was 0.3 mL/min. The gradient elution procedure was as follows: 0 min, 25% A, 75% B; 1 min, 40% A, 60% B; 10 min, 40% A, 60% B; 13 min, 25% A, 75% B; 13.1 min, 25% A, 75% B. The interface

voltage in MS was set at 4.0 kV, and the other settings in MS were the same as those in LC/MS for other metabolites, as described above.

The concentrations of short-chain fatty acids and acetic acids were measured using GC/MS: GC-2030 equipped with a TQ 8040NX mass spectrometer. For the sample preparation of fatty acids, 20 μL of 5% aqueous phosphoric acid solution was added to 50 μL of the sample, and each sample was vortexed for 5 min. Then, 100 μL of methyl tert-butyl ether was added, and each sample was vortexed for 5 min and centrifuged with 12000 rpm/min for 5 min at 4 ℃. The 70 μL of supernatant was transferred to the injection vial. For the sample preparation of acetic acid, 20 μL of 5% aqueous phosphoric acid solution was added to 50 μL of the sample, and each sample was vortexed for 5 min. Then, 100 μL of methyl tert-butyl ether was added, and each sample was vortexed for 5 min and centrifuged with 12000 rpm/min for 5 min at 4 ℃. The 5 μL of the supernatant was added into 495 μL of methyl tert-butyl ether, and transferred to the injection vial. For the settings in GC, the column was DB-FFAP (30 m*0.32 mm*0.25 μm), the injection volume was 1 μL, the flow rate was 1 mL/min, the loading gas was helium, and the temperature was 230 ℃. The gradient elution procedure was as follows: 0 ℃/min, 60 ℃, 2 min holding time; 2 ℃/min, 100 ℃, 0 min holding time; 5 ℃/min, 110 ℃, 0 min holding time; 10 ℃/min, 160 ℃, 0 min holding time; 20 ℃/min, 240 ℃, 2 min holding time. For the settings in MS, the acquisition mode was Q3 SIM, the source temperature was 230 ℃, the interface temperature was 200 ℃, and the detector voltage was 1.04 kV.

6.2.4 Structural analysis of LP-HMX2 derived exopolysaccharide

The isolation and purification of LP HMX2 derived exopolysaccharide (LP-HMX2-EPS) followed the experimental procedure in Yang et al., 2018 [35]. Then, the monosaccharide and glycosyl linkage compositions of the purified LP-HMX2-EPS were quantitatively determined (**see**

Appendix D, Table D.1), and were used to build pseudo-reactions of EPS biosynthesis in the modified GSMM (**see Appendix D, Supplementary method D.1.1**).

The mass fraction of each monosaccharide in LP-HMX2-EPS was measured using LC/MS: ThermoU3000 HPLC system. The 5 mg purified EPS sample was added to 1 ml 2 M trifluoroacetic (TFA) acid solution, and heated at 121 °C for 2 hours. The sample was washed by 3 ml methanol and dried with nitrogen gas for 3 times. The dried sample was dissolved in 5ml sterile deionized water. 0.2 ml 0.5 M NaOH and 0.5 ml 0.5 M PMP-methanol were added to 0.2 ml sample to react at 70 °C for 1 hour to form sugar-PMP derivatives. 0.2 ml 0.5 M HCl was added to neutralize NaOH. Vortex-assisted extraction with chloroform was used to remove excess PMP. The sample volume was adjusted to 1 ml with sterile deionized water before detection. For the settings in ThermoU3000 HPLC system, the column was ZORBAX EclipseXDB-C18, the flow rate of the mobile phase (acetonitrile:12 g/L monobasic potassium phosphate/2 M NaOH=17:83) was 0.8 mL/min for isocratic elution, the column temperature was 30 °C, the detection wavelength was 250 nm, and the injection volume was 10 μl. The mass fractions of mannose, glucosamine, rhamnose, glucuronic acid, galacturonic acid, galactosamine, glucose, galactose, xylose, arabinose and fucose were quantified using external standard calibration.

Methylation analysis and GC/MS (Agilent 7890A GC system equipped with Agilent 5977B quadrupole mass spectrometer) were used to determine the glycosyl linkage composition through the quantification of different monosaccharide residues (e.g., 1,2-Manp). The 10 mg EPS sample was dissolved in 1 ml deionized water, and 1 ml 100 mg/ml carbodiimide was added to react for 2 hours. 1ml 2M imidazole and 1 ml 30 mg/ml sodium borodeuteride (NaBD4) were added to react for 3 hours. 100 μL acetic acid was added to terminate the reaction. The sample was filtered and freeze-dried. Then, the methylation of LP-HMX2-EPS followed the experimental procedure in Ciucanu & Kerek, 1984 [36]. Next, the methylated sample was injected into the Agilent 7890A-5977B GC/MS system. The column was

HP-5MS capillary column (30 m × 0.25 mm × 0.25 μm, Agilent J&W Scientific, Folsom, CA, USA), the loading gas was helium, the flow rate was 1.0 mL/min, the injection volume was 1μL, and the temperature of the injection port was 260 °C. The gradient elution procedure was as follows: 0 ℃/min, 50 ℃, 1 min holding time; 50 ℃/min, 130 ℃, 3 ℃/min, 230 ℃, 2 min holding time. The Agilent 5977B quadrupole mass spectrometer was equipped with an electron impact ion source and a MassHunter workstation. The temperature of the injection port and the quadrupole were 230 ℃ and 150 ℃ respectively. The electron energy was 70 eV. The scan mode was full scan with a range of m/z 30 to m/z 600.

6.2.5 Quantitative proteomic analysis

For each condition, samples were taken at the exponential phase (i.e., 6 hr), and SDT (4% SDS, 100 mM Tris-HCl, pH 7.6) buffer was used for sample lysis and protein extraction. 20 μg of protein for each sample was mixed with the 5X loading buffer and boiled for 5 min, and then separated on 4%-20% SDS-PAGE gel (constant voltage 180 V, 45 min). Protein digestion by trypsin was performed following filter-aided sample preparation (FASP) procedure [37]. The digested peptides of each sample were desalted on MCX, concentrated by vacuum centrifugation and reconstituted in 40 μL 0.1% (v/v) formic acid. The peptide mixture of each sample was labeled using TMT reagent and mixed equally. The labeled peptides were then fractionated by High pH Reversed-Phase Peptide Fractionation Kit (Thermo Scientific). The peptide mixture was reconstituted and acidified with 0.1% TFA solution and loaded to the equilibrated, high-pH, reversed-phase fractionation spin column. Peptides were bound to the hydrophobic resin under aqueous conditions and desalted by washing the column with water by low-speed centrifugation. A step gradient of increasing acetonitrile concentrations in a volatile high-pH elution solution was then applied to the columns to elute bound peptides into 10/20/30

different fractions collected by centrifugation. Finally, the collected fractions of peptides were vacuum dried and lyophilized with 12 µL 0.1% FA.

LC-MS/MS analysis was performed on a Q Exactive mass spectrometer (Thermo Scientific) coupled to Easy nLC (Thermo Scientific). The columns were a reverse phase trap column (Thermo Scientific Acclaim PepMap100, 100 µm*2 cm, nanoViper C18) and a C18-reversed phase analytical column (Thermo Scientific Easy Column, 10 cm long, 75 µm inner diameter, 3 µm resin). The flow rate of the mobile phase (A=0.1% formic acid/water; B=84% acetonitrile and 0.1% formic acid) was 300 nL/min. The mass spectrometer was operated in positive ion mode. MS data was acquired using a data-dependent top20 method that dynamically chose the most abundant precursor ions from the survey scan (300–1800 m/z) for HCD fragmentation. Automatic gain control (AGC) target was set to 1e6, and maximum injection time to 50 ms. Dynamic exclusion duration was 30.0 s. Survey scans were acquired at a resolution of 60,000 at m/z 200 and resolution for HCD spectra was set to 15,000 at m/z 200, and isolation width was 1.5 m/z. Normalized collision energy was 30 eV and the underfill ratio was set as 0.1%. The MS instrument was run with peptide recognition mode enabled.

The MS raw data was searched using MASCOT (Matrix Science, London, UK; version 2.2) [38] for identification and quantitation of proteins. MASCOT matched tandem mass spectra to a protein sequence database of *Lactiplantibacillus plantarum* WCFS1 obtained from the Uniprot database [39]. The protein mass ratio was calculated as the median of only unique peptides of the protein. The quantitative proteomic data can be found in https://github.com/SizheQiu/LbPtEPS/blob/main/data/Proteomics/Proteomics_B.xlsx. Next, gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) [30] annotations were performed for identified proteins using Blast2GO [40] and eggNOG-mapper [41]. The protein subcellular localization was predicted using CELLO [42]. Differential expression analysis was performed on proteomics data at different pH values with PyDESeq2 [43].

6.2.6 Regulatory proteome constrained flux balance analysis

The GSMM used in this study was the modified iBT721 [20], and FBA [19] was used to compute metabolic fluxes (**see Appendix D, Supplementary method D.1.1**). The objective function to maximize was the growth rate normalized to 1 gram dry weight (gDW) of biomass, v_{growth} (Eq. 6.1). The basic constraints were mass conservation (Eq. 6.2) and default lower/upper bounds (v_{lb}, v_{ub}) of reaction fluxes (Eq. 6.3). COBRApy [44] was used to perform FBA.

$$Max\ v_{growth}\ (Eq. 6.1)$$

$$S * v\ =\ 0\ (Eq. 6.2)$$

$$v_{lb} \leq v_i \leq v_{ub}\ (Eq. 6.3)$$

Proteome constrained FBA tightens the metabolic flux solution space by integrating proteome constraints of reactions into conventional FBA [45]. In this work, the reaction flux (v_i, $\frac{mmol}{hr * gDW}$) was constrained by the enzyme activity, $a_i(\frac{mmol}{hr * g\ E})$ (Eq. 6.4). Some values of a_i (e.g., Ribose-5-phosphate isomerase) were missing in BRENDA and SABIO-RK, and thus were estimated by DLTKcat [46] (**see Appendix D, Supplementary method D.1.2**). The proteome of LP HMX2 was divided into sectors of inflexible housekeeping (Q), anabolism (A), transportation (T), catabolism (C) and secondary metabolism (U, EPS biosynthesis). The upper bound of the summation of the 4 flexible sector fractions ($\phi_C + \phi_A + \phi_T + \phi_U$) was assumed to be 50% of the total proteome (Eqs. 6.5 & 6.6) [47–49]. ϕ_j represented the mass fraction of the sector j for $j = A, C, T, U$. P_{TOT} was the total mass of the proteome normalized to 1 gDW of biomass ($\frac{g}{gDW}$) and set as 0.299 $\frac{g}{gDW}$ [20] for LP.

$$v_i \leq a_i[E_i]\ \ (Eq. 6.4)$$

$$\phi_Q(50\%) + \phi_C + \phi_A + \phi_T + \phi_U \leq 100\%\ (Eq. 6.5)$$

$$\phi_j * P_{TOT} = \sum_i \frac{v_{i\in j}}{a_{i\in j}}, j = C, A, T, U \ (Eq. 6.6)$$

To account for the acid stress induced proteome trade-off between primary and secondary metabolism, the ratio between ϕ_U and $\phi_C + \phi_A$ was modeled as a sigmoid function of the pH value (Eq. 6.7) [22]. r_0 was the base ratio at pH 6.5, and r_1, k_{pH} and k_1 were derived based on the quantitative proteomic analysis (**section 6.2.5**). In addition, the changes in enzyme activities caused by the decrease of pH were considered with the pH-dependent function F_{pH} (Eqs. 6.8 & 6.9). The F_{pH} of A, C, T sectors were approximated using the sigmoid function, while that of the U sector was approximated using the quadratic function (**see Appendix D, Supplementary method D.1.2**).

$$\frac{\phi_U}{\phi_C + \phi_A} \geq r_0 + \frac{r_1}{1 + e^{k_{pH} * (pH - k_1)}} \ (Eq. 6.7)$$

$$a_i = F_{pH} a_i^0, F_{pH} = \frac{1}{1 + e^{-c_1 * (pH - c_2)}} \ \forall \ i \in C, A, T \ (Eq. 6.8)$$

$$F_{pH} = c_3 * pH^2 + c_4 * pH + c_5 \ \forall \ i \in U \ (Eq. 6.9)$$

6.3 Results

6.3.1 Inference of EPS structure and characterization of EPS biosynthesis in *Lactiplantibacillus plantarum* HMX2

From the assembled genome sequence of LP HMX2 (NCBI RefSeq assembly, GCF_025144505.1), we predicted the biosynthetic gene cluster of EPS (EPS-BGC), from 1017036 bp to 1063626 bp, that encompassed 20 coding genes using antiSMASH [34] (**Figure 6.1A**). The core genes of EPS-BGC were mainly those encoding for glycosyltransferases (GTs) and biosynthetic proteins of activated monosaccharides (e.g., 1_955, UDP-glucose 4-epimerase), which provide precursors for EPS biosynthesis. 1_947, a core gene of EPS-BGC,

was found to be a homolog of Wzx flippase for polysaccharide transportation with 67.7% identity. The additional genes contained 3 genes encoding for EPS biosynthetic proteins, that were 1_940/1_953 homologous to cps2/4B and 1_952 homologous to cps2/4A.

The monosaccharide composition of LP HMX2 derived exopolysaccharide (LP-HMX2-EPS) was determined to find out its basic precursors: 44.7% mannose, 40.9% glucose, 8.6% galactose, 5.6% glucosamine and 0.2% other monosaccharides (**Figure 6.1B**). The high abundance of mannose in LP-HMX2-EPS indicated that the biosynthesis of GDP-mannose using Mannose-6-phosphate isomerase (MAN6PI), Phosphomannomutase (PMANM), and Mannose 1 phosphate guanylyltransferase (MAN1PT) would direct a larger carbon flux from glycolysis than the biosynthesis of other activated monosaccharides. Methylation analysis was conducted to quantitatively determine the glycosyl linkages in LP-HMX2-EPS (**Figure 6.1C**). The dominance of mannosyl linkages (36.1% t-Manp, 15.5% 1,2,6-Manp and 14.9) suggested that the backbone was mainly composed of 1-2 glycosidic bonds between mannoses while glucosyl and galactosyl linkages likely constituted the branches of the EPS.

Both monosaccharide and glycosyl linkage compositions showed that the EPS mainly consisted of mannose and glucose, but the molar fraction of mannose in glycosyl linkage composition was higher than that in monosaccharide composition. This study considered the glycosyl linkage composition detected by methylation analysis to be more accurate, as the measurement of monosaccharide composition was susceptible to the high glucose concentration in culture media. Subsequently, a pseudo EPS repeating unit was defined (mannose:glucose:galactose = 14:6:1) together with pseudo glycosylation reactions in the GSMM (**Figure 6.1D**). In this study, we assumed that the composition of EPS was invariant under different conditions.

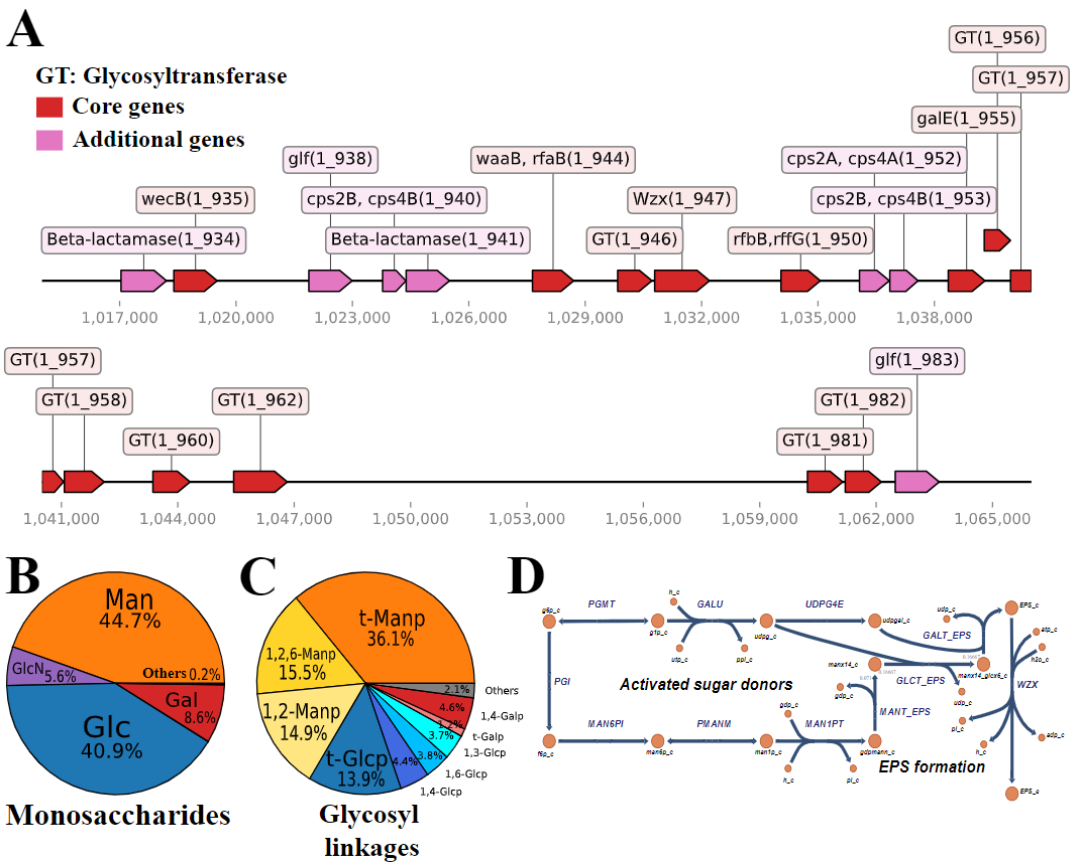

Figure 6.1. The genome-scale biosynthetic pathway of EPS in LP HMX2. (A) The biosynthetic gene cluster of LP HMX2 EPS. GT: glycosyltransferase, Red: core genes, Pink: additional genes. (B) The molar fractions of monosaccharides in LP HMX2 EPS. (C) The glycosyl linkage compositions of LP HMX2 EPS. (D) The metabolic pathway of EPS biosynthesis. Reaction and enzyme information can be found in **Appendix D, Table D.2**. MANT_EPS: mannosyltransferase, GLCT_EPS: glucosyltransferase, GALT_EPS: galactosyltransferase, WZX: Wzx flippase.

6.3.2 The influence of acid stress on cellular metabolism

As the pH decreased from 6.5 to 4.5, the growth coupled primary metabolism and EPS biosynthesis in LP HMX2 showed distinctive responses. The LP HMX2 at pH 6.5 reached exponential and stationary phases faster than those at pH 4.5, 5.0 and 5.5 (**Figure 6.2A**),

suggesting that LP HMX2 was a neutrophile. In contrast, the accumulations of EPS at pH 4.5, 5.0 and 5.5 were faster than that at pH 6.5 in the early growth stage (before 4 hr), though the inhibition effects of low pH and undissociated lactic acid [50,51] on enzyme activities later hampered the EPS accumulation at pH 4.5 and 5.0 (**Figure 6.2B**). The computed average growth rates and EPS production fluxes, normalized to unit biomass, revealed that the highest growth rate at pH 6.5 was approximately twice that at pH 4.5 (**Figure 6.2C**), while the highest EPS production flux at pH 5.0 exceeded twice that at pH 6.5 (**Figure 6.2D**).

The higher consumption rate of glucose and production rate of lactic acid at pH 6.5 than those under acidic conditions further confirmed that the overall turnover rate of primary metabolism was highest at pH 6.5 (**Figure 6.2EF**). Apart from indicating the kinetics of primary metabolism, extracellular metabolomics revealed other interesting physiological information of LP HMX2. The decrease of formic acid concentrations in all conditions during fermentation showed that LP HMX2 in MRS media would consume formic acid instead of producing it (**Figure 6.2G**), possibly for the biosynthesis of purine and pyrimidine [52]. The consumption rate of formic acid also appeared to be inhibited by the low pH. The relatively stable concentrations of acetic acid in all conditions suggested that LP HMX2 is homofermentative (**Figure 6.2H**), like most strains of LP [53].

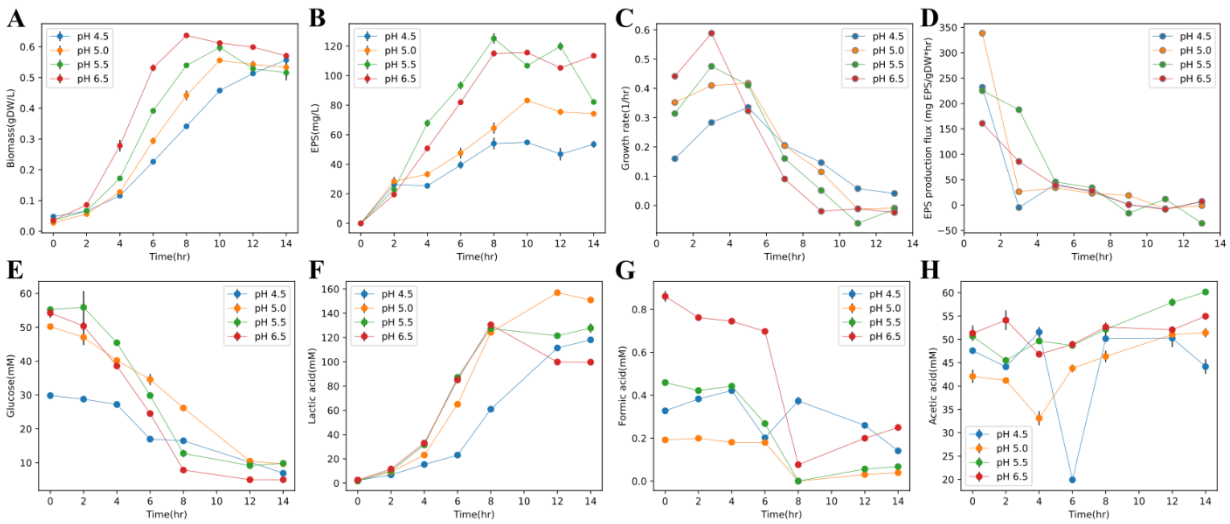

Figure 6.2. Quantification of primary metabolism and EPS production at fixed pH values of 4.5, 5.0, 5.5, 6.5. (A) The concentrations of cellular biomass converted from OD600. (B) The concentrations of EPS. (C) The average growth rates. (D) The average EPS production fluxes normalized to 1gDW cellular biomass. (E) The extracellular concentrations of glucose. (F) The extracellular concentrations of lactic acid. (G) The extracellular concentrations of formic acid. (H) The extracellular concentrations of acetic acid. The number of replicates was three. Error bars represent standard deviations of replicates (error bars of some data points are invisible due to small differences in replicates).

