THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Constraint-based modeling of yeast metabolism and protein secretion

FEIRAN LI

CHALMERS
UNIVERSITY OF TECHNOLOGY

Systems and Synthetic Biology
Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Constraint-based modeling of yeast metabolism and protein secretion
FEIRAN LI
ISBN 978-91-7905-557-8

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5024
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover illustration: Using genome-scale models to study yeast metabolism, protein secretion and evolution

Printed by Chalmers digitaltryck
Gothenburg, Sweden 2021

Constraint-based modeling of yeast metabolism and protein secretion

Feiran Li
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

Yeasts are extensively exploited as cell factories for producing alcoholic beverages, biofuels, bio-pharmaceutical proteins, and other value-added chemicals. To improve the performance of yeast cell factories, it is necessary to understand their metabolism. Genome-scale metabolic models (GEMs) have been widely used to study cellular metabolism systematically. However, GEMs for yeast species have not been equally developed. GEMs for the well-studied yeasts such as *Saccharomyces cerevisiae* have been updated several times, while most of the other yeast species have no available GEM. Additionally, classical GEMs only account for the metabolic reactions, which limits their usage to study complex phenotypes that are not controlled by metabolism alone. Thus, other biological processes can be integrated with GEMs to fulfill diverse research purposes.

In this thesis, the GEM for *S. cerevisiae* was updated to the latest version Yeast8, which serves as the basic model for the remaining work of the thesis including two dimensions: 1) Yeast8 was used as a template for generating GEMs of other yeast species/strains, and 2) Yeast8 was expanded to account for more biological processes. Regarding the first dimension, strain-specific GEMs for 1,011 *S. cerevisiae* isolates from diverse origins and species-specific GEMs for 343 yeast/fungi species were generated. These GEMs enabled explore the phenotypic diversity of the single species from diverse ecological and geographical origins and evolution tempo among diverse yeast species. Regarding the second dimension, other biological processes were formulated within Yeast8. Firstly, Yeast8 was expanded to account for enzymatic constraints, resulting in enzyme-constrained GEMs (ecGEMs). Secondly, Yeast8 was expanded to the model CofactorYeast by accounting for enzyme cofactors such as metal ions, which was used to simulate the interaction between metal ions and metabolism, and the cellular responses to metal ion limitation. Lastly, Yeast8 was expanded to include the protein synthesis and secretion processes, named as pcSecYeast. pcSecYeast was used to simulate the competition of the recombinant protein with the native secretory-pathway-processed proteins. Besides that, pcSecYeast enabled the identification of overexpression targets for improving recombinant protein production.

When developing these complex models, issues were identified among which the lack of enzyme turnover rates, i.e., k_{cat} values, needs to be solved. Accordingly, a machine learning method for k_{cat} prediction and automated incorporation into GEMs were developed, facilitating the generation of functional ecGEMs in a large scale.

Keywords: genome-scale metabolic model, metabolism, phenotype diversity, protein secretion, yeast evolution.

List of Publications

This thesis is based on the following publications and manuscript:

Paper I: A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism.
Lu H†, **Li F**†, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, Beber ME, Kerkhoven EJ and Nielsen J. (2019) Nature communications

Paper II: Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection.
Lu H†, **Li F**†, Yuan L†, Domenzain I, Yu R, Wang H, Li G, Chen Y, Ji B, Kerkhoven EJ and Nielsen J. (2021) Molecular Systems Biology, *In press*

Paper III: Yeast optimizes metal utilization based on metabolic network and enzyme kinetics.
Chen Y, **Li F**, Mao J, Chen Y and Nielsen J. (2021) Proceedings of the National Academy of Sciences

Paper IV: Genome-scale modeling of the protein secretory pathway reveals novel targets for improved recombinant protein production by yeast.
Li F, Chen Y, Qi Q, Wang Y, Yuan L, Huang M, Elsemman IE, Feizi A, Kerkhoven EJ and Nielsen J. (2021) *Manuscript*

Paper V: Deep learning based k_{cat} prediction enables improved enzyme constrained model reconstruction.
Li F†, Yuan L†, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ and Nielsen J. (2021) *Under review*

Additional papers and manuscripts not included in this thesis:

Paper VI: Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species.
Domenzain I†, **Li F**†, Kerkhoven EJ and Siewers V. (2021) FEMS Yeast Research

Paper VII: Different Routes of Protein Folding Contribute to Improved Protein Production in *Saccharomyces cerevisiae*.
Qi Q, **Li F**, Yu R, Engqvist MKM, Siewers V, Fuchs J and Nielsen J. (2020) mBio

Paper VIII: SLIMEr: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework.
Sánchez BJ, **Li F**, Kerkhoven EJ and Nielsen J. (2019) BMC Systems Biology

Paper IX: Recombinant protein production requires metabolic reprogramming to provide more NADPH via activation of kinase Gcn2p.
Qi Q, **Li F**, Vorontsov E and Nielsen J. (2021) *Manuscript*

† co-first authorship

Contribution summary

Paper I. I co-designed the study, contributed to YeastGEM update, designed and generated panYeast8 and the 1,011 strain-specific GEMs, analyzed the data and wrote the paper.

Paper II. I constructed the GEMs for 343 yeast/fungi species, performed model related analysis and wrote the paper.

Paper III. I contributed to the CofactorYeast model construction.

Paper IV. I designed the study, constructed the pcSecYeast model, analyzed the data and wrote the paper.

Paper V. I co-designed the study, constructed the Bayesian pipeline for large-scale ecGEMs reconstruction, analyzed the data and wrote the paper.

Paper VI. I co-reviewed the literature and wrote the paper.

Paper VII. I performed the model simulation.

Paper VIII. I co-designed the mathematical formulation and processed the literature data.

Paper IX. I performed the model simulation.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between September 2017 and October 2021 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The project was co-supervised by Eduard J Kerkhoven and examined by Ivan Mijakovic. The project was funded by the European Union's Horizon 2020 research and innovation program projects DD-DeCaF (grant no. 686070), the Novo Nordisk Foundation (grant no. NNF10CC1016517) and VINNOVA center CellNova (2017-02105).

Feiran Li
September 2021

Table of Contents

Abbreviations

ABC	Approximate Bayesian computation
ATP	Adenosine triphosphate
BGL	β-glucosidase
BYCA	Budding yeast common ancestor
coreGEM	Core-genome-scale metabolic model
DL	Deep learning
EC number	Enzyme Commission number
ecGEM	Enzyme-constrained genome-scale metabolic model
ER	Endoplasmic reticulum
ERAD	Endoplasmic reticulum-associated protein degradation
ETFL	Expression and Thermodynamics Flux models
FCC	Flux control coefficient
FSEOF	Flux Scanning based on Enforced Objective Function
GECKO	GEM with Enzymatic Constraints using Kinetic and Omics data
GEM	Genome-scale metabolic model
GPR	Gene-protein-reaction
hGCSF	Human granulocyte colony stimulating factor
HGT	Horizontal gene transfer
HSA	Human serum albumin
HTF	Human transferrin
IP	Insulin precursor
ISCs	iron-sulfur clusters
KO	KEGG Orthology
memote	Metabolic model tests
NADH	Reduced nicotinamide adenine dinucleotide
panGEM	Pan-genome-scale metabolic model
PCA	Principal component analysis
pcGEM	Proteome-constrained genome-scale metabolic model
PHO	Acid phosphatase
PTM	Post-translational modifications
RBA	Resource Balance Analysis
RMSE	Root mean square error
SLIME	Split lipid into measurable entities
SMC-ABC	Sequential Monte Carlo Approximate Bayesian computation
t-SNE	t-Distributed Stochastic Neighbor Embedding
TCA cycle	Tricarboxylic acid cycle

Acknowledgements

It has been a wonderful experience studying at SysBio, Chalmers. I remember the day when I came to Sweden. It was a shiny afternoon in Sep 2017. Four years have passed, and there comes graduation. I would like to express my gratitude for that has been in this journey.

Thanks to my supervisor, Jens, for giving me the chance to join SysBio, for those exciting projects, for the trust and freedom you give to me to perform the research with my interest, for recognizing the value of my ideas, and for helping me to develop those ideas into good quality research. You are not only a source of academic wisdom, but also a caring mentor giving me a lot of advice about personal development. I would benefit from this experience for the rest of my life. Thanks to my co-supervisor, Ed, for your kindness and encouragement, for invaluable discussions and suggestions in ~200 meetings, and for reviews and constructive comments of all my reports and manuscripts. Without you, I cannot make it.

During these four years, I got many amazing collaborations. Thanks to Hongzhong for the collaboration of Yeast8 and for inviting me to join the evolution project for yeast species. Your hardworking exemplifies and motivates me to become a better researcher. Thanks to Le for the great work in the DLkcat project. Your intelligence and problem-solving ability made the project and life in that period much easier and enjoyable. Thanks to Qi and Yanyan for your excellent experiment for pcSecYeast validation. Thanks to Ben and Gang for your kindness and patience in tolerating my frequent drop-off office visits with endless questions. Thanks to Iván for all valuable discussions in the ecGEM reconstruction and GEM review writing. Thanks to Amir for teaching me protein secretion at the beginning of my PhD. Thanks to Ibrahim for the guidance about the proteome-constrained concept. Thanks to all my co-authors for your valuable input. Thanks to Ben and Yu for reviewing my thesis and giving valuable comments. Thanks to MSB group for attending the meeting and giving me valuable feedback.