The quantification of intracellular metabolomics can provide a direct insight into the metabolic status of LP HMX2 at different pH values. The relatively high concentrations of glycolytic intermediates (i.e., 3-phosphoglyceric acid, phosphoenolpyruvate, pyruvic acid, lactic acid and acetic acid) at pH 5.5, 5.0 and 4.5 indicated the slowdown of central carbon catabolism, caused by the inhibition of glycolytic enzyme activities by the low pH (**Figure 6.3A-E**). When LP HMX2 entered the stationary phase (12 hr), the high extracellular concentration of undissociated lactic acid (**Figure 6.2F**) and low pH exhibited a combined inhibitory effect on central carbon catabolism, reflected by the increased intracellular concentrations of glycolytic intermediates. In addition, the relatively high concentrations of formic acid at pH 5.0 and pH 4.5 implied that the biosynthesis of purine and pyrimidine, which consumes formic acid, was inhibited (**Figure 6.3F**).

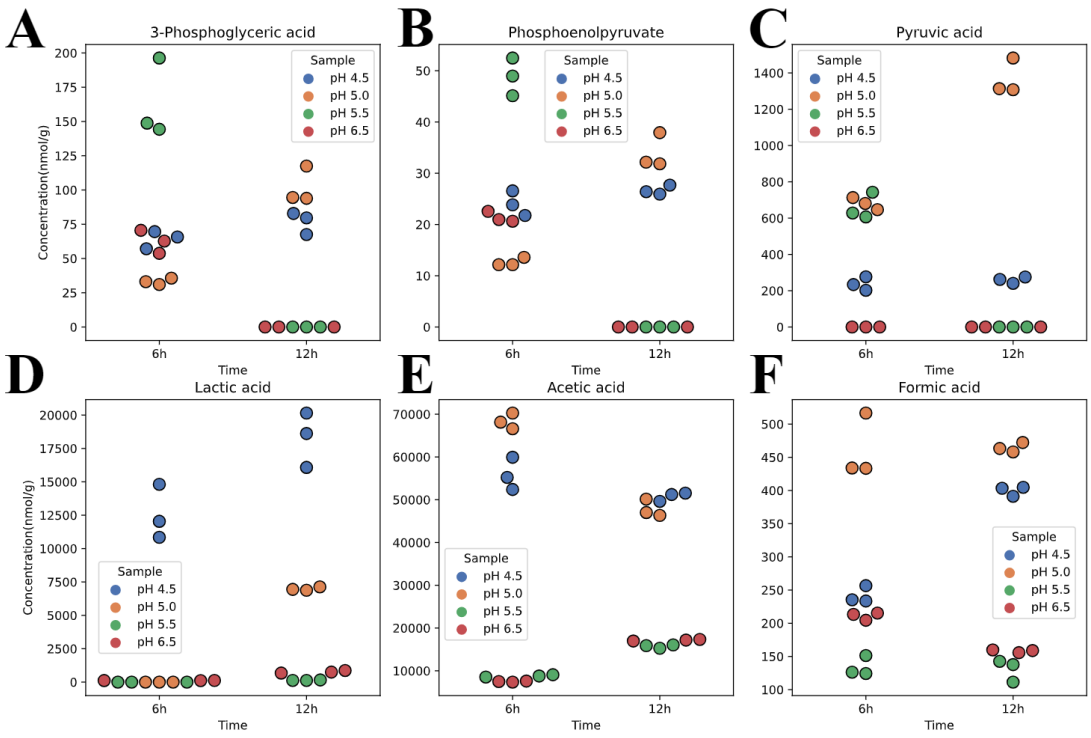

Figure 6.3. Quantification of intracellular metabolomics at different pH values. (A) The intracellular concentrations of 3-phosphoglyceric acid at 6hr and 12hr. (B) The intracellular concentrations of phosphoenolpyruvate at 6hr and 12hr. (C) The intracellular concentrations of pyruvic acid at 6hr and 12hr. (D) The intracellular concentrations of lactic acid at 6hr and 12hr. (E) The intracellular concentrations of acetic acid at 6hr and 12hr. (F) The intracellular concentrations of formic acid at 6hr and 12hr. The number of replicates was three.

6.3.3 Differential protein expression under acidic conditions

The protein expressions of different pH conditions had observable distinguishing distributions in the principal component analysis (PCA) with high explained variance (PC1: 60.24%, PC2: 17.37%) (**Figure 6.4A**). The PCA also identified outlier samples at pH 4.5 and pH 5.0, which were replaced with the averages of other samples from the same conditions. With pH 6.5 as the reference condition, differential expression analysis was performed to identify significantly up- and down-regulated proteins (p-value<0.05, absolute log2 fold

change(|LFC|)>0.5) (**Figure 6.4B**). In comparison with the pH 5.5 condition, the amounts of significantly differentially expressed proteins in pH 4.5 and pH 5.0 conditions were much larger. Notably, some up-regulated proteins had extremely large LFCs, whose expressions were presumably induced by acid stress. Regarding proteins encoded by the EPS-BGC (**section 6.3.1**), significantly up-regulated proteins were only found in pH 4.5 and pH 5.0 conditions (**Figure 6.4C**), which were mainly GTs (e.g., tagE4 encoded by 1_944, lp_1233 encoded by 1_956) and EPS biosynthetic proteins (e.g., cps4B encoded by 1_953).

Samples at pH 4.5 and 5.0 were considered as under acid stress, because the significantly differentially expressed proteins in those two conditions had much higher consistency with each other than with those at pH 5.5 (**Figure 6.4D**). The proteins that were significantly up- or down-regulated at both pH 4.5 and 5.0 were extracted to perform the enrichment analysis for functional pathways, subcellular locations, and regulons. The up-regulated proteins enriched in ATP-binding cassette (ABC) transporters, carbohydrate metabolism, and stress response, while the down-regulated proteins enriched in purine metabolism, ABC transporters, and biosynthesis of cofactors (**Figure 6.4E**), which were highly consistent with previous research on acid stress induced gene expression regulation of LP [12,14,15]. For subcellular locations, most up-regulated proteins (e.g., ABC transporters, stress response proteins, EPS biosynthetic proteins) were membrane proteins, while most down-regulated proteins (e.g., enzymes in purine and pyrimidine metabolism) were located in the cytoplasm (**Figure 6.4F**).

With regulons of LP obtained from RegPrecise [54], both up- and down-regulated proteins were found to be enriched in CcpA regulon (**Figure 6.4G**). Five stress response proteins encoded by genes in LexA regulon were significantly up-regulated. Most proteins encoded by genes in PyrR and PurR regulons were significantly down-regulated. Among 49 identified transcriptional factors (TFs) of LP, 8 and 13 TFs were significantly up- and down-regulated under acid stress, respectively (**Figure 6.4H**). The significant down-regulation of PyrR

and PurR was in agreement with the enrichment analysis of functional pathways and regulons. For TFs controlling nutrient uptake, significantly down-regulated GlnR and FruR are repressors of glutamine ABC transporter and fructose-specific PTS, respectively, while up-regulated MtlR and MdxR are activators of mannitol-specific PTS and maltose/maltodextrin ABC transporter, respectively. Regarding the regulation of EPS biosynthesis, Fur was found to be significantly down-regulated under acid stress, in agreement with previous studies suggesting that Fur directly or indirectly represses EPS biosynthesis [55–57] and mediates acid tolerance [58,59]. Besides, the systematic transcriptomic analysis of LP also indicated that Fur had significant correlations with acid stress and EPS biosynthetic genes [14].

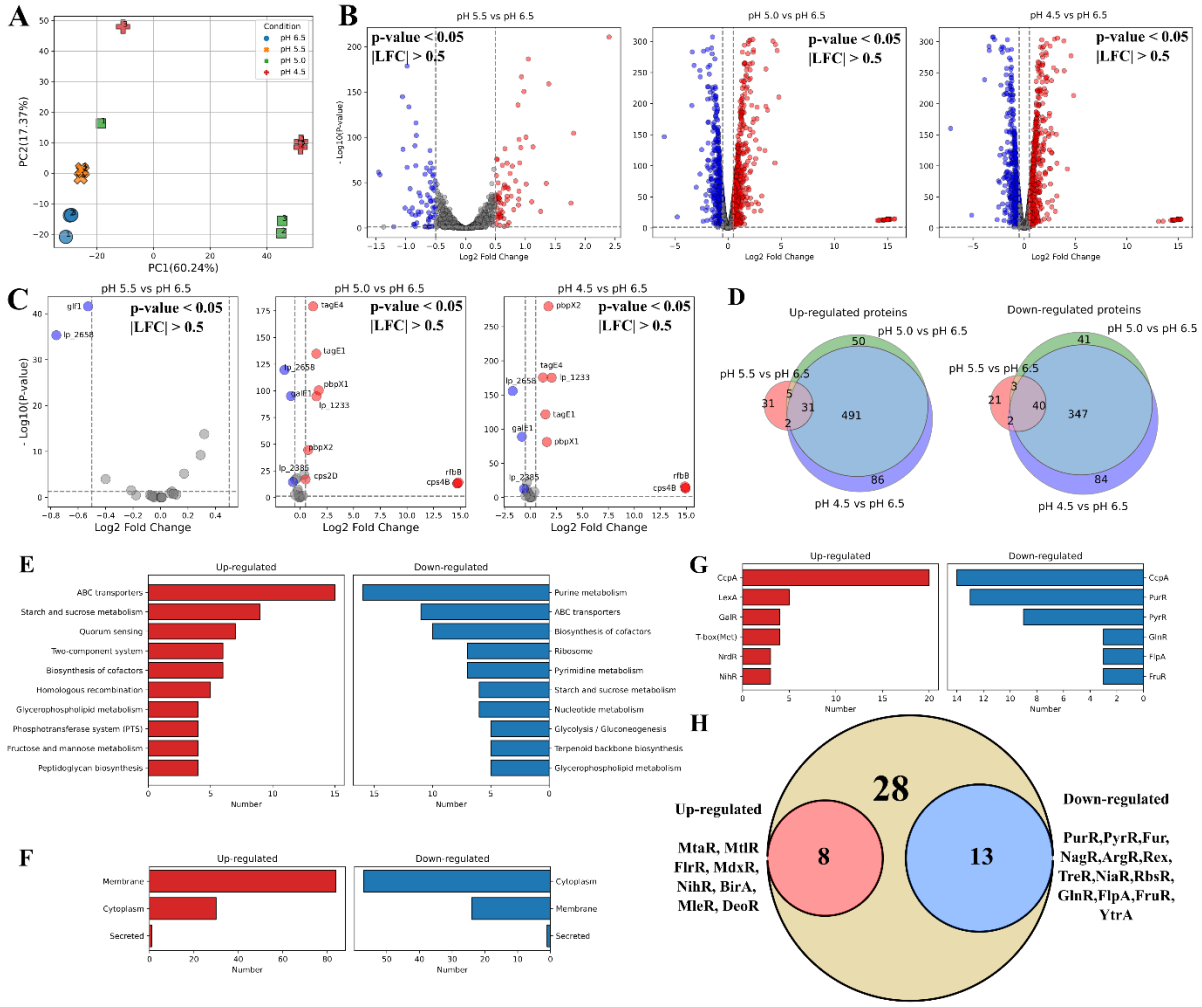

Figure 6.4. Differential expression analysis of proteomics at different pH values. (A) The PCA biplot of proteomics data. (B) Volcano plots of differential expression analysis of pH 5.5 vs pH 6.5, pH 5.0 vs pH 6.5 and pH 4.5 vs pH 6.5. LFC: log2-fold change. (C) Volcano plots of differential expression analysis for proteins encoded by the EPS BGC of pH 5.5 vs pH 6.5, pH 5.0 vs pH 6.5 and pH 4.5 vs pH 6.5. (D) Venn diagrams of up- and down-regulated proteins of pH 5.5 vs pH 6.5, pH 5.0 vs pH 6.5 and pH 4.5 vs pH 6.5 (p-value<0.05, |LFC|>0.5). (E) KEGG pathway enrichment analysis of differentially expressed proteins (p-value<0.05, |LFC|>0.5). (F) Subcellular location enrichment analysis of differentially expressed proteins (p-value<0.05, |LFC|>0.5). (G) Regulon enrichment analysis of differentially expressed proteins (p-value<0.05, |LFC|>0.5). (H) Up- and down-regulated transcription factor proteins under acidic conditions (p-value<0.05, |LFC|>0.5). The p-value was adjusted with Benjamini-Hochberg method. The number of replicates was three.

6.3.4 Acid stress induced proteome resource allocation

With gene-reaction-protein rules in the GSMM of LP HMX2, the protein expression level changes of metabolic reactions between pH 4.5 and pH 6.5 were computed to investigate how the proteome resource allocation of metabolic pathways was affected by acid stress (**Figure 6.5**). At pH 4.5, most glycolytic enzymes were only slightly down-regulated compared to pH 6.5, except that L-lactate dehydrogenase (LDH_L) had a LFC of -0.78. Different from other glycolytic enzymes, Glyceraldehyde-3-phosphate dehydrogenase (GAPD) was significantly up-regulated. In the pentose phosphate pathway (PPP), Phosphogluconate dehydrogenase (GND) and Ribulose 5-phosphate 3-epimerase (RPE) were significantly down-regulated, while Transaldolase (TALA) was significantly up-regulated. For purine and pyrimidine biosynthesis, most enzymes were significantly down-regulated, particularly Dihydroorotase (DHORTS), Dihydoorotic acid dehydrogenase (DHORD) and Orotate phosphoribosyltransferase (ORPT). In

general, acid stress resulted in the reduction of proteome resources allocated to glycolysis and the biosynthesis of DNA/RNA materials.

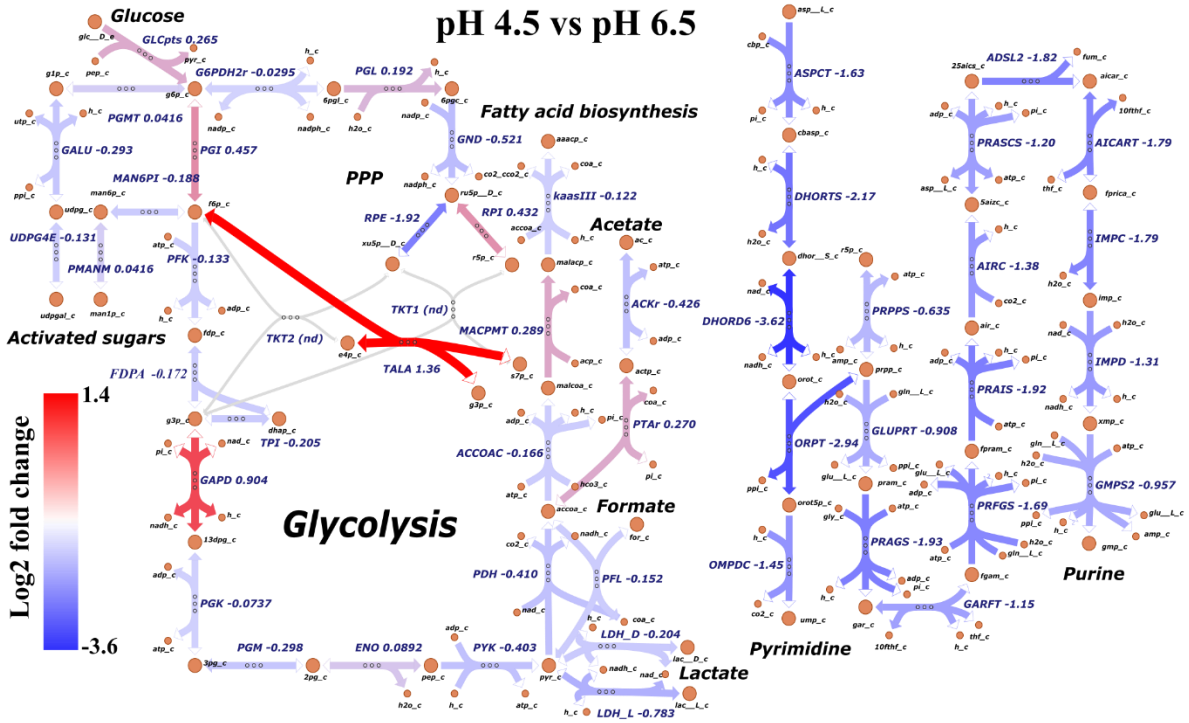

Figure 6.5. Protein expression level changes (pH 4.5 vs pH 6.5) of metabolic reactions in central carbon metabolism and pyrimidine/purine metabolism. Detailed reaction and enzyme information can be found in **Appendix D, Table D.2**.

In the theoretical model of proteome resource allocation proposed in this study, the proteome of LP HMX2 was divided into an inflexible housekeeping sector (Q) and four flexible sectors (A, C, T, U) (**section 6.2.6**). The C sector included all glycolytic enzymes; the T sector included glucose uptake via phosphoenolpyruvate-dependent sugar PTS and acid transporters; the A sector included all ribosomal and chaperone proteins; and the U sector contained all proteins encoded by the EPS-BGC (**Figure 6.6A**). The mass fractions of four flexible sectors at different pH values were computed with protein expression levels (**Figure 6.6B**). The fraction of the C sector dropped substantially when the pH decreased from 5.5 to 5.0, but the declining

trend then became smaller from 5.0 to 4.5. The fraction of the A sector dropped when the pH decreased from 6.5 to 5.5, but only had slight decreases for lower pH values. The fraction of the T sector increased continuously with the decrease of pH, which indicated that the inhibition of transporter activity at low pH demanded the cell to reallocate more proteome resources for maintaining the uptake of nutrients. When the pH decreased from 5.5 to 5.0, the fraction of the U sector had a ~2-fold increase, in agreement with the observed ~2-fold increase of the EPS production flux when the pH decreased from 6.5 to 5.0 (**Figure 6.2D**).

Furthermore, the significantly negative correlations of A and C sectors with the U sector depicted an explicit proteome trade-off between primary and secondary metabolism (**Figure 6.6CD**), which was also reported in the elucidation of independently modulated genes (iModulons) in LP [14]. When the pH decreased from 6.5 to 4.5, the proteome resource occupied by energy metabolism and anabolism decreased, while more proteome resources were allocated to EPS biosynthetic pathway. In addition, a sigmoid function was employed to approximate the variation in the mass fraction of the U sector in response to the pH decrease (**Figure 6.6E**) and was subsequently incorporated into the proteome constrained FBA to simulate the EPS production flux (**section 6.2.6, Eq. 6.7**).

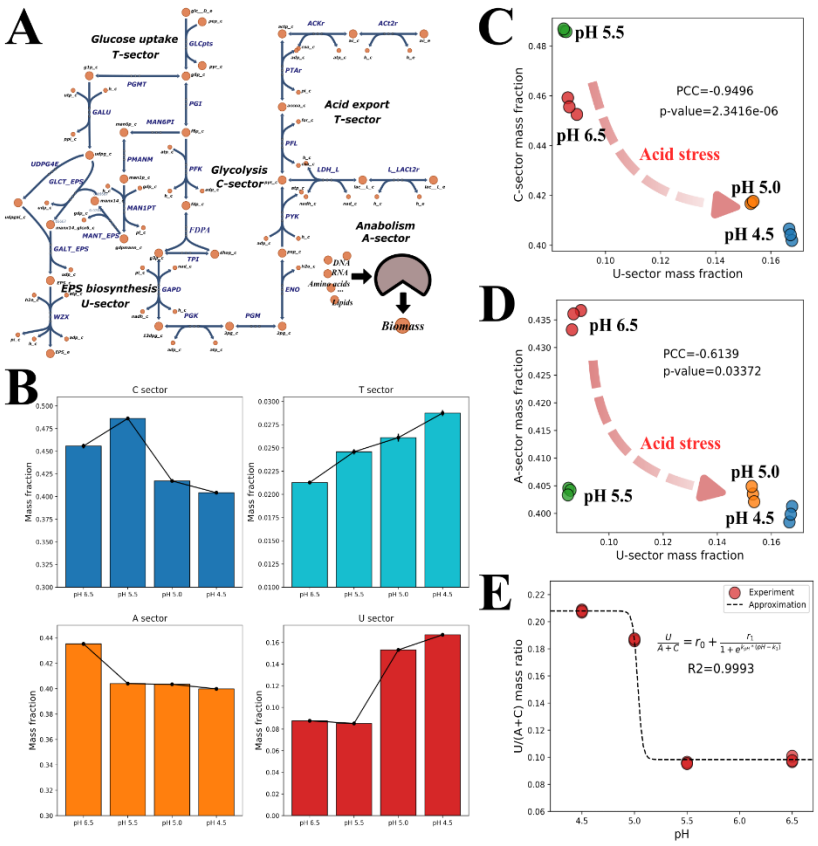

Figure 6.6. Proteome resource allocation among functional sectors under acid stress. (A) Illustration of flexible proteome sectors in the proteome resource allocation model. A: anabolism; T: transportation; C: catabolism; U: EPS biosynthesis. (B) Proteome sector mass fractions at different pH values. (C) The Pearson correlation of C and U sectors (p-value<0.05). (D) The Pearson correlation of A and U sectors (p-value<0.05). (E) The numerical approximation of $\frac{\phi_U}{\phi_C+\phi_A}$ as a sigmoid function of pH values (R2=0.9993). r_0=0.0876, r_1=0.07953, k_{pH}=44.9457, k_1=5.0344.

6.3.5 Simulation of primary metabolism and EPS production

Based on measured metabolomics data (**section 6.3.2**) and acid stress induced proteome resource allocation observed in this study (**section 6.3.4**), a regulatory proteome

constrained FBA (RPCFBA) model was constructed to simulate the primary metabolism and the EPS production flux of LP HMX2. The simulation predicted the maximum growth rates, EPS production and glucose uptake fluxes at different pH values ranging from 4.5 to 6.5, without considering the inhibition of undissociated lactic acid. The simulation results were compared with the experimental data to evaluate the accuracy of model predictions.

When pH decreased from 6.5 to 5.5, the simulated growth rate consecutively dropped from 0.79 $\frac{1}{hr}$ to 0.61 $\frac{1}{hr}$ due to the decrease of enzyme activities and the increase of $\frac{\phi_U}{\phi_C+\phi_A}$ (**Figure 6.7A**). Oppositely, the EPS production flux was enhanced by the increase of the proteome resource allocated to the EPS biosynthetic pathway (**Figure 6.7B**). Around pH 5.0, the sudden jump in the sigmoid function of $\frac{\phi_U}{\phi_C+\phi_A}$ (**Figure 6.6E**) led to the sudden decrease of simulated growth rate and sudden increase of EPS production flux (**Figure 6.7AB**). When the pH continued to decrease below 5.0, the simulated EPS production flux decreased, as the activities of EPS biosynthetic enzymes decreased, but the simulation predicted that the growth rate would slightly increase because the carbon flux and energy (activated sugars) consumed by EPS biosynthesis were reduced. However, the experimental growth rate at pH 4.5 did not become higher than that at pH 5.0, which could be resulted from the inaccuracy of the pH-dependency of enzyme activities parameterized in this model. The trend of simulated glucose uptake fluxes was close to that of simulated growth rates, but there appeared to be an observable inconsistency between simulated and experimental glucose uptake rates (**Figure 6.7C**). Such inconsistency possibly resulted from the underrating of the energy demand (ATP) for growth in the GSMM. In general, the simulated growth rates and EPS production fluxes were close to the experimental data (R2=0.579 for growth rates, R2=0.775 for EPS production fluxes).

Previous studies have indicated that different carbon sources can influence the EPS production rate of lactic acid bacteria [60,61]. Therefore, we conducted an *in-silico* perturbation

on the carbon source of the growth medium, where the sole carbon source was set as glucose, mannose or lactose, respectively. Similar to glucose, mannose can also be uptake via PTS, as the proteome of LP HMX2 contained PTS subunits specific for mannose (1_479, 1_480, and 1_481), but the low activity of Mannose-6-phosphate isomerase (MAN6PI) (in contrast to Glucose-6-phosphate isomerase (PGI)) will theoretically increase the proteome cost of glycolysis (**Figure 6.7D**). For lactose, LP needs to first break it down to glucose and galactose, and thus requires two more steps than utilizing glucose (**Figure 6.7D**). However, the activities of Hexokinase (HEX) and Beta-galactosidase (LACZ) are high. With respect to EPS biosynthesis, the change of carbon source from glucose to lactose or mannose can directly provide galactose or mannose, which will reduce the proteome cost of using Mannose-6-phosphate isomerase (MAN6PI), Phosphoglucomutase (PGMT), UTP-glucose-1-phosphate uridylyltransferase (GALU) and UDPglucose 4-epimerase (UDPG4E) to synthesize UDP-galactose and GDP-mannose from glucose. As expected, the simulation predicted that the maximum growth rates of using glucose and lactose were close, while the maximum growth rates of using mannose were noticeably lower (**Figure 6.7E**). In contrast, the EPS production fluxes of using mannose and lactose were predicted to be higher than that of using glucose (**Figure 6.7F**). In conclusion, our *in-silico* perturbation study predicted that lactose will be the most favorable carbon source, compared to glucose and mannose, for optimal bacterial growth and EPS productivity.

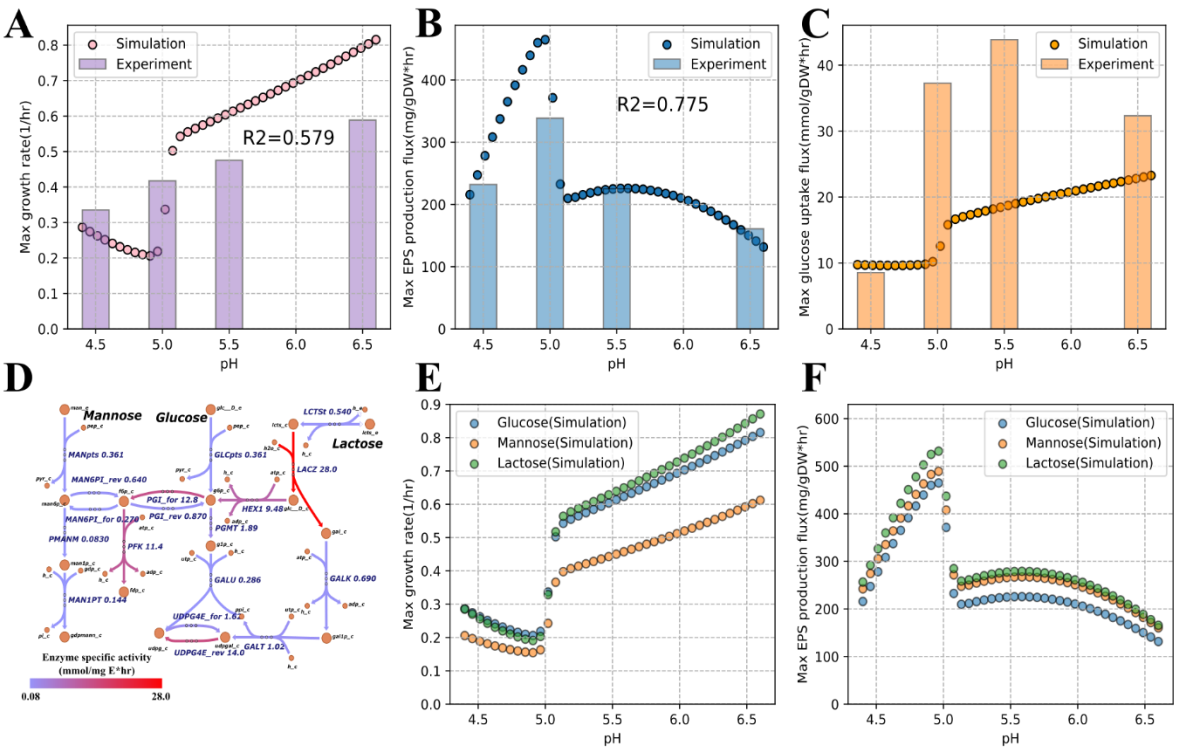

Figure 6.7. The simulation of primary metabolism and EPS production at different pH values. (A) The simulated and experimental growth rates ($\frac{1}{hr}$). (B) The simulated and experimental EPS production fluxes ($\frac{mg\ EPS}{gDW*hr}$). (C) The simulated and experimental glucose uptake fluxes ($\frac{mmol\ Glucose}{gDW*hr}$). (D) Enzyme activities of metabolic pathways for glucose, lactose and mannose. The details of enzyme activities can be found in **Appendix D, Table D.3**. (E) The simulated growth rates with different carbon sources. (F) The simulated growth rates with different carbon sources.

6.4 Conclusions and discussion

Although the high value of LP derived EPS has been demonstrated in various areas, the mechanistic understanding of the driving force behind its biosynthesis still remains to be uncovered. The observation that EPS production can be enhanced by a low pH in a recent

study [11], together with the development of proteome allocation theory [22,62,63], shed light on resolving this issue. This study confirmed that the EPS production flux indeed increased under acidic conditions. Subsequently, the proteomic analysis found that EPS biosynthetic proteins were significantly up-regulated by acid stress. It also identified the most probable TF, namely Fur, regulating the expression of EPS biosynthetic proteins (**section 6.3.3**). Unfortunately, due to the lack of knowledge on transcription units of LP, this study currently cannot derive a detailed regulatory mechanism for proteins encoded by the EPS-BGC. Nevertheless, the analysis of expression levels of functional proteome sectors quantitatively identified an acid stress induced proteome trade-off between primary metabolism and EPS biosynthesis in LP HMX2 (**section 6.3.4**), which unveiled the driving force behind EPS production in lactic acid bacteria.

To quantitatively describe the trade-off between primary and secondary metabolism, existing theoretical models such as Grime's competitor-stress-ruderal triangle [64,65] and Synthetic Chemostat Model [63] have made great contributions to propose a resource allocation framework, but none of them has been implemented to model metabolic fluxes yet. The RPCFBA model, constructed based on multi-omics data in this study, formulated the acid stress induced proteome resource allocation into a concise mathematical function (Eq. 6), and accurately predicted the changes of growth rates and EPS production fluxes in response to the decrease of pH (**section 6.3.5**). Also, the *in-silico* perturbation study on carbon sources demonstrated the broad potential and practical usefulness of RPCFBA on designing and optimizing complex fermentation processes, especially those involving secondary metabolisms for producing high-value bioproducts.

Nonetheless, some limitations remain in the presented modeling framework. The large deviation between experimental and simulated glucose uptake rates (**Figure 6.7C**) indicated that the energy demand (ATP) for growth in the GSMM might need to be re-evaluated for the specific LP strain, though the inaccuracy in the measurement of glucose concentrations was

also a possible error source. Another weakness of the current model is its inability to accurately represent pH dependent changes in enzyme activities. Enzyme k_{cat} predictors considering pH might resolve this issue, but the existing most accurate predictor can only achieve a RMSE of 0.594 [66]. Most importantly, the RPCFBA model cannot yet account for the inhibition of undissociated lactic acid on cellular growth of LP, thus is incapable of simulating the decrease of growth rate in time when the carbon source is abundant. The experimental data of this study reflected that the empirical function used in Özcan et al., 2021 [67] and Qiu et al., 2023 [47] to model the inhibition of undissociated lactic acid is not universally applicable. The growth rates at constant pH 6.5 and pH 5.5 dropped fast in time (**Figure 6.2C**), though the theoretical concentrations of undissociated lactic acid should be small based on Henderson-Hasselbalch equation. Consequently, a better method to model the complicated inhibitory effect of undissociated lactic acid on growth needs to be developed to enhance the performance of metabolic modeling of lactic acid bacteria grown in complex and changing conditions.

In a nutshell, this study provides insights into the growth strategy of LP HMX2 under acid stress, particularly on how it balances growth and stress response via reallocating its proteome resources between primary and secondary metabolism. In future studies, determining transcription units on the genome of LP and conducting ChIP-seq experiments for TFs associated with acid stress response (e.g., Fur) might uncover a detailed regulatory signaling pathway of how acid stress activates the expression of EPS biosynthetic proteins. The RPCFBA model established in this study can satisfactorily capture the changes of growth rates and EPS production fluxes induced by acid stress, though the quantitative accuracy can be further improved. With further refinement and tailored modifications, we can envisage that the RPCFBA model can potentially become a generic modeling framework for the design and control of biosynthesis of various valuable secondary metabolites associated with stress response (e.g., antibiotics produced by actinomycetes [68]).

References

1. Seddik HA, Bendali F, Gancel F, Fliss I, Spano G, Drider D. Lactobacillus plantarum and Its Probiotic and Food Potentialities. Probiotics Antimicrob Proteins. 2017;9: 111–122.

2. De Angelis M, de Candia S, Calasso MP, Faccia M, Guinee TP, Simonetti MC, et al. Selection and use of autochthonous multiple strain cultures for the manufacture of high-moisture traditional Mozzarella cheese. Int J Food Microbiol. 2008;125: 123–132.

3. Zago M, Fornasari ME, Carminati D, Burns P, Suàrez V, Vinderola G, et al. Characterization and probiotic potential of Lactobacillus plantarum strains isolated from cheeses. Food Microbiol. 2011;28: 1033–1040.

4. Rigaux P, Daniel C, Hisbergues M, Muraille E, Hols P, Pot B, et al. Immunomodulatory properties of Lactobacillus plantarum and its use as a recombinant vaccine against mite allergy. Allergy. 2009;64: 406–414.

5. Silva LA, Lopes Neto JHP, Cardarelli HR. Exopolysaccharides produced by Lactobacillus plantarum: technological properties, biological activity, and potential application in the food industry. Ann Microbiol. 2019;69: 321–328.

6. Arasu MV, Al-Dhabi NA, Ilavenil S, Choi KC, Srigopalram S. In vitro importance of probiotic Lactobacillus plantarum related to medical field. Saudi J Biol Sci. 2016;23: S6–S10.

7. Korcz E, Varga L. Exopolysaccharides from lactic acid bacteria: Techno-functional application in the food industry. Trends Food Sci Technol. 2021;110: 375–384.

8. Gacem MA, Krantar K, Hadef S, Boudjemaa B. Chapter 6 - Secondary metabolites from lactic acid bacteria as a source of antifungal and antimycotoxigenic agents. In: Abd-Elsalam

KA, Mohamed HI, editors. Bacterial Secondary Metabolites. Elsevier; 2024. pp. 107–122.

9. Scott M, Hwa T. Bacterial growth laws and their applications. Curr Opin Biotechnol. 2011;22: 559–565.

10. Nguyen P-T, Nguyen T-T, Bui D-C, Hong P-T, Hoang Q-K, Nguyen H-T. Exopolysaccharide production by lactic acid bacteria: the manipulation of environmental stresses for industrial applications. AIMS Microbiol. 2020;6: 451–469.

11. Nguyen P-T, Nguyen T-T, Vo T-N-T, Nguyen T-T-X, Hoang Q-K, Nguyen H-T. Response of Lactobacillus plantarum VAL6 to challenges of pH and sodium chloride stresses. Sci Rep. 2021;11: 1301.

12. Jung S, Lee J-H. Characterization of transcriptional response of Lactobacillus plantarum under acidic conditions provides insight into bacterial adaptation in fermentative environments. Sci Rep. 2020;10: 19203.

13. Zhai Z, Yang Y, Wang H, Wang G, Ren F, Li Z, et al. Global transcriptomic analysis of Lactobacillus plantarum CAUH2 in response to hydrogen peroxide stress. Food Microbiol. 2020;87: 103389.

14. Qiu S, Huang Y, Liang S, Zeng H, Yang A. Systematic elucidation of independently modulated genes in Lactiplantibacillus plantarum reveals a trade-off between secondary and primary metabolism. Microb Biotechnol. 2024;17: e14425.

15. Heunis T, Deane S, Smit S, Dicks LMT. Proteomic profiling of the acid stress response in Lactobacillus plantarum 423. J Proteome Res. 2014;13: 4028–4039.

16. Seesuriyachan P. Statistical modeling and optimization for exopolysaccharide production by Lactobacillus confusus in submerged fermentation under high salinity stress.

Food Sci Biotechnol. 2012;21: 1647–1654.

17. Imran MYM, Reehana N, Jayaraj KA, Ahamed AAP, Dhanasekaran D, Thajuddin N, et al. Statistical optimization of exopolysaccharide production by Lactobacillus plantarum NTMI05 and NTMI20. Int J Biol Macromol. 2016;93: 731–745.

18. Elmansy EA, Elkady EM, Asker MS, Abdallah NA, Khalil BE, Amer SK. Improved production of lactiplantibacillus plantarum RO30 exopolysaccharide (REPS) by optimization of process parameters through statistical experimental designs. BMC Microbiol. 2023;23: 361.

19. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248.

20. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem. 2006;281: 40041–40048.

21. Xu N, Liu J, Ai L, Liu L. Reconstruction and analysis of the genome-scale metabolic model of Lactobacillus casei LC2W. Gene. 2015;554: 140–147.

22. Qiu S, Yang A, Zeng H. Flux balance analysis-based metabolic modeling of microbial secondary metabolism: Current status and outlook. PLoS Comput Biol. 2023;19: e1011391.

23. Hu G, Wang Y, Xue R, Liu T, Zhou Z, Yang Z. Effects of the Exopolysaccharide from Lactiplantibacillus plantarum HMX2 on the Growth Performance, Immune Response, and Intestinal Microbiota of Juvenile Turbot, Scophthalmus maximus. Foods. 2023;12. doi:10.3390/foods12102051

24. Kemp TL, Karim MN, Linden JC, Tengerdy RP. Response surface optimization ofLactobacillus plantarum batch growth. Biotechnol Lett. 1989;11: 817–820.

25. Wegkamp A, Teusink B, de Vos WM, Smid EJ. Development of a minimal growth medium for Lactobacillus plantarum. Lett Appl Microbiol. 2010;50: 57–64.

26. DuBois M, Gilles KA, Hamilton JK, Rebers PA, Smith F. Colorimetric Method for Determination of Sugars and Related Substances. Anal Chem. 1956;28: 350–356.

27. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13: 1050–1054.

28. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119.

29. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35: D61–5.

30. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

31. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43: D261–9.

32. Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res.

2022;50: D571–D577.

33. Saier MH, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, et al. The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 2021;49: D461–D467.

34. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339–46.

35. Yang Y, Feng F, Zhou Q, Zhao F, Du R, Zhou Z, et al. Isolation, purification and characterization of exopolysaccharide produced by Leuconostoc pseudomesenteroides YF32 from soybean paste. Int J Biol Macromol. 2018;114: 529–535.

36. Ciucanu I, Kerek F. A simple and rapid method for the permethylation of carbohydrates. Carbohydr Res. 1984;131: 209–217.

37. Wiśniewski JR. Filter-Aided Sample Preparation for Proteome Analysis. Methods Mol Biol. 2018;1841: 3–10.

38. Hervey WJ 4th, Khalsa-Moyers G, Lankford PK, Owens ET, McKeown CK, Lu T-Y, et al. Evaluation of affinity-tagged protein expression strategies using local and global isotope ratio measurements. J Proteome Res. 2009;8: 3675–3688.

39. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–D531.

40. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.

Bioinformatics. 2005;21: 3674–3676.

41. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38: 5825–5829.

42. Yu C-S, Chen Y-C, Lu C-H, Hwang J-K. Prediction of protein subcellular localization. Proteins. 2006;64: 643–651.

43. Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad547

44. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7: 74.

45. Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained Allocation Flux Balance Analysis. PLoS Comput Biol. 2016;12: e1004913.

46. Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. Brief Bioinform. 2023;25. doi:10.1093/bib/bbad506

47. Qiu S, Zeng H, Yang Z, Hung W-L, Wang B, Yang A. Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. Biotechnol Bioeng. 2023. doi:10.1002/bit.28492

48. Zeng H, Yang A. Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. Sci Rep. 2020;10: 4283.

49. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM,

Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

50. Wu H, Zhao Y, Du Y, Miao S, Liu J, Li Y, et al. Quantitative proteomics of Lactococcus lactis F44 under cross-stress of low pH and lactate. J Dairy Sci. 2018;101: 6872–6884.

51. Wemmenhove E, van Valenberg HJF, van Hooijdonk ACM, Wells-Bennik MHJ, Zwietering MH. Factors that inhibit growth of Listeria monocytogenes in nature-ripened Gouda cheese: A major role for undissociated lactic acid. Food Control. 2018;84: 413–418.

52. Nishimura J, Kawai Y, Aritomo R, Ito Y, Makino S, Ikegami S, et al. Effect of Formic Acid on Exopolysaccharide Production in Skim Milk Fermentation by *Lactobacillus delbrueckii* subsp. *bulgaricus* OLL1073R-1. Biosci Microbiota Food Health. 2013;32: 23–32.

53. Behera SS, Ray RC, Zdolec N. *Lactobacillus plantarum* with Functional Properties: An Approach to Increase Safety and Shelf-Life of Fermented Foods. BioMed Research International. 2018. pp. 1–18. doi:10.1155/2018/9361614

54. Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, et al. RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics. 2013;14: 745.

55. Sun F, Gao H, Zhang Y, Wang L, Fang N, Tan Y, et al. Fur is a repressor of biofilm formation in Yersinia pestis. PLoS One. 2012;7: e52392.

56. González A, Bes MT, Peleato ML, Fillat MF. Expanding the Role of FurA as Essential Global Regulator in Cyanobacteria. PLoS One. 2016;11: e0151384.

57. Ma J, An C, Jiang F, Yao H, Logue C, Nolan LK, et al. Extraintestinal pathogenic Escherichia coli increase extracytoplasmic polysaccharide biosynthesis for serum resistance in response to bloodstream signals. Mol Microbiol. 2018;110: 689–706.

58. Hall HK, Foster JW. The role of fur in the acid tolerance response of Salmonella typhimurium is physiologically and genetically separable from its role in iron acquisition. J Bacteriol. 1996;178: 5683–5691.

59. Pflock M, Kennard S, Finsterer N, Beier D. Acid-responsive gene regulation in the human pathogen Helicobacter pylori. J Biotechnol. 2006;126: 52–60.

60. Yuksekdag ZN, Aslim B. Influence of different carbon sources on exopolysaccharide production by Lactobacillus delbrueckii subsp. bulgaricus (B3, G12) and Streptococcus thermophilus (W22). Braz Arch Biol Technol. 2008;51: 581–585.

61. Zhang Y, Dai X, Jin H, Man C, Jiang Y. The effect of optimized carbon source on the synthesis and composition of exopolysaccharides produced by Lactobacillus paracasei. J Dairy Sci. 2021;104: 4023–4032.

62. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Mol Syst Biol. 2015;11: 784.

63. Panikov NS. Genome-Scale Reconstruction of Microbial Dynamic Phenotype: Successes and Challenges. Microorganisms. 2021;9. doi:10.3390/microorganisms9112352

64. Grime JP. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory. Am Nat. 1977;111: 1169–1194.

65. Bruggeman FJ, Teusink B, Steuer R. Trade-offs between the instantaneous

growth rate and long-term fitness: Consequences for microbial physiology and predictive computational models. Bioessays. 2023;45: e2300015.

66. Jiang H, Wang J, Yang Z, Chen C, Yao G, Bao S, et al. MPEK: a multi-task learning based on pre-trained language model for predicting enzymatic reaction kinetic parameters. Research Square. 2024. doi:10.21203/rs.3.rs-3916417/v1

67. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. Biotechnol Bioeng. 2021;118: 223–237.

68. Mohite OS, Weber T, Kim HU, Lee SY. Genome-Scale Metabolic Reconstruction of Actinomycetes for Antibiotics Production. Biotechnol J. 2019;14: e1800377.