I would like to thank to Gang, Yating, Jichen for giving me a lot of help when I arrived in Sweden for the first time. Thanks to Qi, Yanyan, Jiwei, Zhengming, Chunjun, Peishun, and Rosemary for those wonderful afterworks with cooking and exploring Gothenburg. You make my PhD life colorful. Thanks to my friend Yujing, Dan, Na, Wenqin, Lu, Ziyu and Wangsheng for encouragements during all these years. All your encouragements make a better me. Special thanks go to Yu, for the joy, support, encouragement, and accompany especially during the downs in my PhD. Thanks to my younger sister, Dr. Han, for giving me much peer pressure and accompany all the way. Thanks to my mom and dad for your deep love, which brought the optimistic and enthusiastic part of me.

Thanks to everyone that I have met along the journey!
Feiran

1. Background

1.1 Yeast: one of the widely used microbes

Yeast may be one of the earliest domesticated microorganisms as the fermentation by yeast can be predated to 7,000 BC [1]. Since then, yeast has been domesticated for wide use in the industry.

The conventional yeast *Saccharomyces cerevisiae* is by far the most dominant yeast for industrial applications. A key metabolic trait of *S. cerevisiae* is the aerobic fermentation, also referred as the Crabtree effect, shaping its industrial usage for wine production and biofuel production. Besides that, *S. cerevisiae* is dominantly used in fundamental research. As one of the simplest eukaryotes and the first eukaryote with its whole genome sequenced, *S. cerevisiae* is used as a model organism to study eukaryotes. The findings gained from *S. cerevisiae* can be transferred to other eukaryotes such as human cells. For example, *S. cerevisiae* has been used as the model organism to study cell cycle and human diseases such as Parkinson's and Alzheimer's [2], [3].

Besides *S. cerevisiae*, more than 1,500 yeast species have been identified from diverse ecological and geographical habitats [4]. They exhibit remarkably diverse phenotypes, which have enabled them to inhabit every continent and even some extreme biomes. These non-conventional yeasts have shown several advantages over *S. cerevisiae* for metabolic diversity, product profile and growth physiology. For example, thermotolerant yeasts such as *Kluyveromyces marxianus* can grow at 45°C and even tolerate up to 50°C [5], oleaginous yeasts such as *Yarrowia lipolytica*, *Lipomyces sp.*, and *Rhodosporidium sp*. can accumulate up to 50-70% of biomass content as lipid [6], [7], *Zygosaccharomyces rouxii* can grow in medium with up to 90% (w/v) of sugar [8], and a newly isolated extremophilic yeast *Rhodotorula frigidialcoholis* grow as low as 0°C accompanied with ethanol production [9] Substrate utilization tests also showed diverse substrate utilization profiles among yeast species [10]. With the diverse phenotypes, non-conventional yeasts have gained more attention [11]–[13]. That said, many yeast species that could have industrial potential are not characterized yet [14]. Understanding those unexploited yeasts would thus accelerate the yeast industrial application.

1.2 Cellular metabolism

Most phenotypes are affected by metabolism. In order to understand the phenotype, there is, therefore, a requirement to understand cell metabolism. Metabolism represents the sum of thousands of biochemical reactions, describing the conversion of nutrients into different intermediate chemicals and energy for growth that occurs in living cells (**Figure 1**).

Cells dynamically adjust the flow of metabolites, i.e., metabolic flux, to adapt to environmental conditions. Therefore, cells could also be adjusted to direct the metabolic

flux toward the metabolites of interest such as value-added chemicals, which is part of the field of metabolic engineering [15]. However, due to the complex and intertwined instincts of metabolism, this requires a systematic understanding of metabolism and remains challenging.

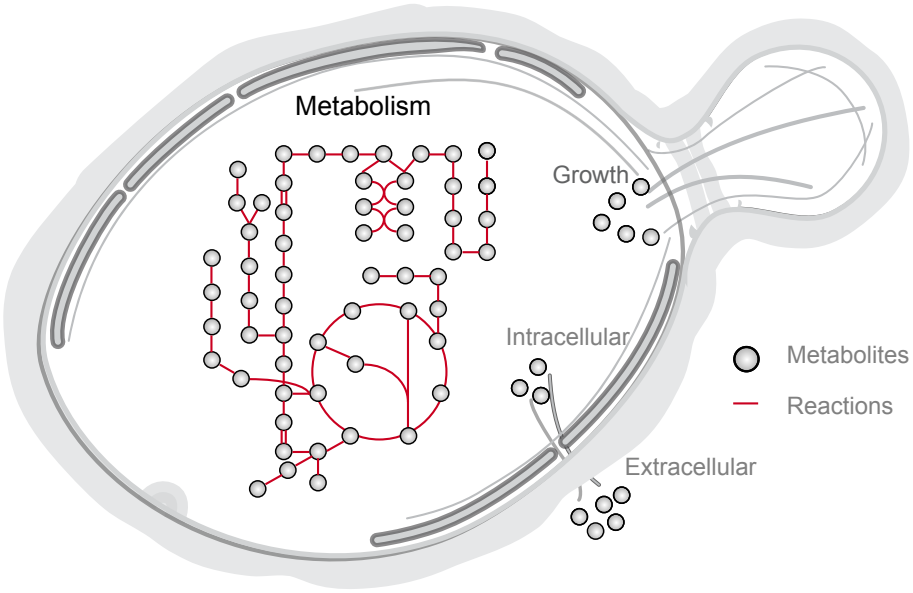

Figure 1 Diagram of metabolism. Yeast compartmentalized figure source: SwissBioPics under CC BY4.0 license.