Chapter 7 Conclusions and perspectives

7.1 Summary of the PhD study and key contributions

The research work accomplished in this PhD study, from October, 2021 to June, 2024, consists of a systematic literature review on using FBA to model microbial metabolism under environmental stress (**Chapter 2**) and four research projects that analyze and model the effects of undissociated weak acid, acidic pH and temperature changes on the growth and metabolism of lactic acid bacteria (**Chapters 3-6**). Three FBA-based models developed in this study (i.e., dynamic proteome constrained community-level FBA, temperature sensitive proteome constrained FBA and regulatory proteome constrained FBA (RPCFBA)) all incorporated environmental stress, though the quantitative accuracy of some models still remained to be improved. Generally speaking, this study demonstrated that proteome allocation based metabolic modeling that reflects the effect of environmental stress is a relatively simple (in contrast to complex regulatory networks), interpretable and useful “white-box” modeling method for microbial metabolism under stress.

This research started with modeling the fermentation kinetics of yogurt with a metagenome-scale metabolic model of dominant species (ST and LB) in the starter culture. Project 1 (**Chapter 3**) presented a novel modeling method that integrated proteome allocation into dynamic FBA and used a feed-back inhibition function on the proteome constraint of carbon source uptake to model the effect of undissociated lactic acid. As the first dynamic community FBA model with proteome constraints, the model in Project 1 could accurately simulate the growth-coupled primary metabolism of ST and LB under the stress of accumulating undissociated lactic acid. Furthermore, this model could account for metabolic interactions between ST and LB, and thus could evaluate the impact of modulating the starter culture composition on the fermentation behavior.

Subsequently, this research investigated the modeling of the metabolic response of microorganisms to temperature changes using a combination of deep learning and proteome constrained FBA. In Project 2 (**Chapter 4**), a deep learning-based predictor of temperature dependent enzyme k_{cat}, DLTKcat, was constructed and achieved a superior accuracy in comparison with previous models (RMSE ($log_{10}\ scale$)=0.88). For the first time, the significant feature importance of temperature in predicting k_{cat} was demonstrated. Then, Project 2 demonstrated that temperature sensitive proteome constrained FBA with predicted k_{cat} at different temperatures could capture the metabolic response to temperature changes. In short, DLTKcat has the potential to become a computational tool to quantitatively model temperature dependent microbial metabolism.

Next, Project 3 (**Chapter 5**) and Project 4 (**Chapter 6**) investigated the driving force behind the biosynthesis of EPS in LP, a secondary metabolic pathway to mediate acid stress. In Project 3, 45 iModulons in LP were obtained using ICA of transcriptomic data to reconstruct the TRN, and the analysis of iModulon activities revealed a trade-off relationship between the regulatory activities of primary metabolism and acid stress induced EPS biosynthesis. Building upon the results of Project 3 and with the support of multi-omics analysis, Project 4 identified a proteome trade-off between primary metabolism and acid stress induced EPS biosynthesis in LP HMX2, and then constructed the RPCFBA model that can accurately simulate the cellular growth and EPS production flux simultaneously. The acid stress induced proteome resource reallocation, unveiled in this study, contributes to the understanding of microbial secondary metabolism and lactic acid bacteria physiology, and the RPCFBA model has the potential to become a generic mechanistic modeling framework for stress induced microbial secondary metabolism.

In conclusion, this PhD study developed a set of proteome allocation based models to simulate the growth-coupled primary metabolism under stress and stress-induced secondary metabolism. The performance of these models has been demonstrated through applications on

lactic acid bacteria, while the underlying approach holds the potential to be applicable on the prediction of other microbial systems where the impact of environmental stress on primary and/or secondary metabolism is of great importance.

7.2 Limitations and future perspectives

Despite the achievements presented above, the quantitative accuracy of computational models and the quality of omics-based analysis in this PhD were limited due to drawbacks such as the neglect of some key environmental factors. Although the metagenome-scale metabolic model presented in Project 1 characterized the product inhibition of undissociated lactic acid, the influence of pH decrease on activities of metabolic enzymes was not incorporated into proteomic constraints due to the lack of enzyme kinetic data for ST and LB. Similarly, the RPCFBA model for *L. plantarum* HMX2 in Project 4 could not accurately account for the pH dependent change of enzyme activities. Using deep learning to predict pH dependent enzyme k_{cat} is probably the optimal solution, as enzyme kinetic assay is expensive and labor intensive.

With respect to deep learning based k_{cat} prediction, the prediction error of DLTKcat is not low enough (RMSE ($log_{10}\ scale$)=0.88) to enable accurate temperature sensitive proteome constrained FBA. The condition parameters (e.g., temperature and pH) are not involved in attention weighted features of protein sequences and substrates. Also, the influence of pH on enzyme k_{cat} is not included in the current deep learning model. There are two potential approaches to enhance the prediction accuracy: 1. use pre-trained large language models for protein sequences and substrates to generate embeddings in place of dictionary-based embeddings; 2. compute attention weighted features between condition parameters (temperature and pH) and protein residues in addition to bi-directional attention features between protein residues and substrate atoms.

Another important factor neglected in the metagenome-scale metabolic model in Project 1 was inter-species gene expression regulatory interactions in the lactic acid bacteria community, such as the stimulation of ST's growth by formic acid, casitone, pyruvic acid, folic acid, and polysorbate 20. Such regulatory interactions can be discovered from the integrative analysis of meta-transcriptomics and metabolomics. Regarding *in-silico* reconstruction of gene expression regulatory network, the ICA based inference of TRN is recognized as a state of art computational method, but the quality of reconstructed TRN of LP in Project 3 was restricted by the sample size and the lack of operon annotation. Appending the sample size with more transcriptomic data under different conditions (especially conditions of interests) will theoretically enhance the quality of ICA based inference of TRN.

Expectedly, limitations discussed above will be resolved in the future with the accumulation of experimental data (e.g., enzyme kinetic parameters or gene expression levels under different conditions) and further improvement of proteome constrained FBA modeling framework. The inclusion of new data will benefit the deep learning-based prediction of enzyme k_{cat}, possibly leading to accurate prediction of temperature and pH dependent enzyme k_{cat}. Furthermore, the enhanced predictive accuracy on enzyme kinetic parameters might allow us to gradually convert FBA-based metabolic models to fine-grained whole-cell kinetic models, which are currently only implementable for well-studied cells like *E. coli* or red blood cells. With high quality reconstruction of TRNs, a novel method to combine TRN and metabolic networks can potentially be developed to bring fine-grained gene regulatory interactions into proteome allocation based metabolic modeling, enabling FBA to account for metabolic phenomena that cannot be modeled for now (e.g., the production of methyl ketones by lactic acid bacteria in dairy fermentation).

As envisaged, a proteome allocation-based modeling framework incorporating major types of stress responses of lactic acid bacteria will be developed for both mono- and co-cultures, with much-improved quantitative accuracy. Such a modeling framework can also be

applied to other microbial processes involving stress responses with case-specific modifications. Due to the high interpretability of white-box metabolic models, researchers will be able to perform *in-silico* perturbations on reaction pathways, growth media and co-culture compositions, facilitating computer aided design and control of microbial processes.

Appendix A

A.1 Supplementary methods

A.1.1 Genome scale metabolic model refinement

The growth function, that converts amino acids, DNA, RNA, lipids, polysaccharides, vitamines, and metal ions into cellular biomass, is essential for the prediction of growth rate by FBA. For ST, the biomass composition of *Streptococcus thermophilus* LMG18311 was used, which is 46% protein, 2.3% DNA, 10.7% RNA, 12% polysaccharides 3.4% lipids and 25.6% other components [1]. With the relative molar fractions in the same type of metabolites fixed, the stoichiometric coefficients of 20 essential amino acids, 4 DNAs, 4 RNAs, lipids, and polysaccharides in the growth function of ST were adjusted to fit the measured biomass composition. For LB, the growth function of *Lactobacillus delbrueckii subsp. bulgaricus* ATCC 11842 in AGORA [2] was adopted. The growth associated ATP requirement of *Streptococcus thermophilus* CH8, that is 18.15 mmol/gDW [3], was used for ST's growth function. For LB's growth function, the growth associated ATP requirement of *Lactobacillus plantarum* WCFS1, 27.4 mmol/gDW [4], was used.

To refine reconstructed metabolic networks, reactions added erroneously were removed. Methionine synthase (METS) was removed from ST's GSMM, because this reaction does not have any gene associated with it and gene functional annotation indicated that there is no methionine synthase in the MAG of ST. UDPglucose--hexose-1-phosphate uridylyltransferase (UGLT) was removed from LB's GSMM because LB is reported to be galactose negative [5], but UGLT allows the metabolism of galactose.

Missing Reactions were added manually. The pentose phosphate pathways in both ST and LB's raw GSMMs were incomplete initially, and therefore, Transketolase (TKT1, TKT2) and

Transaldolase (TALA) were added to LB's GSMM, and Phosphogluconate dehydrogenase (GND), 6-phosphogluconolactonase (PGL) and Glucose 6-phosphate dehydrogenase (G6PDH2r) were added to ST's GSMM. Based on gene functional annotation, Galactokinase (GALKr) was added to ST's GSMM, and Pyruvate dehydrogenase (PDH) was added to LB's GSMM. In addition, transport reactions were added for lactate (lac), diacetyl (diacet), formate (for), acetate (ac), succinate (succ), adenine (ade), 4-Hydroxy-benzyl alcohol (4hba), 3-Methyl-2-oxobutanoate (3mob), (S)-3-Methyl-2-oxopentanoate (3mop), and 4-Methyl-2-oxopentanoate (4mop). To model casein peptide utilization, a reaction named "**PROLYSIS**" that decomposes casein peptide into amino acids was added:

caspep_c --> 0.87 his__L_c + 1.15 ile__L_c + 2.48 leu__L_c + 1.99 lys__L_c + 0.74 met__L_c + 1.06 phe__L_c + 1.53 thr__L_c + 0.19 trp__L_c + 1.56 val__L_c + 1.42 ala__L_c + 0.72 arg__L_c + 2.1 asn__L_c + 0.26 cys__L_c + 5.53 glu__L_c + 0.84 gly_c + 2.13 pro__L_c + 2.11 ser__L_c + 1.01 tyr__L_c

A.1.2 Simulation and validation

FBA was performed using COBRApy [6]. For dFBA, the time range was 6 hours and the number of time steps was 1000. The code of model implementation can be found in https://github.com/SizheQiu/MetaStLbCom. Three growth medium parameters were used in this study: complete CDM [3] , MPL medium [7] and milk environment, in all of which lactose is abundant to fully saturate the transporter protein. The first two have been clearly defined in previous works. The composition of the milk environment was defined based on the measurement of free amino acid concentrations in 12% (w/w) reconstituted whole milk (**Table B.2**), and formic acid concentration was assumed to be 2 mM [8].

For dFBA in section 3.3 in main text, the initial condition settings at t=0 were: the biomass concentrations of ST and LB were both 0.09 gDW/L, the concentration of lactose was 191.4 mM [9], the concentrations of free amino acids and formic acid were the same as the milk

environment defined above. To assess the accuracy of dynamic simulation, R-squared values were computed with concentrations of ST and LB's biomass, lactic acid, lactose in reference data and simulation at 0, 1.5, 3, 4.5, 6 hr. For dFBA in section 3.4 in main text, the initial condition settings at t=0 were the same as those in section 3.3, except varying initial ST:LB inoculation ratios.

A.1.3 Parameter estimation

LB's growth rate in MPL medium was determined to be 0.7/hr [7], and the value of growth rate was used to estimate its v_{max} of amino acid uptake rate. The v_{max} of amino acid uptake in LB was estimated to be 0.288 mmol/gDW*hr (**Figure B.3A**). The growth rate 0.64/hr, when Streptococcus thermophilus CH8 only utilizes casein peptide as the nitrogen source [3], was used to estimate ST's v_{max} of casein peptide utilization, and the v_{max} was estimated to be 0.091 mmol/gDW*hr (**Figure B.3B**). And LB's maximum flux of casein peptide utilization in the milk environment was found to be 0.653 mmol/gDW*hr with uptake fluxes of all individual amino acids set to 0, which was set as LB's v_{max} of casein peptide utilization.

For the inhibition of undissociated lactic acid, k_{LacH} for the uptake of lactose (Eq. 10 in main text), free amino acid and casein peptide were set at 0.1/mM to apply mild inhibition on bacterial growth and metabolism [10]. The minimal growth rates of both ST and LB at pH = ~ 4.5 (lactic acid concentration = ~ 120mM) were reported to be around 0.1/hr in previous studies [9] and measurement in this study (**Table S1**), and this value was used to estimate the minimal activity of lactose transporter (**Figure S3C**). The minimal activity of the lactose transporter was estimated to be 10 for ST and 16 for LB. The v_{min} of amino acid uptake was set at 0.0277 mmol/gDW*hr for ST and 0.09 mmol/gDW*hr for LB so as not to limit the growth rate when the activity of the lactose transporter is inhibited to the minimal value.

A.2 Tables

Table A.1. Measured concentration profiles of lactic acid, lactose and total bacterial biomass, and pH changes in 300 minutes

T(min)	Lactic acid (mg/g milk)	Lactose (mg/g milk)	Biomass (CFU/ml)	pH
0	0.0316+/-0.0145	49.4731+/-1.3653	5800000	6.332
60	0.8903+/-0.0756	48.6823+/-1.6307	11000000	6.116
105	1.9703+/-0.1582	47.0768+/-0.9449	46700000	5.592
120	2.3740+/-0.1436	44.4647+/-1.4251	268000000	5.476
180	6.9361+/-0.0380	39.8116+/-3.7565	350000000	4.752
240	9.5818+/-0.3944	37.1539+/-2.8049	392000000	4.516
300	10.0638+/-0.2164	30.4263+/-1.1584	485000000	4.396

Table A.2. Concentrations of free amino acid in 12%(w/w) reconstituted milk

Amino acid	Concentration ($\mu g/g$)	Amino acid	Concentration ($\mu g/g$)
Asn	2.2146+/-0.1293	Tyr	2.3302+/-0.6454
Gln	2.7007+/-0.7595	Met	0.2298+/-0.1983
Ser	12.9733+/-3.9325	Val	3.4507+/-0.8787
Gly	10.5210+/-0.4277	Ile	1.5413+/-0.2401
Asp	3.8261+/-0.1823	Leu	1.9131+/-0.4222
Glu	35.8685+/-6.6227	Phe	1.1631+/-0.2211
Thr	3.6426+/-0.6210	Trp	1.1458+/-0.1658
Ala	6.9639+/-1.3911	Orn	4.1740+/-0.1615
Pro	4.2520+/-1.0392	_	_
Lys	2.9785+/-0.3263	_	_

* His, Arg and Cys were not detected.

Table A.3. Taxonomic profiles of the yogurt starter culture computed by mOTUs and KRAKEN-2.0

mOTUs		KRAKEN-2.0	
Species (top 10)	Abundance (%)	Species (top 10)	Abundance (%)
Streptococcus thermophilus	95.1774	Streptococcus thermophilus	98.4683
Lactobacillus delbrueckii	1.3427	Lactobacillus delbrueckii	0.8790
Anoxybacillus flavithermus	0.1390	Anoxybacillus flavithermus	0.1580
Geobacillus stearothermophilus	0.0238	Streptococcus infantarius	0.1513
Streptococcus salivarius	0.0148	Streptococcus gallolyticus	0.0900
Lactococcus piscium	0.0086	Streptococcus salivarius	0.0590
Anoxybacillus flavithermus	0.0083	Streptococcus macedonicus	0.0480
Vibrio sp.	0.0033	Streptococcus pneumoniae	0.0157
Streptococcus salivarius	0.0032	Geobacillus thermoleovorans	0.0137
Vibrio alginolyticus	0.0023	Clostridioides difficile	0.0067

Table A.4. Metabolic reaction information

ID	Name	EC number
LACZ	Beta-galactosidase	3.2.1.23
GALKr	Galactokinase	2.7.1.6
HEX	Hexokinase (D-glucose:ATP)	2.7.1.2
PGI	Glucose-6-phosphate isomerase	5.3.1.9
PFK	Phosphofructokinase	2.7.1.11
FBA	Fructose-bisphosphate aldolase	4.1.2.13
TPI	Triose-phosphate isomerase	5.3.1.1
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12
PGK	Phosphoglycerate kinase	2.7.2.3
PGM	Phosphoglycerate mutase	5.4.2.11
ENO	Enolase	4.2.1.11
PYK	Pyruvate kinase	2.7.1.40
LDH	Lactate dehydrogenase	1.1.1.27
PFL	Pyruvate formate lyase	2.3.1.54
PDH	Pyruvate dehydrogenase	1.2.7.1
PTAr	Phosphotransacetylase	2.3.1.8
ACKr	Acetate kinase	2.7.2.1
ACALD	Acetaldehyde dehydrogenase (acetylating)	1.2.1.10
ACCOAC	Acetyl-CoA carboxylase	6.4.1.2
G1PACT	Glucosamine-1-phosphate N-acetyltransferase	2.3.1.157
CYSDS	Cysteine Desulfhydrase	4.4.1.28
VPAMTr	Valine-pyruvate aminotransferase	2.6.1.66
THRA	Threonine aldolase	4.1.2.5

Table A.5. Parameters of the dynamic community-level metabolic model

Parameter	Model species	Value	Unit	Reference
Protein concentration	ST	0.46	g/gDW	[1]
Protein concentration	LB	0.529	g/gDW	Default biomass composition of gram positive bacteria [2]
Activity(EX_lcts_e)	ST&LB	540	mmol/g E * hr	[11]
Minimal activity(EX_lcts_e)	ST	10	mmol/g E * hr	Estimated
Minimal activity(EX_lcts_e)	LB	16	mmol/g E * hr	Estimated
Activity(acid exportation)	ST&LB	6360	mmol/g E * hr	[12]
Activity(LACZ)	ST&LB	27960	mmol/g E * hr	[13]
Activity(GALKr)	ST&LB	690	mmol/g E * hr	[14]
Activity(HEX)	ST&LB	9480	mmol/g E * hr	[15]
Activity(PGI)	ST&LB	12756	mmol/g E * hr	[16]
Activity(PFK)	ST&LB	11400	mmol/g E * hr	[17]
Activity(FBA)	ST&LB	28620	mmol/g E * hr	[18]
Activity(TPI)	ST&LB	3300	mmol/g E * hr	[19]
Activity(GAPD)	ST&LB	2400	mmol/g E * hr	[20]
Activity(PGK)	ST&LB	28800	mmol/g E * hr	[21]
Activity(PGM)	ST&LB	7440	mmol/g E * hr	[22]
Activity(ENO)	ST&LB	15600	mmol/g E * hr	[23]
Activity(PYK)	ST&LB	3300	mmol/g E * hr	[24]
Activity(LDH)	ST&LB	141000	mmol/g E * hr	[25]
Activity(PFL)	ST	9600	mmol/g E * hr	[26]
Activity(PDH)	ST&LB	1500	mmol/g E * hr	[27]
Activity(PTAr)	ST&LB	428400	mmol/g E * hr	[28]
Activity(ACKr)	ST&LB	65220	mmol/g E * hr	[29]

Activity(ribosome)	ST&LB	107.4	mmol/g E * hr	[30]
Km(Lactose)	ST&LB	10	mM	[31]
Km(amino acid)	ST&LB	0.001	mM	[32]
Vmax(amino acid)	ST	0.2	mmol/gDW*hr	[10],[32],[33]
Vmax(amino acid)	LB	0.288	mmol/gDW*hr	Estimated
Vmin(amino acid)	ST	0.0277	mmol/gDW*hr	Estimated
Vmin(amino acid)	LB	0.09	mmol/gDW*hr	Estimated
Vmax(casein utilization)	ST	0.091	mmol/gDW*hr	Estimated
Vmax(casein utilization)	LB	0.653	mmol/gDW*hr	Estimated
Vmax(formic acid)	LB	0.3	mmol/gDW*hr	[34]
Km(formic acid)	LB	15	mM	[34]
klach(lactose)	ST&LB	0.1	1/mM	Estimated
klach(amino acid)	ST&LB	0.1	1/mM	Estimated
klach(casein utilization)	ST&LB	0.1	1/mM	Estimated

Table A.6. Proteome costs for unit flux of acetic acid and lactic acid production pathways

Pathway	Species	Proteome cost
Acetic acid production	ST	$\frac{1\ mmol/gDW*hr}{a_{PFL}}+\frac{1\ mmol/gDW*hr}{a_{PTAr}}+\frac{1\ mmol/gDW*hr}{a_{ACKr}}$ $=0.1218\ mg\ E/gDW$
Acetic acid production	LB	$\frac{1\ mmol/gDW*hr}{a_{PDH}}+\frac{1\ mmol/gDW*hr}{a_{PTAr}}+\frac{1\ mmol/gDW*hr}{a_{ACKr}}$ $=0.6843\ mg\ E/gDW$
Lactic acid production	ST&LB	$\frac{1\ mmol/gDW*hr}{a_{LDH}}=0.0071\ mg\ E/gDW$

A.3 Figures

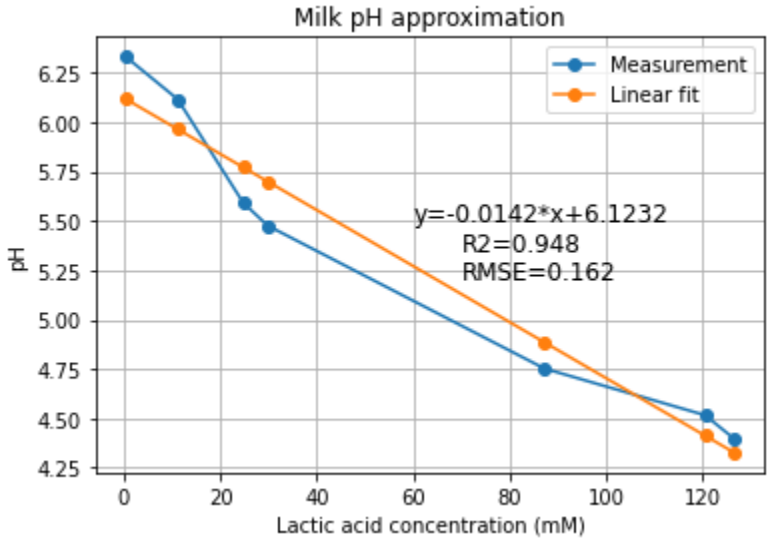

Figure A.1. Linear approximation of milk pH based on the concentration of lactic acid

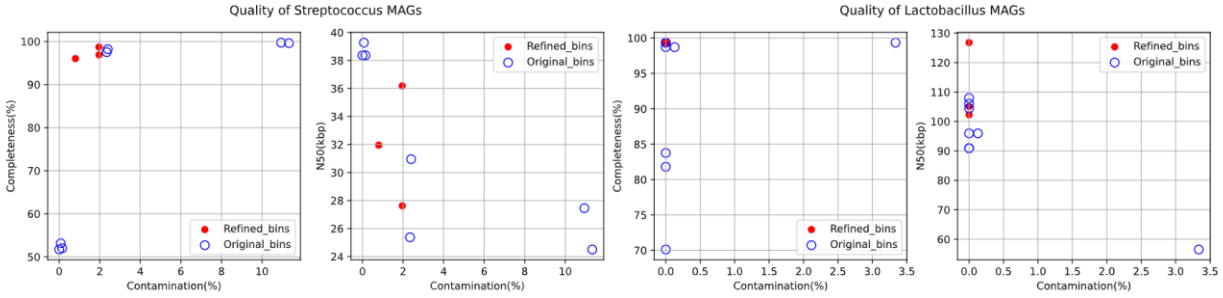

Figure A.2. Quality assessment of metagenome binning for ST and LB, completeness versus contamination, and N50(an assessment of assembly continuity) versus contamination.

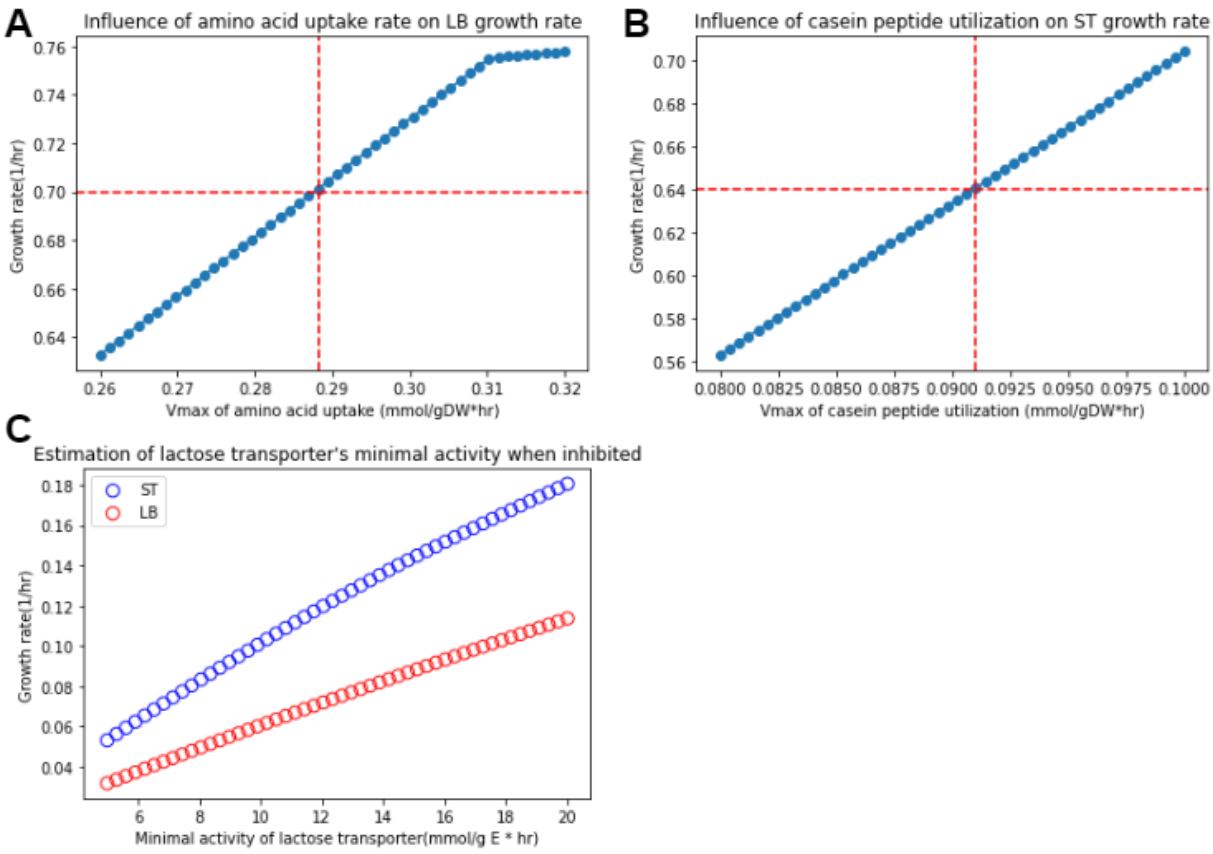

Figure A.3. Estimation of kinetic parameters: (A) Vmax of casein peptide utilization for ST, (B) Vmax of amino acid uptake for LB, (C) minimal activity of the lactose transporter for ST and LB

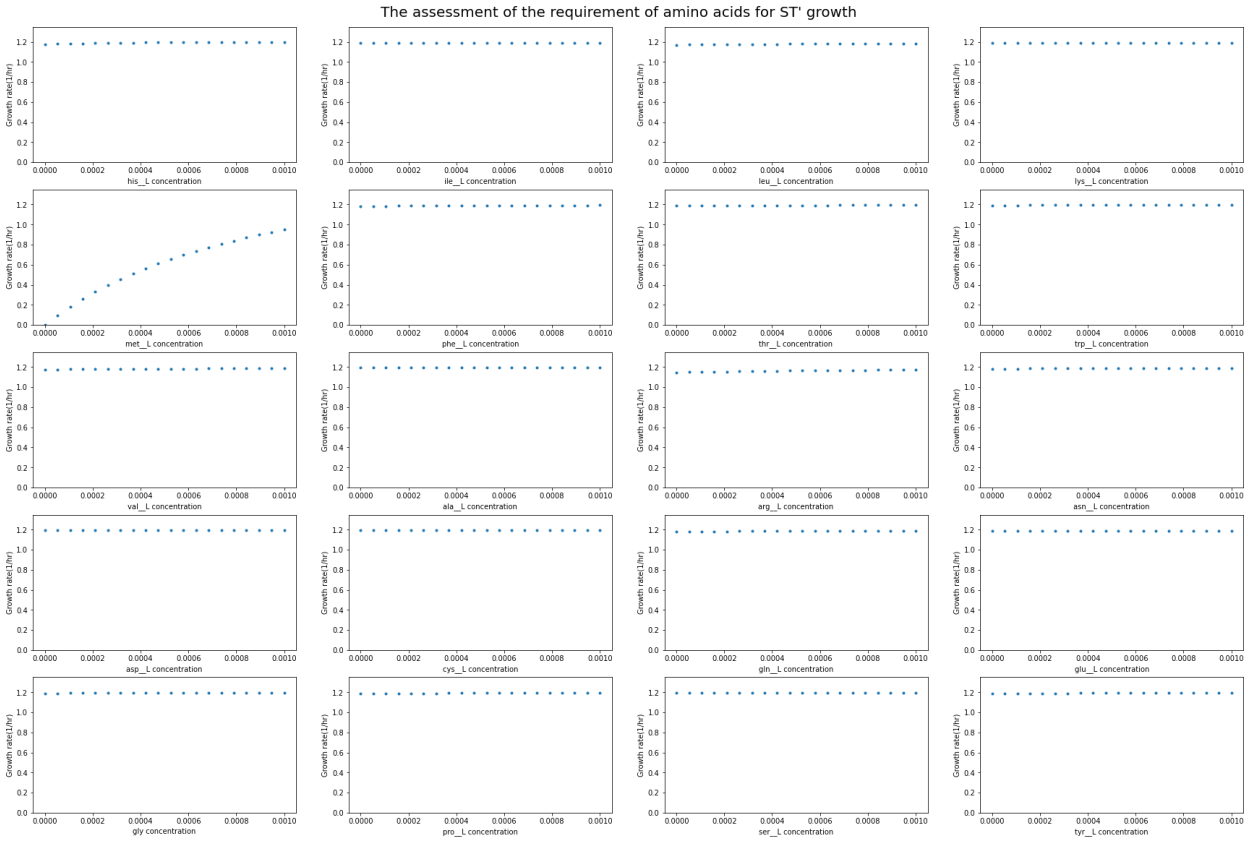

Figure A.4. The assessment of the requirement of amino acids for ST's growth by varying the amino acid concentration in complete CDM from 0 to 0.001 mM. Only methionine shows a significant impact on ST's growth rate.

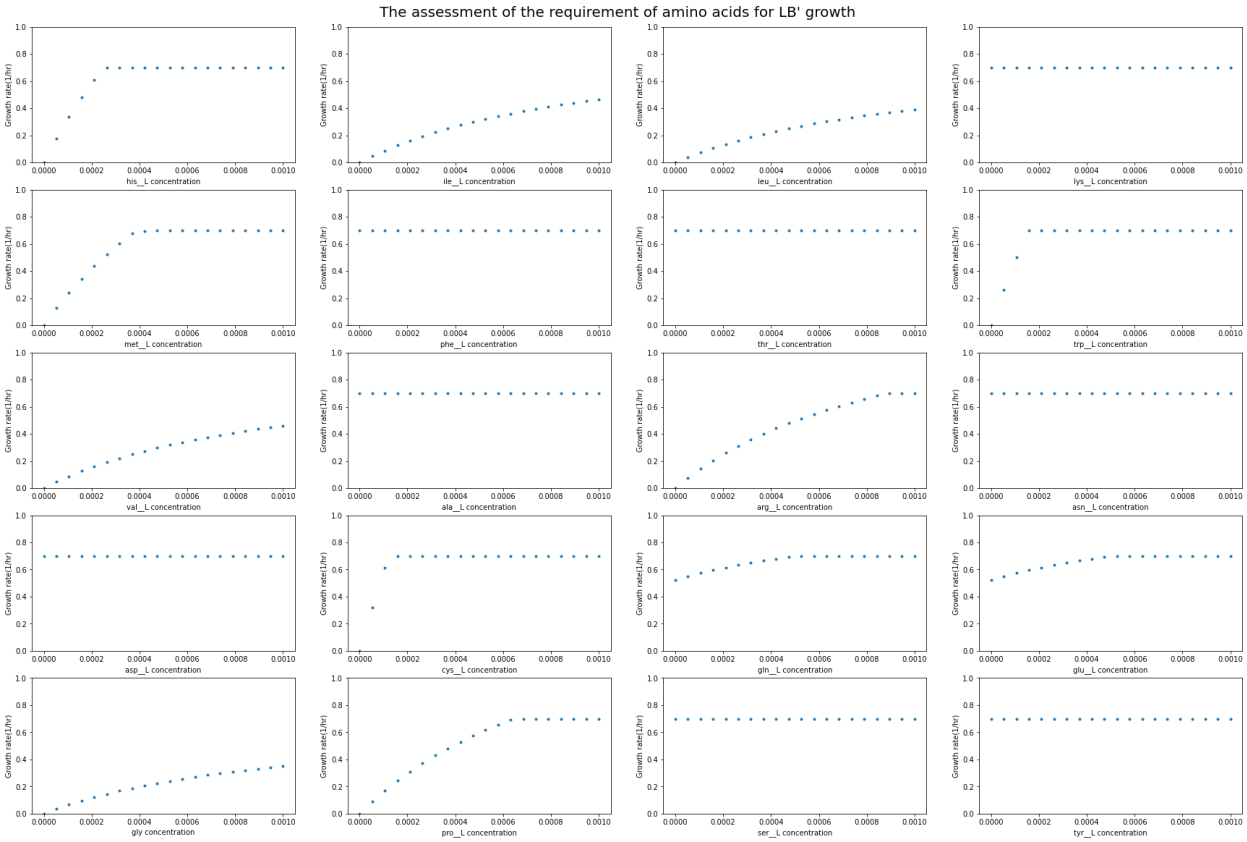

Figure A.5. The assessment of the requirement of amino acids for LB's growth by varying the amino acid concentration in MPL medium from 0 to 0.001 mM. Histidine, isoleucine, leucine, methionine, tryptophan, valine, arginine, cysteine, glutamine/glutamate, glycine, proline are all essential amino acids for LB's growth.

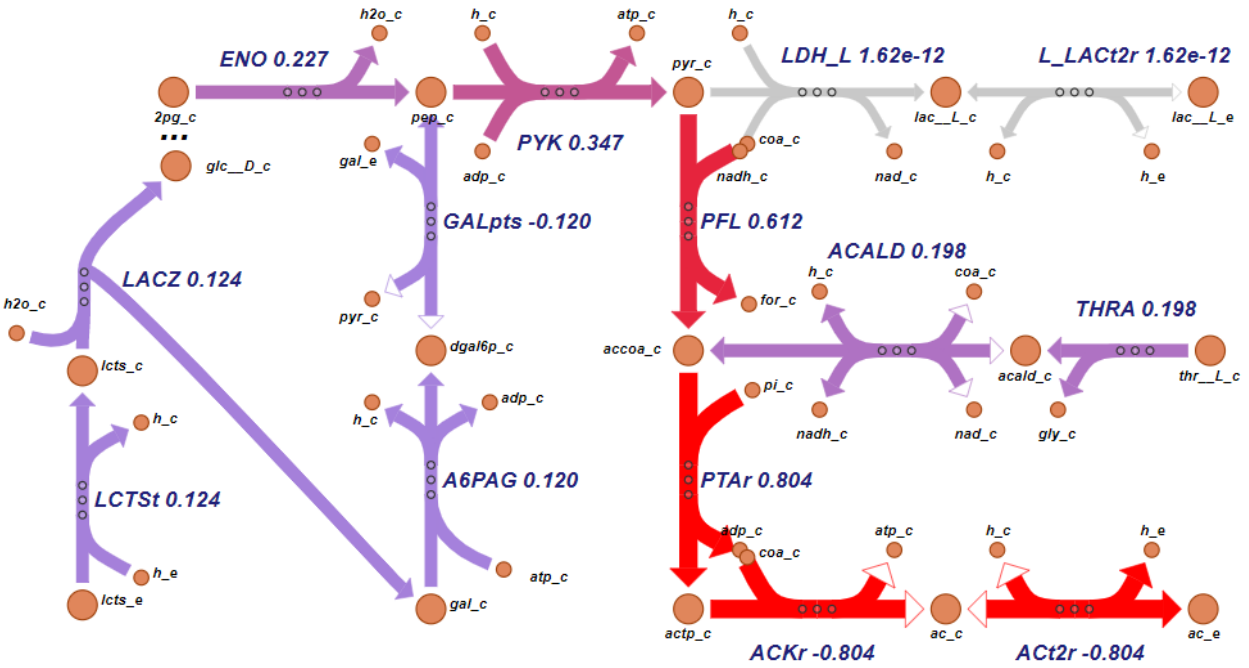

Figure A.6. Predicted metabolic fluxes of ST when lactose concentration is at 0.01mM, growth rate is 0.009 1/hr. The flux through the biosynthesis of acetic acid is from the metabolism of the glucose moiety of lactose, threonine and phosphoenolpyruvate formed in galactose exportation via phosphotransferase system (PTS).

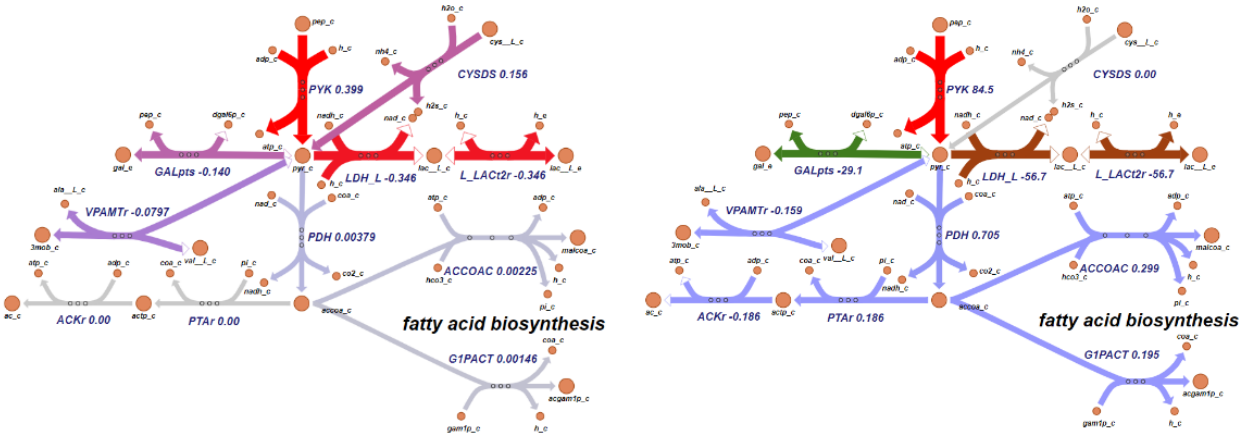

Figure A.7. Predicted metabolic fluxes of LB at low (**left**) and abundant (**right**) lactose concentrations in MPL medium. When lactose concentration is at 0.01mM, the growth rate is predicted to be 0.005 hr^{-1}. The metabolic flux from acetyl-CoA all goes into fatty acid biosynthesis, leaving the flux through acetic acid production to be 0. When lactose

concentration is large enough to saturate the transporter, the growth rate is 0.7 hr^{-1}. The biosynthesis of fatty acid is satisfied, and surplus carbon flux goes into acetic acid production.

References

1. Pastink MI, Teusink B, Hols P, Visser S, de Vos WM, Hugenholtz J. Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria. Appl Environ Microbiol. 2009;75: 3627–3633.

2. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol. 2017;35: 81–89.

3. Rau MH, Gaspar P, Jensen ML, Geppel A, Neves AR, Zeidan AA. Genome-Scale Metabolic Modeling Combined with Transcriptome Profiling Provides Mechanistic Understanding of Streptococcus thermophilus CH8 Metabolism. Appl Environ Microbiol. 2022;88: e0078022.

4. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem. 2006;281: 40041–40048.

5. Turner KW, Martley FG. Galactose fermentation and classification of thermophilic lactobacilli. Appl Environ Microbiol. 1983;45: 1932–1934.

6. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7: 74.

7. Chervaux C, Ehrlich SD, Maguin E. Physiological study of Lactobacillus delbrueckii

subsp. bulgaricus strains in a novel chemically defined medium. Appl Environ Microbiol. 2000;66: 5306–5311.

8. Gould IA, Frantz RS. Formic acid content of milk heated to high temperatures as determined by the distillation procedure. J Dairy Sci. 1946;29: 27–31.

9. Oliveira RP de S, Torres BR, Perego P, Oliveira MN de, Converti A. Co-metabolic models of Streptococcus thermophilus in co-culture with Lactobacillus bulgaricus or Lactobacillus acidophilus. Biochem Eng J. 2012;62: 62–69.

10. Özcan E, Seven M, Şirin B, Çakır T, Nikerel E, Teusink B, et al. Dynamic co-culture metabolic models reveal the fermentation dynamics, metabolic capacities and interplays of cheese starter cultures. Biotechnol Bioeng. 2021;118: 223–237.

11. Geertsma ER, Duurkens RH, Poolman B. The activity of the lactose transporter from Streptococcus thermophilus is increased by phosphorylated IIA and the action of beta-galactosidase. Biochemistry. 2005;44: 15889–15897.

12. Schumacher R. Metabolic trade-offs arising from increased free energy conservation in Saccharomyces cerevisiae. Delft University of Technology. 2018. doi:10.4233/UUID:177E9F4C-F847-436D-9FD4-9ED97BA709D9

13. Smart J, Richardson B. Molecular properties and sensitivity to cations of β-Galactosidase from Streptococcus thermophilus with four enzyme substrates. Appl Microbiol Biotechnol. 1987;26: 177–185.

14. Reinhardt LA, Thoden JB, Peters GS, Holden HM, Cleland WW. pH-rate profiles support a general base mechanism for galactokinase (Lactococcus lactis). FEBS Lett. 2013;587: 2876–2881.

15. Meyer D, Schneider-Fresenius C, Horlacher R, Peist R, Boos W. Molecular characterization of glucokinase from Escherichia coli K-12. J Bacteriol. 1997;179: 1298–1306.

16. Gao H, Chen Y, Leary JA. Kinetic measurements of phosphoglucose isomerase and phosphomannose isomerase by direct analysis of phosphorylated aldose–ketose isomers using tandem mass spectrometry. Int J Mass Spectrom. 2005;240: 291–299.

17. Kotlarz D, Buc H. Phosphofructokinases from Escherichia coli. Methods Enzymol. 1982;90 Pt E: 60–70.

18. Baldwin SA, Perham RN, Stribling D. Purification and characterization of the class-II D-fructose 1,6-bisphosphate aldolase from Escherichia coli (Crookes' strain). Biochem J. 1978;169: 633–641.

19. Garza-Ramos G, Pérez-Montfort R, Rojo-Domínguez A, de Gómez-Puyou MT, Gómez-Puyou A. Species-specific inhibition of homologous enzymes by modification of nonconserved amino acids residues. The cysteine residues of triosephosphate isomerase. Eur J Biochem. 1996;241: 114–120.

20. D'Alessio G, Josse J. Glyceraldehyde phosphate dehydrogenase of Escherichia coli. Structural and catalytic properties. J Biol Chem. 1971;246: 4326–4333.

21. Fifis T, Scopes RK. Purification of 3-phosphoglycerate kinase from diverse sources by affinity elution chromatography. Biochem J. 1978;175: 311–319.

22. D'Alession G, Josse J. Phosphoglycerate kinase and phosphoglyceromutase from Escherichia coli. Methods Enzymol. 1975;42: 139–144.

23. Kühnel K, Luisi BF. Crystal structure of the Escherichia coli RNA degradosome

component enolase. J Mol Biol. 2001;313: 583–592.