1.3 Genome-scale metabolic model

A genome-scale metabolic model (GEM) incorporates the genome-scale metabolic enzymes, metabolic reactions, and metabolites through gene-protein-reaction (GPR) associations, describing the whole set of metabolic conversions of the cell and serving as a platform for systematic analysis for metabolism (**Figure 2a**). Prerequisite to the reconstruction of a GEM is the whole-genome sequence. With the development of the sequencing techniques, numerous GEMs have been reconstructed for different organisms in the past [16], [17], which find use in different fields, such as predicting metabolic engineering targets, uncovering interspecies interaction and evolution process, analyzing coupling reaction sets, understanding strain phenotype, and guiding model-driven discoveries (**Figure 2b**) [17]–[19].

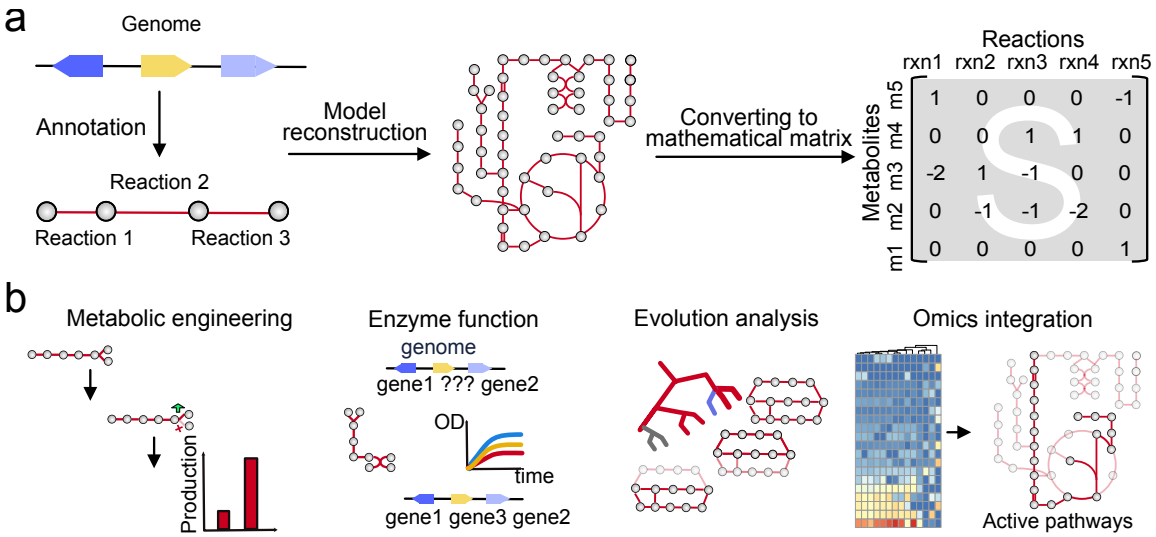

Figure 2 GEM reconstruction and its applications. a) GEM reconstruction. Metabolic genes from the genome are annotated to metabolic reactions. All reactions in the GEM interact together through the shared metabolites. The reconstructed GEM is then converted into a mathematical matrix to indicate the substrates and products of each reaction, where metabolites with negative coefficients represent substrates and positive coefficients represent products. The coefficients stand for the stoichiometry of the metabolites in the reaction. b) GEM applications. GEMs can identify the knockout and overexpression targets for metabolic engineering through several developed algorithms. GEMs can also identify reaction gaps in biological pathways, guiding the new gene identification. Through GEM comparison of multiple species, metabolic diversity for those species during the evolution can be identified. Lastly, GEMs can be used to identify the active pathways and reactions for a certain purpose by integration of omics data, which would benefit the understanding of cell metabolism.

Initially, GEM reconstruction required much manual work, as stated in the well-documented protocol [20], which indicates that reconstruction and curation of a GEM can take several months to years. More recently, various automatic and semi-automatic tools have been developed, such as RAVEN [21], [22] and ModelSEED [23], which can generate draft GEMs based on genome annotations and therefore accelerate the reconstruction processes. Since then, other template-based homology searching toolboxes have been developed, including CarveME [24], RAVEN version 2 [21] and AuReMe [25], which use one or several well-curated models as the template and thus are more preferred in large-scale GEM reconstruction [24], [26]. With the development of these tools, the effort for draft GEM reconstruction has decreased tremendously. However, additional time and effort for gap-filling and manual curation are still required to generate robust and high-quality GEMs.

When simulating GEMs, optimization problems are solved to estimate metabolic fluxes. Since the number of metabolic reactions is larger than the number of metabolites, the optimization problem is underdetermined. Constraints are imposed to reduce the solution space, such as constraints for the mass balance and reaction bounds that define the limits on the rate of a reaction (**Figure 3**). A unique flux distribution from the solution space is determined through optimization of an objective function, which is usually to maximize growth based on the assumption that microbial cells have evolved with the objective to grow as fast as possible. The simulated metabolic fluxes can mathematically represent the

metabolic state of cells under certain conditions, which supplies a snapshot of the overall metabolic conversion.

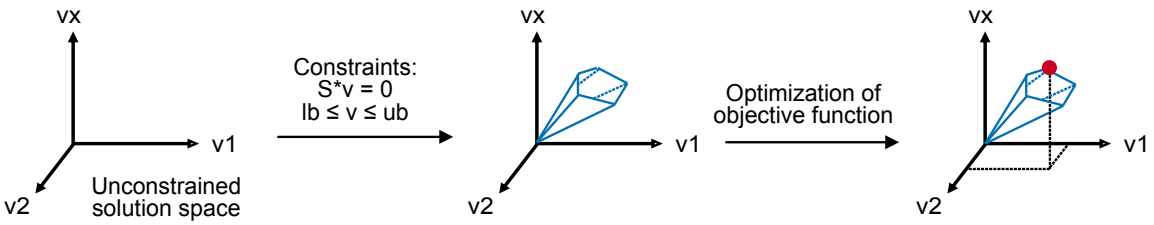

Figure 3 Constraints-based approach for flux simulation. The solution space is infinite without any constraint. In order to limit the solution space to mimic the *in vivo* flux distribution, multiple constraints are imposed, such as mass balance by assuming a steady state. The second constraint is on the reaction rates, which constrains the upper and lower bound for each reaction. Applying those two basic constraints to the GEM would shrink the solution space. The optimal solution can be found by optimizing the objective function in the allowable solution space. The red dot in the figure represents the optimal solution space. vx means the flux for a common objective function: growth.

The first GEM for yeast species was developed in 2003 for *S. cerevisiae* [27]. Since then, GEMs have been reconstructed for various yeast species [19], [28], although these models are not all equally developed. There are currently 12 yeast species with available GEMs (**Table 1**). Among them, five species were updated through multiple rounds to provide additional annotation or improve the scope. The GEM for *S. cerevisiae* has been updated 19 times, demonstrating the importance of continuous curation in GEM development.

As large parts of (central carbon) metabolism is conserved across species, there is intense crossing referencing in the GEM development for yeast species. A new GEM for a certain species thereby adopts information from an existing GEM from another species. On the other hand, the GEMs of *S. cerevisiae* were used as templates for generating GEMs of other yeast species such as *Y. lipolytica* [29], [30], *Kluyveromyces lactis* [31], *K. marxianus* [32], *Candida glabrata* [33], and *Pichia pastoris* [34], indicating the potential of using a well-curated GEM as the template for reconstruction of other GEMs.

Those developments aided the elucidation of biological processes and inspired yeast cell factory design, such as the production of sesquiterpenes [35], vanillin [36], 3-hydroxypropionic acid [37] and fumaric acid [38] by *S. cerevisiae*, malate [39] and acetoin [40] by *C. glabrata* and human recombinant protein by *P. pastoris* [41]. Detailed applications have been reviewed in previous references [19], [28], [42]–[44].

Table 1 Genome-scale metabolic model of yeast species.

Organism	Available GEMs	Period
S.cerevisiae	17	2003-2018
P.pastoris	6	2010-2017
Y.lipolytica	5	2012-2018
S.stipitis	4	2012-2018
C.glabrata	1	2012
C.tropicalis	1	2016
K.lactis	1	2014
K.marxianus	1	2019
R.toruloides	2	2019
S.pombe	1	2012
Z.parabailii	1	2018
L.kluyveri	1	2020
O.polymorpha	1	2021

1.4 Advanced GEM development: proteome-constrained GEM

Even though GEMs have been used as excellent platforms for understanding metabolism, they only consider stoichiometry constraints of metabolism, limiting their usage for simulating complex phenotypes that are not constrained by pure metabolism. Therefore, other constraints have been integrated into GEMs to extend their prediction potential. One of the critical constraints is the proteome constraint, which relies on the assumption that cells allocate their finite proteome resource towards diverse biological processes for faster growth or better fitness. With this assumption, corresponding proteome constraints are integrated into GEMs, which enables simulation and explanation for specific phenotypes such as overflow metabolism [45] and Crabtree effect [46], suggesting that it could be a valuable addition towards basic GEMs. In this section, I will introduce two types of proteome-constrained GEMs (pcGEMs): coarse- and fine-grained pcGEMs (**Figure 4**).