24. Waygood EB, Sanwal BD. The Control of Pyruvate Kinases of Escherichia coli: I. PHYSICOCHEMICAL AND REGULATORY PROPERTIES OF THE ENZYME ACTIVATED BY FRUCTOSE 1,6-DIPHOSPHATE. J Biol Chem. 1974;249: 265–274.

25. Hensel R, Mayr U, Fujiki H, Kandler O. Comparative studies of lactate dehydrogenases in lactic acid bacteria. Amino-acid composition of an active-site region and chemical properties of the L-lactate dehydrogenase of Lactobacillus casei, Lactobacillus curvatus, Lactobacillus plantarum, and Lactobacillus acidophilus. Eur J Biochem. 1977;80: 83–92.

26. Conradt H, Hohmann-Berger M, Hohmann HP, Blaschkowski HP, Knappe J. Pyruvate formate-lyase (inactive form) and pyruvate formate-lyase activating enzyme of Escherichia coli: isolation and structural properties. Arch Biochem Biophys. 1984;228: 133–142.

27. Meinecke B, Bertram J, Gottschalk G. Purification and characterization of the pyruvate-ferredoxin oxidoreductase from Clostridium acetobutylicum. Arch Microbiol. 1989;152: 244–250.

28. Bergmeyer HU, Holz G, Klotzsch H, Lang G. Phosphotransacetylase from Clostridium kluyveri. Culture of the bacterium, isolation, crystallization and properties of the enzyme. Biochem Z. 1963;338: 114–121.

29. Winzer K, Lorenz K, DÜrre P. Acetate kinase from Clostridium acetobutylicum: a highly specific enzyme that is actively transcribed during acidogenesis and solventogenesis. Microbiology. 1997;143 (Pt 10): 3279–3286.

30. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM,

Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

31. Poolman B, Knol J, Lolkema JS. Kinetic Analysis of Lactose and Proton Coupling in Glu379 Mutants of the Lactose Transport Protein of Streptococcus thermophilus∗. J Biol Chem. 1995;270: 12995–13003.

32. Kaiser JC, Omer S, Sheldon JR, Welch I, Heinrichs DE. Role of BrnQ1 and BrnQ2 in branched-chain amino acid transport and virulence in Staphylococcus aureus. Infect Immun. 2015;83: 1019–1029.

33. Özcan E, Selvi SS, Nikerel E, Teusink B, Toksoy Öner E, Çakır T. A genome-scale metabolic network of the aroma bacterium Leuconostoc mesenteroides subsp. cremoris. Appl Microbiol Biotechnol. 2019;103: 3153–3165.

34. Schmidt JDR, Beitz E. Mutational widening of constrictions in a formate–nitrite/H+ transporter enables aquaporin-like water permeability and proton conductance. J Biol Chem. 2022;298. doi:10.1016/j.jbc.2021.101513

Appendix B

B.1 Supplementary methods

B.1.1 Software and code availability

All scripts were written in python. The deep learning model was implemented using PyTorch v1.7.1. The computer used in this work was a Dell Latitude Laptop with intel core i7 CPU. The model was trained with GPU RTX8000 provided by Advanced Research Computing (ARC) service in the University of Oxford [1]. Figures were edited using InkScape (https://inkscape.org/). The code and data used to generate results of this paper are available at https://github.com/SizheQiu/DLTKcat.

B.1.2 Deep learning model evaluation

The original dataset was randomly split into the test and the train dataset with a ratio of 1:9. The test dataset was held out to examine the accuracy of the model. Before the training, 10% of the train dataset was randomly split as the validation dataset (also called dev set). During the training process, R2 (Eq. B.1) and RMSE (Eq. B.2) scores of k_{cat} predictions were computed at each epoch for the test and validation datasets.

$$R^2 = \frac{\sum_{i=1}^{n}(y_{ie} - y_{ip})^2}{\sum_{i=1}^{n}(y_{ie} - \bar{y})^2} \quad (Eq.B.1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ie} - y_{ip})^2} \quad (Eq.B.2)$$

For distributions of prediction errors for a certain sub-dataset, this study randomly selected 50 data entries for 200 times to obtain the distribution of RMSE (Eq. B.2), R2 (Eq. B.1) and mean absolute error (MAE, Eq. B.3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{ie} - y_{ip}\right| \ (Eq.B.3)$$

B.1.3 Hyperparameter optimization for latent dimension and number of dense layers

To find the optimal combinations of the latent dimension, the dimension of protein and compound features in bi-directional attention process, and the number of dense layers after feature concatenation, training processes were performed for latent dimension = 40, 64 and number of dense layers = 3, 4, 5, 6. The latent dimension and number of dense layers that achieved the lowest RMSE score on the dev set were used to generate the trained deep learning model (**Figure B.4**). The other hyperparameters were reported in the main text.

B.1.4 Assessment of the feature importance of temperature by random shuffling

1026 entries in the curated dataset had the same protein sequence and substrate as other entries but different temperatures. To assess the feature importance of temperature, they were extracted to obtain R2 (Eq. B.1), RMSE (Eq. B.2) and MAE (Eq. B.3) scores of k_{cat} predictions with shuffled and unshuffled temperature related features. The feature shuffling was performed by randomly sampling from the original list of feature values, so that the shuffled and unshuffled features still had the same distribution.

B.2 Tables

Table B.1. Metabolic reaction information of *Lactococcus lactis MG1363*

ID	name	EC number
GLCpts	D-glucose transport via PEP:Pyr phosphotransferase system	_
G6PDH	Glucose-6-phosphate dehydrogenase	1.1.1.49
PGL	6-phosphogluconolactonase	3.1.1.31
PGDH	6-phosphogluconate dehydrogenase	1.1.1.351, 1.1.1.44
GALU	UTP-glucose-1-phosphate uridylyltransferase	2.7.7.9
PGMT	Phosphoglucomutase	5.4.2.2, 5.4.2.5
UDPG4E	UDPglucose 4-epimerase	5.1.3.2
PGI	Glucose-6-phosphate isomerase	5.3.1.9
PFK	Phosphofructokinase	2.7.1.11
FBA	Fructose-bisphosphate aldolase	4.1.2.13
TPI	Triose-phosphate isomerase	5.3.1.1
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12
PGK	Phosphoglycerate kinase	2.7.2.3
PGM	Phosphoglycerate mutase	5.4.2.11
ENO	Enolase	4.2.1.11
PYK	Pyruvate kinase	2.7.1.40
LDH	Lactate dehydrogenase	1.1.1.27
PFL	Pyruvate formate lyase	2.3.1.54
PDH	Pyruvate dehydrogenase	1.2.7.1
PTAr	Phosphotransacetylase	2.3.1.8
ACKr	Acetate kinase	2.7.2.1

ACCOAC	Acetyl-CoA carboxylase	6.4.1.2
MACPMT	Malonyl CoAacyl carrier protein S malonyltransferase	2.3.1.39

Table B.2. Metabolic reaction information of *Streptococcus thermophilus LMG18311*

ID	name	EC number
LCTSGALex	Lactose galactose exchange via antiporter	_
LACZ	Beta-galactosidase	3.2.1.23
GALK	Galactokinase	2.7.1.6
HEX	Hexokinase (D-glucose:ATP)	2.7.1.2
GALK	Galactokinase	2.7.1.6
GALM	Aldose 1-epimerase	5.1.3.3
GALT	Galactose 1 phosphate uridylyltransferase	2.7.7.10
GALU	UTP-glucose-1-phosphate uridylyltransferase	2.7.7.9
PGMT	Phosphoglucomutase	5.4.2.2, 5.4.2.5
UDPG4E	UDPglucose 4-epimerase	5.1.3.2
PGI	Glucose-6-phosphate isomerase	5.3.1.9
PFK	Phosphofructokinase	2.7.1.11
FBA	Fructose-bisphosphate aldolase	4.1.2.13
TPI	Triose-phosphate isomerase	5.3.1.1
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12
PGK	Phosphoglycerate kinase	2.7.2.3
PGM	Phosphoglycerate mutase	5.4.2.11
ENO	Enolase	4.2.1.11
PYK	Pyruvate kinase	2.7.1.40
LDH	Lactate dehydrogenase	1.1.1.27
PFL	Pyruvate formate lyase	2.3.1.54

PDH	Pyruvate dehydrogenase	1.2.7.1
PTAr	Phosphotransacetylase	2.3.1.8
ACKr	Acetate kinase	2.7.2.1
ACCOAC	Acetyl-CoA carboxylase	6.4.1.2

* Reaction IDs, names, and enzyme EC numbers are all obtained from the BIGG database (http://bigg.ucsd.edu/) [2].

B.3 Figures

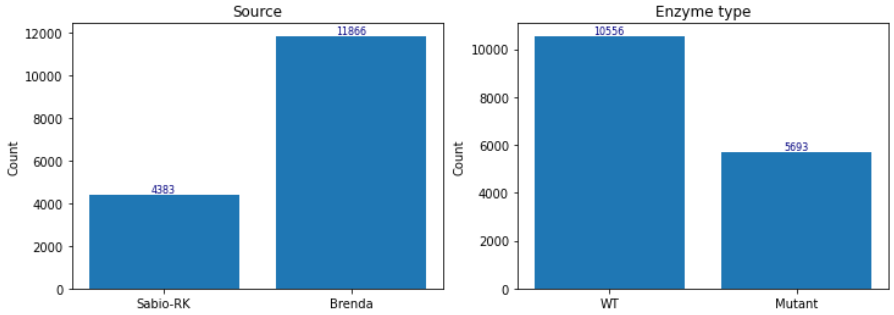

Figure B.1. Statistics of collected data from enzyme databases. Left: 4383 entries are from Sabio-RK, 11866 entries are from Brenda. Right: enzymes in 10556 entries are wild types, in 5693 entries are mutants.

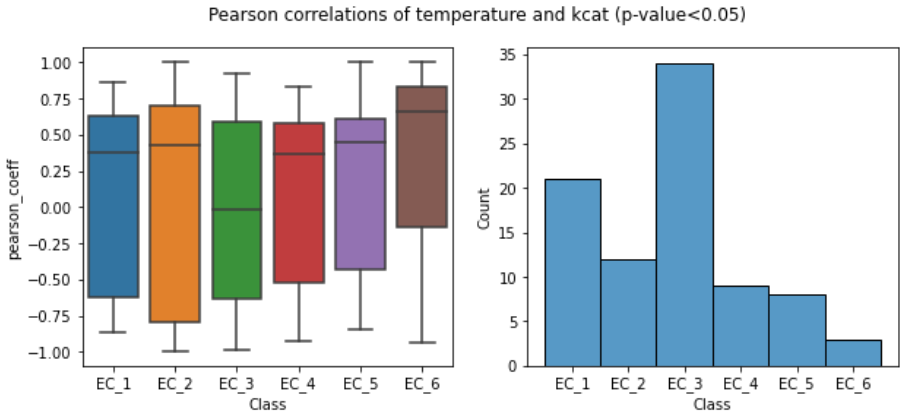

Figure B.2. Significant Pearson correlations of k_{cat} of 87 enzyme classes (EC numbers) covering 2430 entries and temperature.

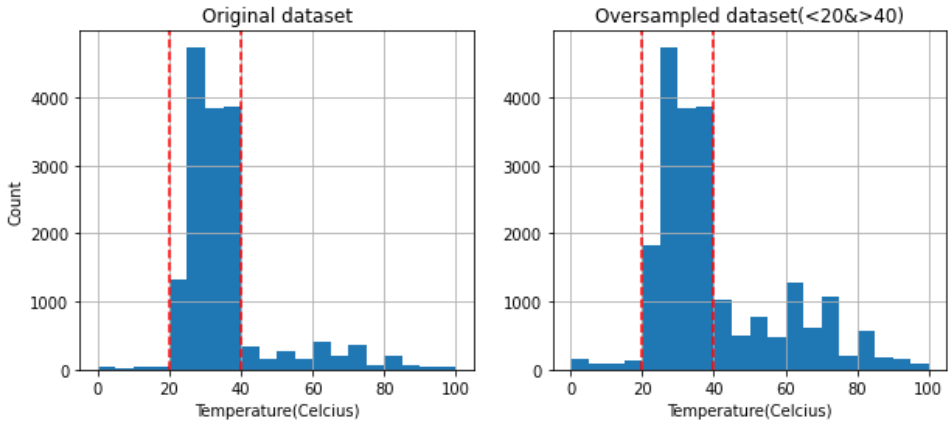

Figure B.3. Training datasets before and after oversampling of entries at low and high temperature ranges.

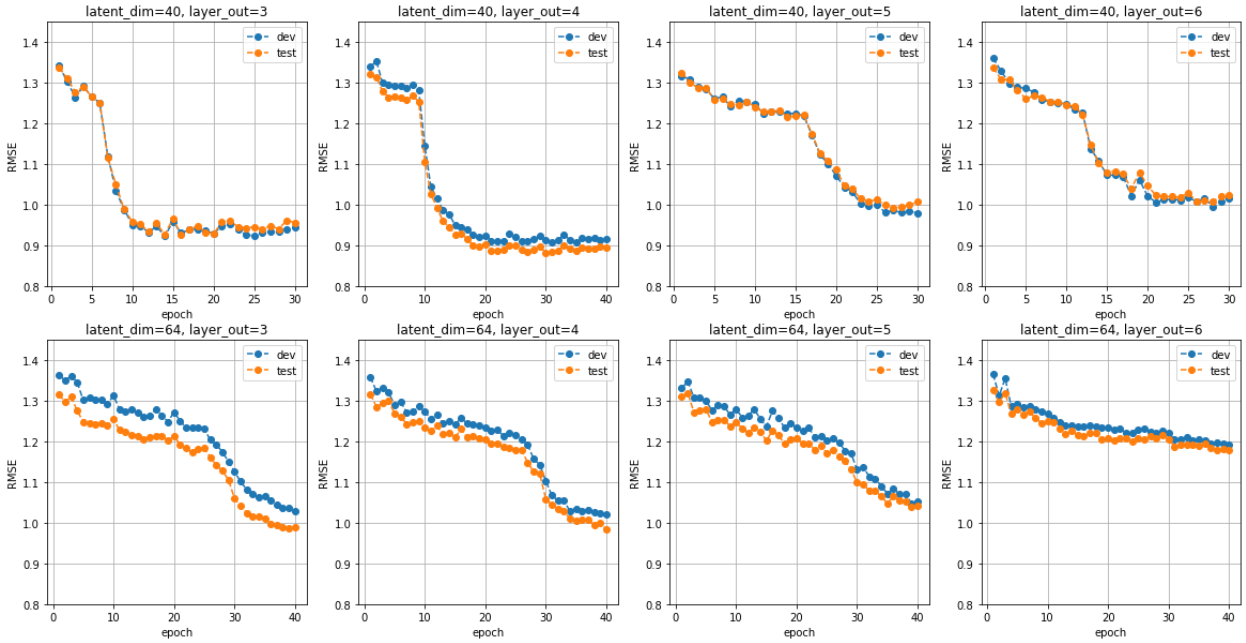

Figure B.4. Results of hyperparameter optimization for latent dimension (latent_dim) and number of dense layers (layer_out). The trial with latent_dim=40 and layer_out=4 achieved the lowest RMSE. test: the test set; dev: the validation set (the development set).

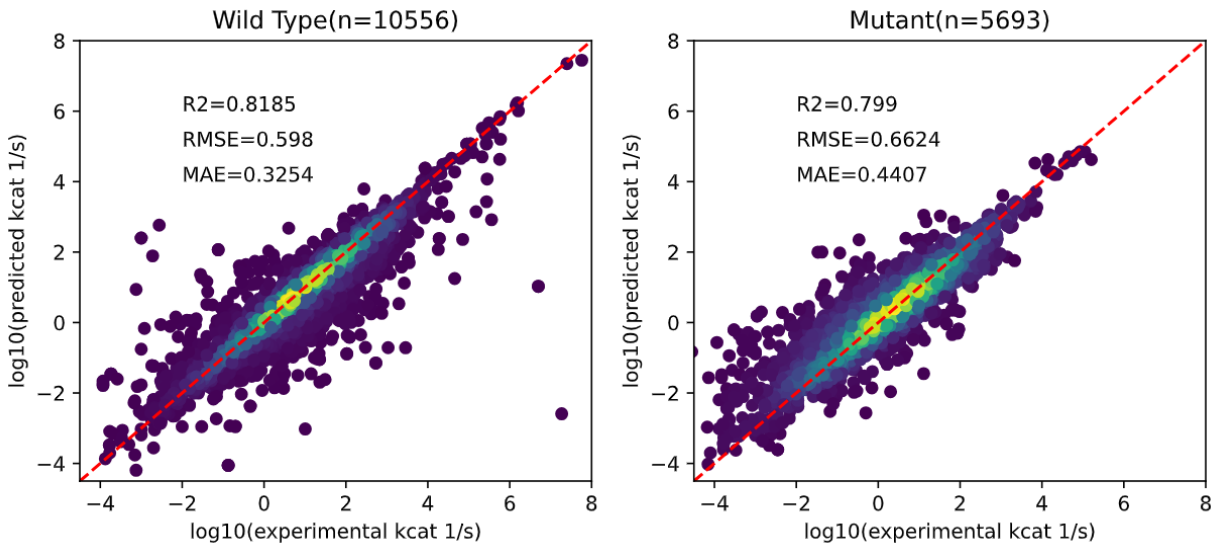

Figure B.5. R2, RMSE and MAE scores of $log_{10}(k_{cat})$ prediction for WT and mutated enzymes.

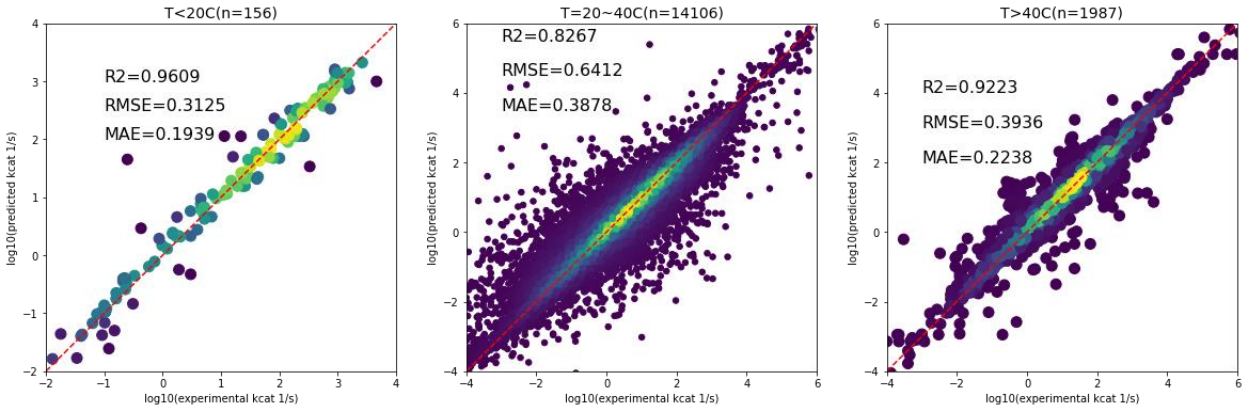

Figure B.6. RMSE, R2, MAE scores of predicted $log_{10}(k_{cat})$ values at low (<20℃), middle (20~40℃) and high (>40℃) temperature ranges.

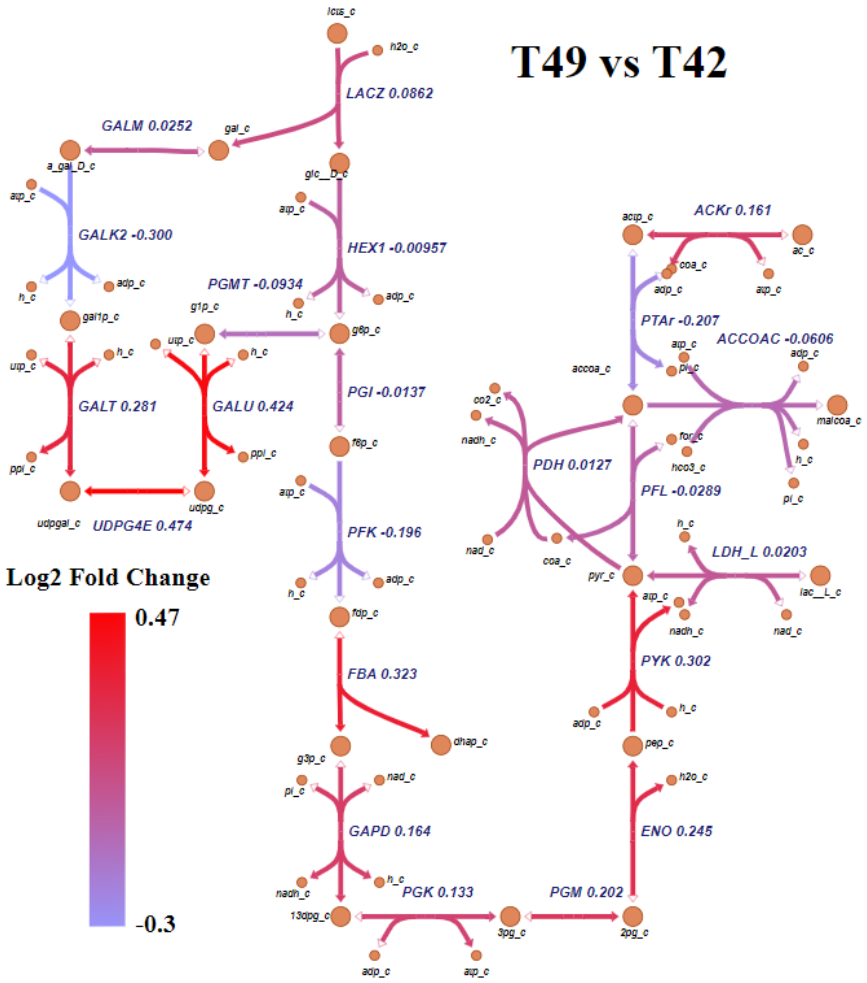

Figure B.7. Log2-fold change of predicted k_{cat} values for ST at 49 ℃ and 42 ℃ (49 ℃ vs 42 ℃).