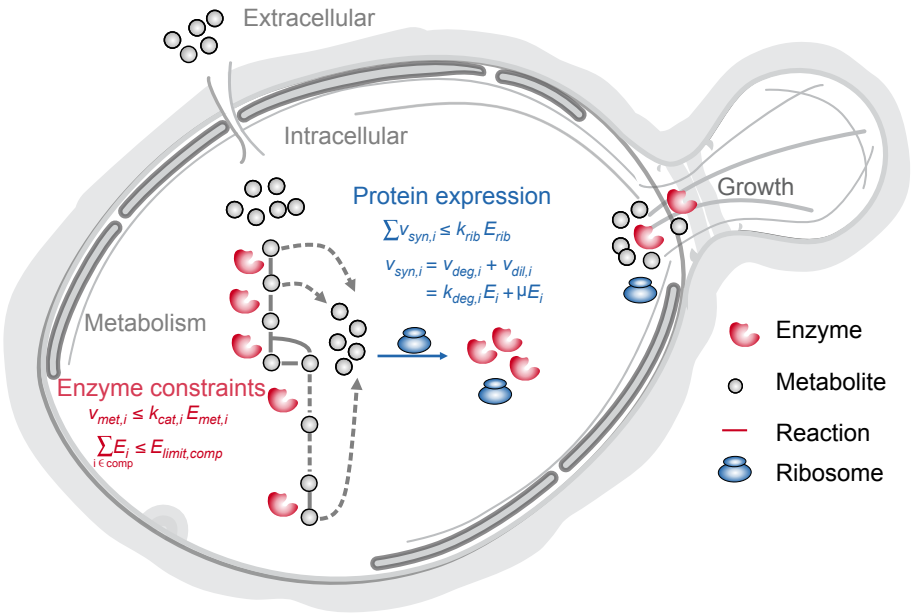

Figure 4 Schematic figure for the reconstruction of proteome-constrained GEMs (pcGEMs) which incorporates more constraints/processes into the basic GEM. There are two main types of pcGEMs: coarse- and fine-grained pcGEMs. The coarse-grained pcGEM adds genome-scale enzymatic constraints into the basic GEM and assumes an upper limit for the total metabolic enzymes (red constraints in the figure). On the other hand, the fine-grained pcGEM contains the additional processes such as protein expression (additional constraint and synthesis reactions marked in blue) besides the same constraints as the coarse-grained pcGEM. Yeast compartmentalized figure source: SwissBioPics under CC BY4.0 license.

1.4.1 Coarse-grained pcGEMs

Coarse-grained proteome-constrained approaches assume that reaction fluxes should not exceed their maximum capacity. Thus, each metabolic reaction is constrained in the GEM with a particular enzymatic cost, mathematically represented by the turnover rate, i.e., k_{cat} value and abundance of the enzyme that catalyzes the reaction. Enzyme costs of individual reactions sum up as the total enzyme cost, which is further constrained by the finite total proteome abundance in a cell. These approaches do not add new processes or genes compared with the basic GEM, but they could shrink the feasible solution space and better predict the cell's state. There are several different pipelines to generate such coarse-grained pcGEMs. The GECKO pipeline adds enzymatic constraints and enzyme concentrations in an explicit way incorporating the enzymatic constraints in each reaction with the corresponding enzyme as a pseudo metabolite, which benefits further proteome data integration [46]. In contrast, the sMOMENT pipeline adds an overall enzyme pool as the pseudo metabolite and assigns the adequate proteome pool to each metabolic reaction. This approach generates much smaller enzyme constrained models, which would decrease the computing demand and improve efficiency [47], especially for large GEMs such as human GEM. The coarse-grained pcGEM generated by the GECKO pipeline is referred to as enzyme-constrained GEM (ecGEM).

An ecGEM for S. cerevisiae was developed with the publication of the GECKO pipeline, named as ecYeast7 [46], in which over 750 enzymes were assigned with enzymatic constraints. ecYeast7 can accurately predict the maximum growth rate without defining a limitation on the substrate uptake rate which is incapable in basic GEM as enzymatic costs do not constrain its reaction rates. ecYeast7 can also simulate the metabolic shift that is observed with increasing energy demands in *S. cerevisiae*, i.e., Crabtree effect, demonstrating that the proteome allocation theory could explain the Crabtree phenotype.

Recently, ecGEMs for *Y. lipolytica* and *K. marxianus* were developed with GECKO version 2.0 [48]. It was demonstrated that these ecGEMs could be integrated with proteomics data to compute enzyme usage. It was found that there is enforced proteome allocation towards the central carbon metabolism and diversified utilization of isozymes, which showcased the metabolic robustness of the microbes under environmental stress and nutrient limitation.

1.4.2 Fine-grained pcGEMs

Compared with the coarse-grained proteome-constrained approaches, the fine-grained approaches are more fundamental as they compile detailed protein expression processes such as protein translation and complex assembly. There are two parts in the fine-grained model: the metabolic part derived from the basic GEM and the protein expression part for protein synthesis. These two parts are coupled together since metabolism supplies the substrate and energy for protein expression parts such as ribosome and enzyme synthesis. At the same time, metabolic reactions are catalyzed by enzymes and therefore constrained by the protein synthesis rates. Additionally, the fine-grained approaches no longer contain a biomass equation where the protein content is represented as a fixed amount of amino acids but rather a dynamic protein content comprised by changing composition of enzymes. There are several pipelines to generate fine-grained pcGEMs such as COBRAme [49], RBA [50] and ETFL [51]. The COBRAme pipeline adopts the basic GEM formulation, and all additional constraints are integrated by adding pseudo-metabolites, which generates a more standard model file including the complete information required for the simulation. In contrast, RBA does not directly affect the model stoichiometry but instead extends the linear programming file with the protein-related constraints directly. Thus, RBA generates an easily interpretable model without many modifications in the metabolic reactions derived from the basic GEMs. ETFL is one step further advanced than COBRAme and RBA, since it can incorporate thermodynamics constraints. Instead of using the model default reaction reversibility, ETFL adopts the thermodynamic flux analysis, coupling reaction directionality with Gibbs free energies and metabolite concentrations.

Fine-grained pcGEMs have been developed for the conventional yeast *S. cerevisiae*. The first one is yETFL, which expanded the original ETFL approach to eukaryotes [52]. yETFL was used to simulate the Crabtree effect and gene essentiality. Another fine-grained pcGEM for *S. cerevisiae* is pcYeast [53]. The model contains protein expression,

translation, folding, and protein degradation at genome-scale with fully compartmentalized formation, identifying the compartment constraint, especially mitochondria, towards the Crabtree effect. This model enables the computation of the protein cost to identify active constraints during the growth phase, suggesting that resource optimization can explain the metabolic strategy for eukaryotic cells for growth maximization.

1.5 Other model development from the basic GEM

1.5.1 Whole-cell model

In addition to metabolism and protein expression, other cellular processes such as cell replication could also be integrated towards the concept of a whole-cell model [54], enabling simulate any interval of a cell cycle and simultaneously model multiple processes in the cell. The simulation is based on time increments from the previous simulations dictating enzyme abundances and metabolite concentrations, rather than the steady-state assumption as in the GEMs. This whole-cell modeling approach broadens the predictive scope and gives a comprehensive insight into cellular physiology.

A whole-cell model for *S. cerevisiae* has been developed, named WM_S288C, which includes 15 cellular states and 26 cellular processes [55]. This model enables simulation of the cell cycle and real-time dynamic allocation of intracellular molecules, which was used to understand gene essentiality towards cell function by identifying the impact of experimentally determined essential and non-essential genes.

1.5.2 Specific model for certain purpose

In order to fulfill specific requirements, GEMs can be extended to incorporate specific processes or enrich existing processes [56]–[58]. To various aims, several specific models of yeast were reconstructed, such as to estimate the iron-recruiting enzyme abundance (Yeast7.Fe) [59], to simulate growth under different temperatures (etcYeast7.6) [60], to predict recombinant protein production (ihGlycopastoris) [41] and to incorporate regulatory network [56]. Each of these models adds more constraints or additional information, mostly in terms of the reactions and metabolites. Those extended models can simulate a specific delicate cellular state under a specific condition, enabling a broader application.

Among these models, two models are based on basic GEMs: Yeast7.Fe [59] and ihGlycopastoris [41]. Yeast7.Fe was developed to cover the complete iron metabolism based on the GEM for *S. cerevisiae*, which enables estimate iron requirements roughly from metabolic fluxes. The ihGlycopastoris expands the original basic GEM for *P. pastoris* to the N-glycosylation process, enabling the simulation and identification of metabolic engineering targets of recombinant protein production.

etcYeast7.6 was developed based on ecGEM for *S. cerevisiae* [60]. By linking the enzyme activities for metabolic enzymes to cultivation temperature, this model can simulate the

growth with different temperatures. The regulatory hybrid model was also developed based on the ecGEM for *S. cerevisiae*, integrating Boolean regulatory module, which enables exploration of the interplay between signaling and metabolism [56].

1.6 What model to select: tradeoff between model scope and uncertainty

Limited by existing knowledge, there are always uncertainties in GEMs. As for the basic GEMs, the uncertainties would come from genome annotation, environment specification, biomass formulation, flux distributions[61], which may mask the biological insights. Those uncertainties would accumulate to any models derived from the basic GEMs.