References

1. Richards A. University of Oxford Advanced Research Computing. Zenodo; 2015. doi:10.5281/zenodo.22558

2. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22.

Appendix C

C.1 Supplementary methods

C.1.1 Functional annotation of iModulons

Coding genes in the genome of Lactobacillus plantarum were annotated using eggNOG-mapper [1], and COG (Clusters of Orthologous Groups), KEGG [2], CAZy [3] and BiGG [4] annotations were obtained. The E-value threshold was 1e-5. The thresholds of query and subject coverages were all set as 50%. The functional annotation of each iModulon was determined based on enrichment analysis of cellular pathways.

C.1.2 Metabolic pathway visualization

Metabolic pathways encoded by iModudon member genes were visualized using Escher (https://escher.github.io/) [5]. The genome-scale metabolic model iBT721 [6] was used as the database of gene-reaction rules to query for metabolic reactions encoded by member genes of iModudons.

C.2 Tables

Table C.1. Metadata of sample conditions

Sample	Condition	Cultivation time	Reference
wt_pH6.2	MRS medium, 10 μg/L valinomycin, pH adjusted to 6.2 with addition of lactic acid	12 hours	[7]
wt_pH5.5	MRS medium, 10 μg/L valinomycin, pH adjusted to 5.5 with addition of lactic acid	12 hours	[7]

wt_pH5.0	MRS medium, 10 µg/L valinomycin, pH adjusted to 5.0 with addition of lactic acid	12 hours	[7]
wt_3OCp1	MRS medium, 100 µM 3OC12, 37℃, +1 hour	_	[8]
wt_3OCp4	MRS medium, 100 µM 3OC12, 37℃, +4 hour	_	[8]
wt_3OCp7	MRS medium, 100 µM 3OC12, 37℃, +7 hour	_	[8]
wt_3OCm1	MRS medium, 0.1% DMSO (v/v), 37℃, +7 hour	_	[8]
wt_3OCm4	MRS medium, 0.1% DMSO (v/v), 37℃, +7 hour	_	[8]
wt_3OCm7	MRS medium, 0.1% DMSO (v/v), 37℃, +7 hour	_	[8]
WCFS1_MRS	MRS broth, pH = 5.7, 30℃	24 hours	[9]
LB16_MRS	MRS broth, pH = 5.7, 30℃	24 hours	[9]
BEE	Homogenized bee:water of 3:1 (v/v), pH = 4.7, 30℃	24 hours	[9]
CB	Cheese broth prepared as in Neviani *et al.*, 2009 [10], pH = 6.1, 30℃	24 hours	[9]
DE	Homogenized larvae:water of 1:1 (wt/wt), pH = 6.5, 30℃	24 hours	[9]
FE	Fecal extract, pH = 5.9, 30℃	24 hours	[9]
OE	Homogenized olive extract, pH = 4.7, 30℃	24 hours	[9]
PJ	Pineapple juice medium treated as in Filannino *et al.*, 2014 [11], pH = 3.5, 30℃	24 hours	[9]
TJ	Tomato juice medium treated as in Filannino *et al.*, 2014 [11], pH = 4.7, 30℃	24 hours	[9]
WFH	Wheat flour hydrolyzate treated as in Gobbetti *et al.*, 1994 [12], pH = 5.6, 30℃	24 hours	[9]
wt_GLC	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) glucose, 37℃	48 hours	[13]
wt_HMO	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) human milk oligosaccharides, 37℃	48 hours	[13]
wt_PAC1	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) proanthocyanidin fraction 1, 37℃	48 hours	[13]

wt_PAC2	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) proanthocyanidin fraction 2, 37℃	48 hours	[13]
wt_HMOPAC1	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) human milk oligosaccharides + proanthocyanidin fraction 1, 37℃	48 hours	[13]
wt_HMOPAC2	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) human milk oligosaccharides + proanthocyanidin fraction 1, 37℃	48 hours	[13]
wt_FOSPAC1	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) fructooligosaccharides + proanthocyanidin fraction 1, 37℃	48 hours	[13]
wt_FOSPAC2	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) fructooligosaccharides + proanthocyanidin fraction 2, 37℃	48 hours	[13]
wt_XG	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) xyloglucans, 37℃	48 hours	[13]
wt_XGPAC2	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) xyloglucans + proanthocyanidin fraction 1, 37℃	48 hours	[13]
wt_XGPAC2	MRS medium, 0.05% (w/v) L-cysteine HCl, 1% (w/v) xyloglucans + proanthocyanidin fraction 2, 37℃	48 hours	[13]

Table C.2. Metabolic reaction information

ID	iModulon	Genes	Name
SERAT	McbR	lp_0254	Serine O-acetyltransferase
CYSS	McbR	lp_0265	Cysteine synthase
CYSTL	McbR	lp_0255	Cystathionine-b-lyase
CYSTGL	McbR	lp_0255	Cystathionine-b-lyase
ASPCT	PyrR	lp_2703	Aspartate carbamoyltransferase
DHORTS	PyrR	lp_2702	Dihydroorotase
DHORD6	PyrR	lp_2699	Dihydoorotic acid dehydrogenase
ORPT	_	lp_2697	Orotate phosphoribosyltransferase
OMPDC	PyrR	lp_2698	Orotidine 5 phosphate decarboxylase
CBPS	PyrR, ArgR/MleR	lp_2700/2701, lp_0526/0527	Carbamoyl-phosphate synthase (glutamine-hydrolysing)
ORNTAC	ArgR/MleR	lp_0529	Ornithine transacetylase
ACOTA	ArgR/MleR	lp_0531	Acetylornithine transaminase
ACGK	ArgR/MleR	lp_0530	Acetylglutamate kinase
AGPR	ArgR/MleR	lp_0528	N-acetyl-g-glutamyl-phosphate reductase
OCBT	ArgR/MleR	lp_0532	Ornithine carbamoyltransferase
MALt2r	ArgR/MleR	lp_1119	L-malate reversible transport via proton symport
MALLAC	ArgR/MleR	lp_1118	Malolactic enzyme
PYROX	CcpA	lp_2629	Pyruvate oxidase
PDH	CcpA	lp_2151-2154	Pyruvate dehydrogenase
PFL	CcpA	lp_3313/3314	Formate C acetyltransferase
GLYCt1	CcpA	lp_0372	Glycerol transport via channel
GLYK	CcpA	lp_0370	Glycerol kinase
G3PO	CcpA	lp_0371	Glycerol 3 phosphate oxidase
NDPK	CcpA	lp_0242	Nucleoside diphosphate kinase

G1PTMT	GntR	lp_1186	Glucose 1 phosphate thymidylyltransferase
TDPGDH	GntR	lp_1189	DTDPglucose 4,6-dehydratase
TDPDRE	GntR	lp_1188	DTDP-4-dehydrorhamnose 3,5-epimerase
TDPDRR	GntR	lp_1190	DTDP-4-dehydrorhamnose reductase
CPSS	GntR	lp_1177-1185	Capsular polysaccharide synthase complex
MANpts	GntR	lp_0576/0577	D-mannose transport via PEP:Pyr PTS
ASP1DC	GntR	lp_0579	Aspartate 1-decarboxylase
DB4PS	GntR	lp_1437	3,4-Dihydroxy-2-butanone-4-phosphate synthase
DHDPS	GntR	lp_2685	Dihydrodipicolinate synthase
MELIBt2	GalR/AraR	lp_3486	Melibiose transport in via symport
RAFGH	GalR/AraR	lp_3485	Raffinose galactohydrolase
GALS3	GalR/AraR	lp_3485	A-galactosidase (melibiose)
BGLA1	GalR/AraR	lp_0440	6-phospho-beta-glucosidase
GALM	_	lp_0826/1731/3487	Aldose 1 epimerase
GALK2	GalR/AraR	lp_3482	Galactokinase
GALT	GalR/AraR	lp_3480	Galactose 1 phosphate uridylyltransferase
UDPG4E	GalR/AraR	lp_3481	UDPglucose 4 epimerase
SERt2r	GalR/AraR	lp_0502	L-serine reversible transport via proton symport
SERD_L	GalR/AraR	lp_0505/0506	L-serine deaminase
TKT1	GalR/AraR	lp_3538	Transketolase
TKT2	GalR/AraR	lp_3538	Transketolase
TALA	GalR/AraR	lp_3539	Transaldolase
RBP4E	GalR/AraR	lp_3555	L-ribulose-phosphate 4-epimerase
RBK_L1	GalR/AraR	lp_3556	L-ribulokinase (L-ribulose)
ARAI	GalR/AraR	lp_3554	L-arabinose isomerase

ARAB_Lt	GalR/AraR	lp_3557	L-arabinose extracellular transport
RBLK2	GalR/AraR	lp_3556	L-ribulokinase (ribitol)

C.3 Figures

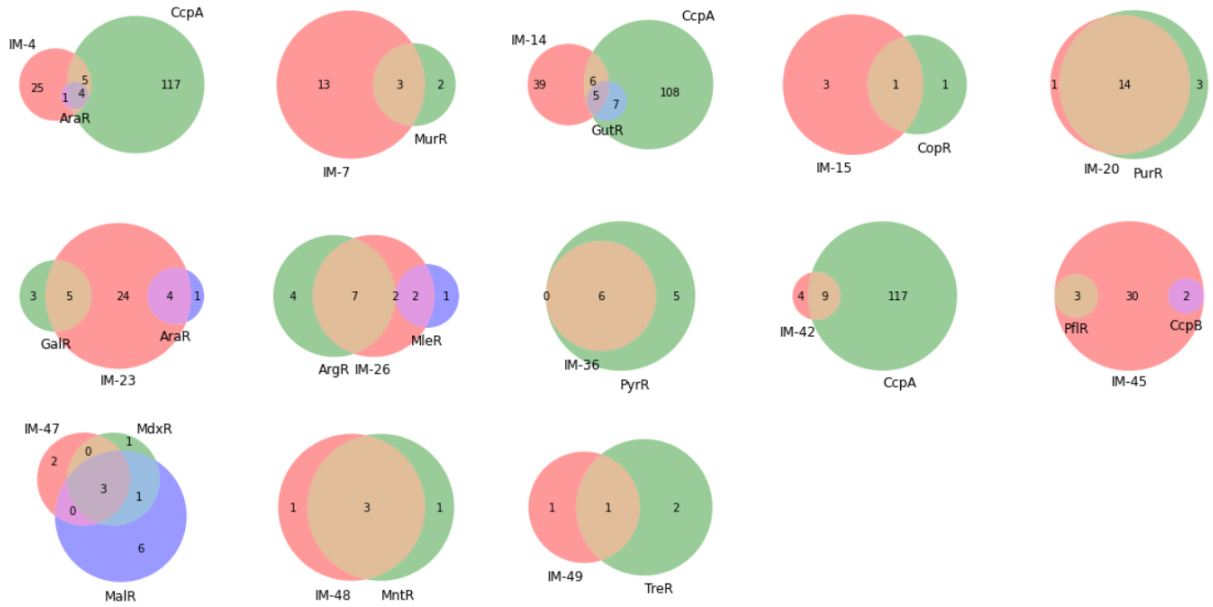

Figure C.1. Venn diagrams of regulon enrichment for IM-4, 7, 14, 15, 20, 23, 26, 36, 42, 45, 47, 48, 49.

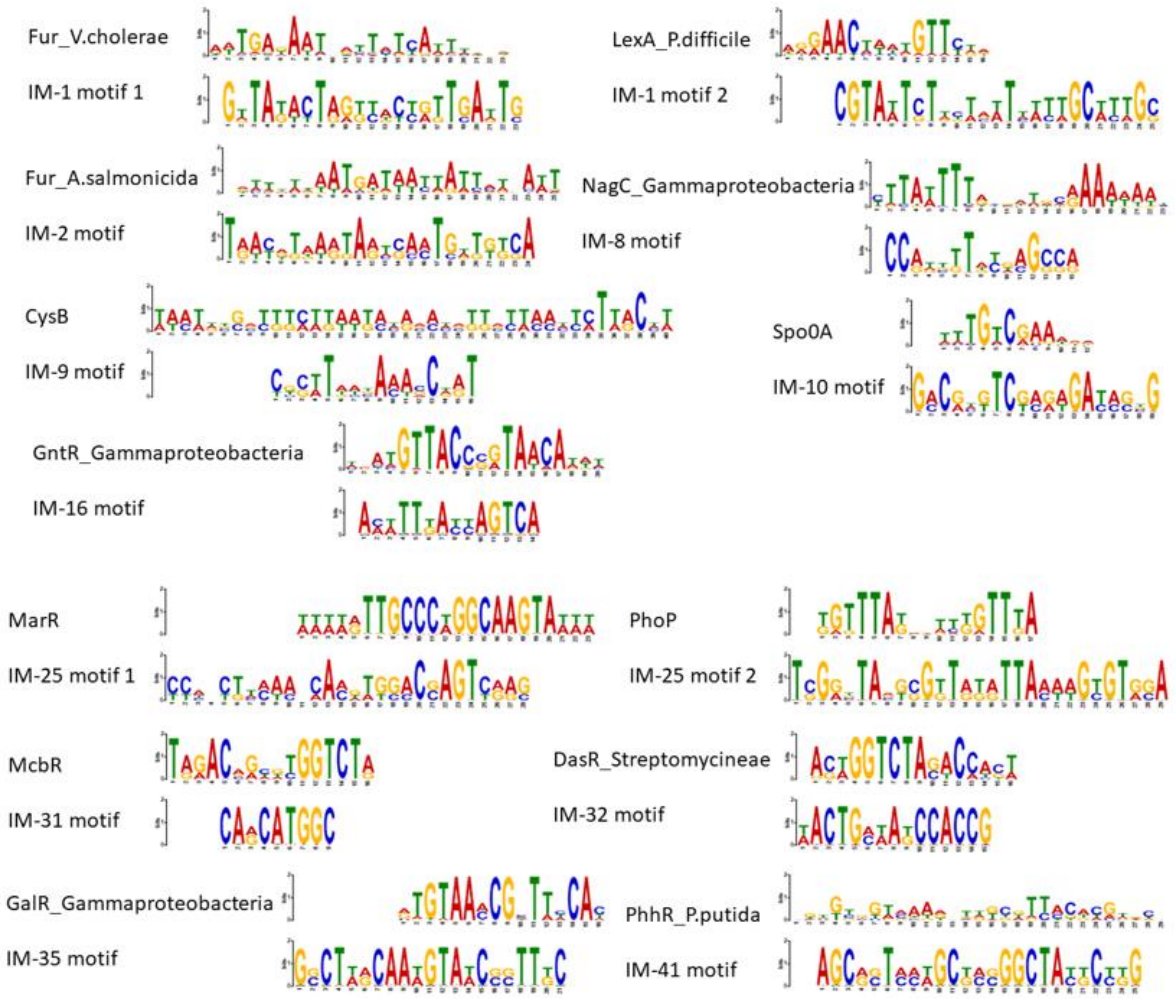

Figure C.2. Motif comparison logo plots for IM-1, 2, 8, 9, 10, 16, 25, 31, 32, 35, 41.

References

1. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38: 5825–5829.

2. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–62.

3. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.

Nucleic Acids Res. 2009;37: D233–8.

4. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22.

5. King ZA, Dräger A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. PLoS Comput Biol. 2015;11: e1004321.

6. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem. 2006;281: 40041–40048.

7. Jung S, Lee J-H. Characterization of transcriptional response of Lactobacillus plantarum under acidic conditions provides insight into bacterial adaptation in fermentative environments. Sci Rep. 2020;10: 19203.

8. Spangler JR, Dean SN, Leary DH, Walper SA. Response of Lactobacillus plantarum WCFS1 to the Gram-Negative Pathogen-Associated Quorum Sensing Molecule N-3-Oxododecanoyl Homoserine Lactone. Front Microbiol. 2019;10: 715.

9. Filannino P, De Angelis M, Di Cagno R, Gozzi G, Riciputi Y, Gobbetti M. How Lactobacillus plantarum shapes its transcriptome in response to contrasting habitats. Environ Microbiol. 2018;20: 3700–3716.

10. Neviani E, De Dea Lindner J, Bernini V, Gatti M. Recovery and differentiation of long ripened cheese microflora through a new cheese-based cultural medium. Food Microbiol. 2009;26: 240–245.

11. Filannino P, Cardinali G, Rizzello CG, Buchin S, De Angelis M, Gobbetti M, et al. Metabolic responses of Lactobacillus plantarum strains during fermentation and storage of vegetable and fruit juices. Appl Environ Microbiol. 2014;80: 2206–2215.

12. Gobbetti M, Corsetti A, Rossi J. The sourdough microflora. Interactions between lactic acid bacteria and yeasts: metabolism of carbohydrates. Appl Microbiol Biotechnol. 1994;41: 456–460.

13. Özcan E, Rozycki MR, Sela DA. Cranberry Proanthocyanidins and Dietary Oligosaccharides Synergistically Modulate Lactobacillus plantarum Physiology. Microorganisms. 2021;9. doi:10.3390/microorganisms9030656

Appendix D

D.1 Supplementary method

D.1.1 Genome-scale metabolic model modification

The GSMM of LP HMX2 was built via the modification of iBT721 [1]. This study assumed that the glucose uptake of LP was carried out by D-glucose transport via PEP:Pyr PTS (GLCpts), and hence, GLCt2r was deleted from the original GSMM. Phosphoketolase (PKL), Fructose 6-phosphate aldolase (F6PA) and Dihydroxyacetone phosphotransferase (DHAPT) were removed due to redundancy. The directions of irreversible reactions were all fixed. For reversible reactions (i.e., RPE, RPI, PGI, UDPG4E and MAN6PI), reverse and forward reactions were created. Because GapMind [2] found that the amino acid auxotrophy of LP WCFS1 and LP HMX2 were the same, amino acid metabolism was not altered in this GSMM.

Mannose-1-phosphate guanylyltransferase (MAN1PT) was added to the GSMM for the biosynthesis of GDP-Mannose.The EPS biosynthetic pathway was separated as 3 pseudo reactions (i.e., MANT_EPS, GLCT_EPS, and GALT_EPS) and WZX flippase, as shown below:

$$MANT\_EPS\colon 14\ GDPMannose[c] \rightarrow\ Man14[c] +\ 14\ GDP[c]$$

$$GLCT\_EPS\colon 6\ UDPGlucose[c]\ +\ Man14[c]\ \rightarrow Man14GLC6[c] +\ \ 14\ UDP[c]$$

$$GALT\_EPS\colon UDPGalactose[c]\ + Man14GLC6[c] \rightarrow EPS[c] + UDP[c]$$

$$WZX\colon EPS[c] +\ ATP[c] + H2O[c] \rightarrow EPS[e] + ADP[c] + Pi[c] + H[c]$$

$Man14[c]$ and $Man14GLC6[c]$ were pseudo intermediates of the biosynthesis of the pseudo EPS repeating unit, $EPS[c]$. The chemical formulas of $Man14[c]$, $Man14GLC6[c]$ and $EPS[c]$ were C84H154O70, C120H220O100 and C126H231O105, respectively. $[c]$ represented the cytoplasm, and $[e]$ represented the extracellular space.

D.1.2 Estimation of essential parameters for proteome constrained FBA

For missing enzyme activities (a_i), DLTKcat [3] was used to predict k_{cat} for the reaction first, and then the k_{cat}(1/s) was converted to a_i (mmol/hr*g E) using Eq. S1. The input variables of DLTKcat were substrate SMILES strings, protein sequences and the constant growth temperature, 37 ℃. The detailed user guide of DLTKcat can be found at https://github.com/SizheQiu/DLTKcat.

$$v_{max} = a_i[E_i] = k_{cat}\frac{[E_i]}{MW_i} \Rightarrow a_i = \frac{k_{cat}}{MW_i} \ (Eq.S1)$$

The activities of PMANM, RPI (reverse) and RPE (reverse) were predicted in this study. For RPI (reverse) and RPE (reverse), the substrates were D-Ribulose 5-phosphate (ru5p__D) and D-Xylulose 5-phosphate (xu5p__D), respectively.

The mathematical functions of F_{pH} were estimated based on approximated relative activities of enzymes in primary metabolism and EPS biosynthesis (**Figure S2**). In Eq. 8, $c_1 = 1.3812, c_2 = 4.3315$. In Eq. 9, $c_3 = -0.3815, c_4 = 4.2847, c_5 = -11.0359$. The average relative enzyme activities for primary metabolism were approximated using normalized growth rates at different pH values. To approximate the relative enzyme activities for EPS biosynthesis, measured EPS production fluxes were divided by total expressions of EPS biosynthetic proteins, and the results were normalized. The enzyme activities of glycosyltransferases (MANT_EPS, GLCT_EPS and GALT_EPS) were estimated based on the maximum EPS production flux at pH 5.5 without the inhibition of undissociated lactic acid (**Figure S3**).

In the simulation, the upper bounds of amino acid uptake fluxes were all set as 0.2 $\frac{mmol}{gDW*hr}$, approximated based on the concentrations of 19 essential amino acids measured during the fermentation (**Figure S4**). As the MRS medium is a rich medium, this study did not set constraints on the uptake fluxes of purines, pyrimidines, metal ions, and vitamins.

D.2 Tables

Table D.1. Molecular structural properties of LP-HMX2-EPS

Monosaccharide	Mass fraction (%)	Glycosyl linkage	Molar fraction (%)
Man	44.69666	t-Manp	36.05515
Glc	40.90727	1,2,6-Manp	15.47439
Gal	8.579008	1,2-Manp	14.88804
GlcN	5.602263	t-Glcp	13.86694
GalA	0.17621	1,4-Galp	4.621861
Ara	0.038584	1,4-Glcp	4.404261
_	_	1,6-Glcp	3.758698
_	_	1,3-Glcp	3.676406
_	_	t-Galp	1.150325
_	_	t-Fucp	0.560417
_	_	1,3,4-GalAp	0.452948
_	_	1,6-Galp	0.404385
_	_	1,3,6-Glcp	0.365454
_	_	1,3,6-Manp	0.320716

Table D.2. Metabolic enzyme information

ID	Name	EC number
GLCpts	D-glucose transport via PEP:Pyr PTS	_
MANpts	D-mannose transport via PEP:Pyr PTS	_
LCTSt	Lactose transport via proton symport	_
LACZ	Beta-galactosidase	3.2.1.23
GALK	Galactokinase	2.7.1.6
GALT	Galactose 1 phosphate uridylyltransferase	2.7.7.10
HEX	Hexokinase (D-glucose:ATP)	2.7.1.2
PGI	Glucose-6-phosphate isomerase	5.3.1.9
PFK	Phosphofructokinase	2.7.1.11
FDPA	Fructose-bisphosphate aldolase	4.1.2.13
TPI	Triose-phosphate isomerase	5.3.1.1
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12
PGK	Phosphoglycerate kinase	2.7.2.3
PGM	Phosphoglycerate mutase	5.4.2.11
ENO	Enolase	4.2.1.11
PYK	Pyruvate kinase	2.7.1.40
LDH	Lactate dehydrogenase	1.1.1.27
PFL	Pyruvate formate lyase	2.3.1.54
PDH	Pyruvate dehydrogenase	1.2.7.1
PTAr	Phosphotransacetylase	2.3.1.8
ACKr	Acetate kinase	2.7.2.1
ACCOAC	Acetyl-CoA carboxylase	6.4.1.2
MAN6PI	Mannose-6-phosphate isomerase	5.3.1.8
PMANM	Phosphomannomutase	5.4.2.8
MAN1PT	Mannose 1 phosphate guanylyltransferase	2.7.7.13

PGMT	Phosphoglucomutase	5.4.2.2
GALU	UTP-glucose-1-phosphate uridylyltransferase	2.7.7.9
UDPG4E	UDP-glucose 4-epimerase	5.1.3.2
G6PDH	Glucose 6-phosphate dehydrogenase	1.1.1.49
PGL	6-phosphogluconolactonase	3.1.1.31
GND	Phosphogluconate dehydrogenase	1.1.1.44
RPE	Ribulose 5-phosphate 3-epimerase	5.1.3.1
RPI	Ribose-5-phosphate isomerase	5.3.1.6
TKT1/2	Transketolase	2.2.1.1
TALA	Transaldolase	2.2.1.2

* FBA (fructose-bisphosphate aldolase) was renamed as FDPA in this study to avoid confusion with FBA (flux balance analysis).

* Detailed enzyme and reaction information can be found in the BIGG database [4].

Table D.3. Enzyme specific activity values

Enzyme	a_i $(\frac{mmol}{hr * g\ E})$	Source
LCTSt	540	[5]
GLCpts/MANpts	361.14	[6]
HEX	9480	[7]
G6PDH2r	6240	[8]
PGL	31200	[9]
GND	1920	[10]
RPE	133886.21 (+)/133.94 (-)	[11]/Predicted by DLTKcat
RPI	316635.95 (+)/4793.05 (-)	[12]/Predicted by DLTKcat
PGI	12756 (+)/870 (-)	[13]/[14]
PFK	11400	[15]
FDPA	28620	[16]
GAPD	2400	[17]
PGK	28800	[18]
PGM	7440	[19]
ENO	15600	[20]
PYK	3300	[21]
LDH_L	141000	[22]
LDH_D	126000	[23]
ACALD	2916	[24]
ALCD2x	11451.6	[25]
PFL	720	[26]
PDH	1500	[27]
PTAr	428400	[28]
ACKr	65220	[29]

ACCOAC	360	[30]
PGMT	1890	[31]
UDPG4E	1620 (+)/13998 (-)	[32]/[33]
GALU	285.6	[34]
GALK	690	[35]
GALT	1020	[36]
MAN6PI	270 (+)/ 639.79 (-)	[37]/[38]
PMANM	82.99	Predicted by DLTKcat
MAN1PT	144	[39]
MANT_EPS/GLCT_EPS/GALT_EPS	250.61	Estimated in this study
WZX	37.44	[40]
Growth (biomass_LPL60)	107.4	[41]
Acid exportation (lactic and acetic acid)	6360	[42]

* (+) means the forward reaction and (-) means the reverse reaction.

D.3 Figures

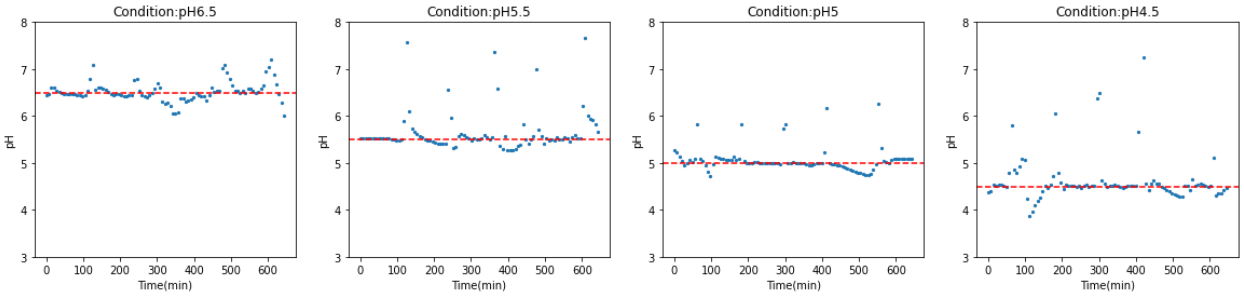

Figure D.1. The control of pH for 4 different conditions: pH=6.5, 5.5, 5.0 and 4.5.

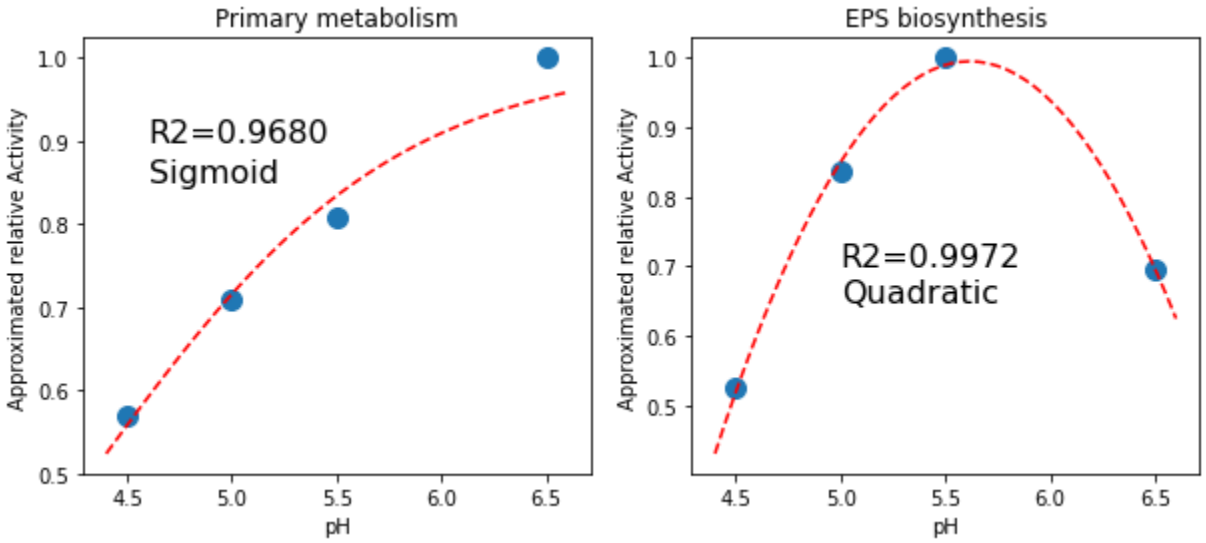

Figure D.2. The approximation of pH dependent enzyme activity coefficients for enzymes in primary metabolism (left) and EPS biosynthesis (right).

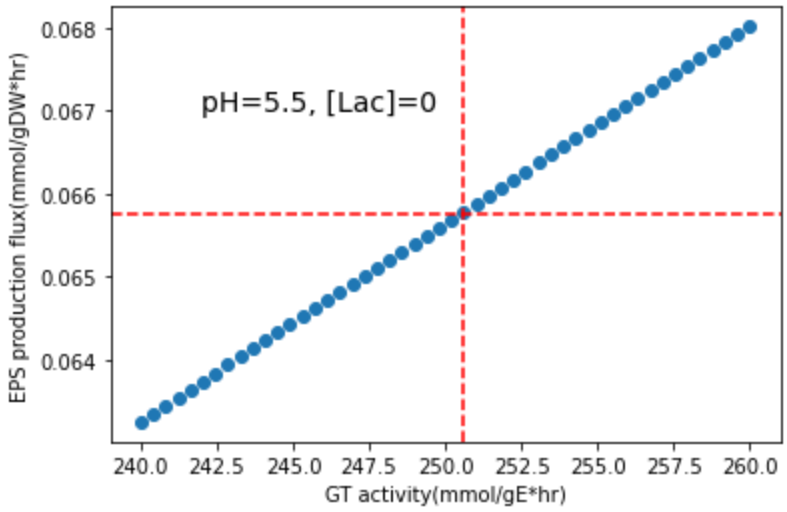

Figure D.3. The estimation of enzyme activities of glycosyltransferases (MANT_EPS, GLCT_EPS and GALT_EPS).

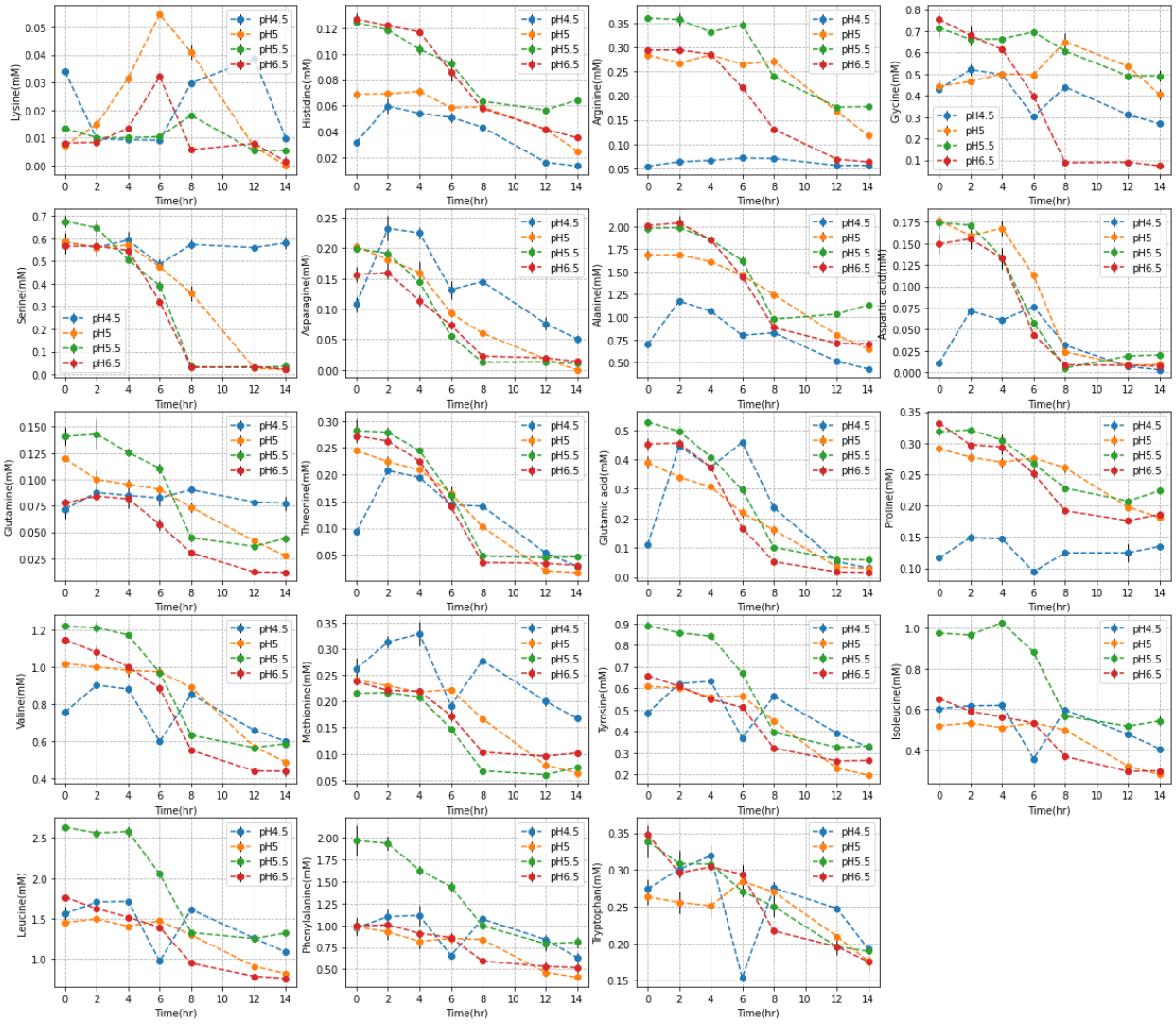

Figure D.4. The concentrations of 19 essential amino acids. Cysteine was not detected.

References

1. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. J Biol Chem. 2006;281: 40041–40048.

2. Price MN, Deutschbauer AM, Arkin AP. GapMind: Automated Annotation of Amino Acid Biosynthesis. mSystems. 2020;5. doi:10.1128/mSystems.00291-20

3. Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. Brief Bioinform. 2023;25. doi:10.1093/bib/bbad506

4. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22.

5. Geertsma ER, Duurkens RH, Poolman B. The activity of the lactose transporter from Streptococcus thermophilus is increased by phosphorylated IIA and the action of beta-galactosidase. Biochemistry. 2005;44: 15889–15897.

6. Christiansen I, Hengstenberg W. Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system--two highly similar glucose permeases in Staphylococcus carnosus with different glucoside specificity: protein engineering in vivo? Microbiology. 1999;145 (Pt 10): 2881–2889.

7. Meyer D, Schneider-Fresenius C, Horlacher R, Peist R, Boos W. Molecular characterization of glucokinase from Escherichia coli K-12. J Bacteriol. 1997;179: 1298–1306.

8. Banerjee S, Fraenkel DG. Glucose-6-phosphate dehydrogenase from Escherichia coli and from a “high-level” mutant. J Bacteriol. 1972;110: 155–160.

9. Holt JF, Kiedrowski MR, Frank KL, Du J, Guan C, Broderick NA, et al. Enterococcus faecalis 6-phosphogluconolactonase is required for both commensal and pathogenic interactions with Manduca sexta. Infect Immun. 2015;83: 396–404.

10. Veronese FM, Boccù E, Fontana A. Isolation and properties of 6-phosphogluconate dehydrogenase from Escherichia coli. Some comparisons with the thermophilic enzyme from Bacillus stearothermophilus. Biochemistry. 1976;15: 4026–4033.

11. Akana J, Fedorov AA, Fedorov E, Novak WRP, Babbitt PC, Almo SC, et al. D-Ribulose 5-phosphate 3-epimerase: functional and structural relationships to members of the ribulose-phosphate binding (beta/alpha)8-barrel superfamily. Biochemistry. 2006;45: 2493–2503.

12. Zhang RG, Andersson CE, Savchenko A, Skarina T, Evdokimova E, Beasley S, et al. Structure of Escherichia coli ribose-5-phosphate isomerase: a ubiquitous enzyme of the pentose phosphate pathway and the Calvin cycle. Structure. 2003;11: 31–42.

13. Gao H, Chen Y, Leary JA. Kinetic measurements of phosphoglucose isomerase and phosphomannose isomerase by direct analysis of phosphorylated aldose–ketose isomers using tandem mass spectrometry. Int J Mass Spectrom. 2005;240: 291–299.

14. Verhees CH, Huynen MA, Ward DE, Schiltz E, de Vos WM, van der Oost J. The phosphoglucose isomerase from the hyperthermophilic archaeon Pyrococcus furiosus is a unique glycolytic enzyme that belongs to the cupin superfamily. J Biol Chem. 2001;276: 40926–40932.

15. Kotlarz D, Buc H. Phosphofructokinases from Escherichia coli. Methods Enzymol. 1982;90 Pt E: 60–70.

16. Baldwin SA, Perham RN, Stribling D. Purification and characterization of the class-II D-fructose 1,6-bisphosphate aldolase from Escherichia coli (Crookes' strain). Biochem J. 1978;169: 633–641.

17. D'Alessio G, Josse J. Glyceraldehyde phosphate dehydrogenase of Escherichia coli. Structural and catalytic properties. J Biol Chem. 1971;246: 4326–4333.

18. Fifis T, Scopes RK. Purification of 3-phosphoglycerate kinase from diverse sources by affinity elution chromatography. Biochem J. 1978;175: 311–319.

19. D’Alession G, Josse J. Phosphoglycerate kinase and phosphoglyceromutase from Escherichia coli. Methods Enzymol. 1975;42: 139–144.

20. Kühnel K, Luisi BF. Crystal structure of the Escherichia coli RNA degradosome component enolase. J Mol Biol. 2001;313: 583–592.

21. Waygood EB, Sanwal BD. The Control of Pyruvate Kinases of Escherichia coli: I. PHYSICOCHEMICAL AND REGULATORY PROPERTIES OF THE ENZYME ACTIVATED BY FRUCTOSE 1,6-DIPHOSPHATE. J Biol Chem. 1974;249: 265–274.

22. Hensel R, Mayr U, Fujiki H, Kandler O. Comparative studies of lactate dehydrogenases in lactic acid bacteria. Amino-acid composition of an active-site region and chemical properties of the L-lactate dehydrogenase of Lactobacillus casei, Lactobacillus curvatus, Lactobacillus plantarum, and Lactobacillus acidophilus. Eur J Biochem. 1977;80: 83–92.

23. Kochhar S, Chuard N, Hottinger H. Cloning and overexpression of the Lactobacillus bulgaricus NAD(+)-dependent D-lactate dehydrogenase gene in Escherichia coli:purification and characterization of the recombinant enzyme. Biochem Biophys Res Commun. 1992;185: 705–712.

24. Smith LT, Kaplan NO. Purification, properties, and kinetic mechanism of coenzyme A-linked aldehyde dehydrogenase from Clostridium kluyveri. Arch Biochem Biophys. 1980;203: 663–675.

25. Ouyang Y, Li Q, Kuang X, Wang H, Wu J, Ayepa E, et al. YMR152W from Saccharomyces cerevisiae encoding a novel aldehyde reductase for detoxification of aldehydes derived from lignocellulosic biomass. J Biosci Bioeng. 2021;131: 39–46.

26. Takahashi S, Abbe K, Yamada T. Purification of pyruvate formate-lyase from

Streptococcus mutans and its regulatory properties. J Bacteriol. 1982;149: 1034–1040.

27. Meinecke B, Bertram J, Gottschalk G. Purification and characterization of the pyruvate-ferredoxin oxidoreductase from Clostridium acetobutylicum. Arch Microbiol. 1989;152: 244–250.

28. Bergmeyer HU, Holz G, Klotzsch H, Lang G. Phosphotransacetylase from Clostridium kluyveri. Culture of the bacterium, isolation, crystallization and properties of the enzyme. Biochem Z. 1963;338: 114–121.

29. Winzer K, Lorenz K, DÜrre P. Acetate kinase from Clostridium acetobutylicum: a highly specific enzyme that is actively transcribed during acidogenesis and solventogenesis. Microbiology. 1997;143 (Pt 10): 3279–3286.

30. Sumper M. Acetyl-CoA carboxylase from yeast. Methods Enzymol. 1981;71 Pt C: 34–37.

31. Maino VC, Young FE. Regulation of glucosylation of teichoic acid. I. Isolation of phosphoglucomutase in Bacillus subtilis 168. J Biol Chem. 1974;249: 5169–5175.

32. Chen X, Kowal P, Hamad S, Fan H, Wang PG. Cloning, expression and characterization of a UDP-galactose 4-epimerase from Escherichia coli. Biotechnol Lett. 1999;21: 1131–1135.

33. Bauer AJ, Rayment I, Frey PA, Holden HM. The isolation, purification, and preliminary crystallographic characterization of UDP-galactose-4-epimerase from Escherichia coli. Proteins. 1991;9: 135–142.

34. Weissborn AC, Liu Q, Rumley MK, Kennedy EP. UTP: alpha-D-glucose-1-phosphate uridylyltransferase of Escherichia coli: isolation and DNA sequence of the galU

gene and purification of the enzyme. J Bacteriol. 1994;176: 2611–2618.

35. Reinhardt LA, Thoden JB, Peters GS, Holden HM, Cleland WW. pH-rate profiles support a general base mechanism for galactokinase (Lactococcus lactis). FEBS Lett. 2013;587: 2876–2881.

36. Lee LJ, Kimura A, Tochikura T. Presence of a single enzyme catalyzing the pyrophosphorolysis of UDP-glucose and UDP-galactose in Bifidobacterium bifidum. Biochim Biophys Acta. 1978;527: 301–304.

37. Yeom S-J, Kim Y-S, Lim Y-R, Jeong K-W, Lee J-Y, Kim Y, et al. Molecular characterization of a novel thermostable mannose-6-phosphate isomerase from Thermus thermophilus. Biochimie. 2011;93: 1659–1667.

38. Roux C, Bhatt F, Foret J, de Courcy B, Gresh N, Piquemal J-P, et al. The reaction mechanism of type I phosphomannose isomerases: new information from inhibition and polarizable molecular mechanics studies. Proteins. 2011;79: 203–220.

39. Preiss J, Wood E. SUGAR NUCLEOTIDE REACTIONS IN ARTHROBACTER. I. GUANOSINE DIPHOSPHATE MANNOSE PYROPHOSPHORYLASE: PURIFICATION AND PROPERTIES. J Biol Chem. 1964;239: 3119–3126.

40. Woebking B, Velamakanni S, Federici L, Seeger MA, Murakami S, van Veen HW. Functional role of transmembrane helix 6 in drug binding and transport by the ABC transporter MsbA. Biochemistry. 2008;47: 10904–10914.

41. Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio-Iglesias M, Lema JM, Kleerebezem R. Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. Biotechnology and Bioengineering. 2021. pp. 745–758. doi:10.1002/bit.27605

42. Schumacher R. Metabolic trade-offs arising from increased free energy conservation in Saccharomyces cerevisiae. Delft University of Technology. 2018. doi:10.4233/UUID:177E9F4C-F847-436D-9FD4-9ED97BA709D9