Reconstruction for both coarse- and fine-grained pcGEMs rely heavily on the enzyme specific k_{cat} values, which are used in constraining the enzyme usage and synthesis [46], [47], [62], but there remain a few challenges. First, compared with the number of organisms and enzymes, the measurement of k_{cat} values is far less than complete even for the well-studied species. For example, only less than 10% of total enzymatic reactions have measured k_{cat} values in *E. coli* [63]. Thus, to generate pcGEMs, numerous k_{cat} values are defined through fuzzy matching with k_{cat} values measured with other substrates, organisms, or by introducing wild cards in the Enzyme Commission number (EC number). For example, only 47 k_{cat} values are fully matched with enzymes and substrates in the ecGEM for *S. cerevisiae* ecYeast7 [46]. Second, the k_{cat} values query process relies heavily on EC number annotation for enzymes since the k_{cat} data mainly stored with EC numbers rather the enzyme sequence. For non-well annotated species, getting reliable EC numbers would be a challenge. Third, k_{cat} values were mostly measured *in vitro* which could differ in magnitude with *in vivo* k_{cat} values [64]. All these three challenges introduce uncertainties that would cause substantial manual work for generating a robust and functional pcGEM.

In addition to the k_{cat} values, fine-grained pcGEMs and whole-cell models require more parameters and detailed knowledge, such as the ribosome subunit composition and the translation scanning process. That knowledge is currently fragmentally collected from literature, while unmeasured parameters are rather assigned by estimation, which would cause further uncertainties in the model.

There is a dilemma for whether to go for a simple model with less uncertainties or a complex model which can describe more processes. The answer depends on the focused scientific questions, albeit it is often a good choice to start with the simpler model.

1.7 Protein synthesis and secretion

Typically, the protein-related part in the fine-grained pcGEMs only contains partial protein synthesis processes to the point of protein translation. However, after the gene transcription and peptide translation, the nascent peptide goes through many steps before maturing.

Besides that, about 30% of proteins in eukaryotic cells are processed by the secretory pathway [65], which spans several organelles such as Endoplasmic reticulum (ER) and Golgi, carrying out peptide translocation, folding, ER-associated protein degradation (ERAD), sorting processes as well as various post-translational modifications (PTMs) to ensure proper protein functionality [66]. The diverse combination of these processes is needed for processing different secretory proteins based on their protein features, which makes the protein secretory pathway complex to describe. Besides that, protein folding is an error-prone process. Mutation in the sequence, errors during the synthesis, or environmental insults cause the newly synthesized protein to misfold. Misfolded proteins are prioritized to be eliminated rapidly by the ERAD pathway. However, they can also be retained and accumulated in the ER. When it exceeds ER tolerance, it will trigger cell disorder, e.g., many diseases such as Alzheimer's and Parkinson's in humans [67], [68]. Unraveling the processing and energy costs for proteins passing through the secretory pathway and how the cell distributes energy and enzymes to process these proteins is therefore desirable, as this would facilitate better understanding of protein secretion. Therefore, there is an urgent requirement to develop a model to include the complete protein synthesis and secretion pathway.

1.8 Aim and significance

In this thesis, I conducted research on modeling yeast species. This thesis focuses on developing yeast models in two dimensions: developing models for more yeast species/strains and integrating more constraints/processes into yeast models (**Figure 5**).

To keep track of evolving understanding of yeast metabolism, I updated the GEM for *S. cerevisiae* to the latest version, Yeast8, which serves as a comprehensive reconstruction of yeast metabolism, and as the foundation stone for other research in this thesis (**Paper I**).

On the one hand, I used Yeast8 as the template to build GEMs for multiple *S. cerevisiae* strains (**Paper I**) and multiple yeast/fungi species (**Paper II**). The first study demonstrates that the GEM reconstruction can systematically investigate and explain variability across *S. cerevisiae* strains within the same species, while the second systematically studied yeast metabolic diversity evolved through a long history. The large-scale model reconstruction in this dimension showcases how a well-curated GEM could facilitate the GEMs reconstruction for phylogenetically close organisms.

On the other hand, I demonstrated that integrating more biological processes and constraints into yeast GEMs could explain complex phenotypes. Firstly, ecGEMs for yeast/fungi species were developed to simulate growth and identify critical controlling enzymes for different conditions (**Paper I & Paper II**). Next, integrating enzyme cofactors, primarily metal ions, into Yeast8, which resulted in the model CofactorYeast, enabled simulation of cellular responses to metal ion limitation (**Paper III**). Lastly, a fine-grained pcGEM of yeast with the expansion of protein secretion, namely pcSecYeast was developed to systematically study the protein secretion on the genome scale (**Paper IV**).

As I have introduced in the background chapter, complex models always come with more uncertainties. Thus, in the last part (**Paper V**), I discuss how to reduce the model uncertainties to accelerate the large-scale functional ecGEM reconstruction.

This thesis aims to exemplify the model development path and compare the predictive potential for models with different complexities. This thesis is expected to be a solid basis for yeast model development.

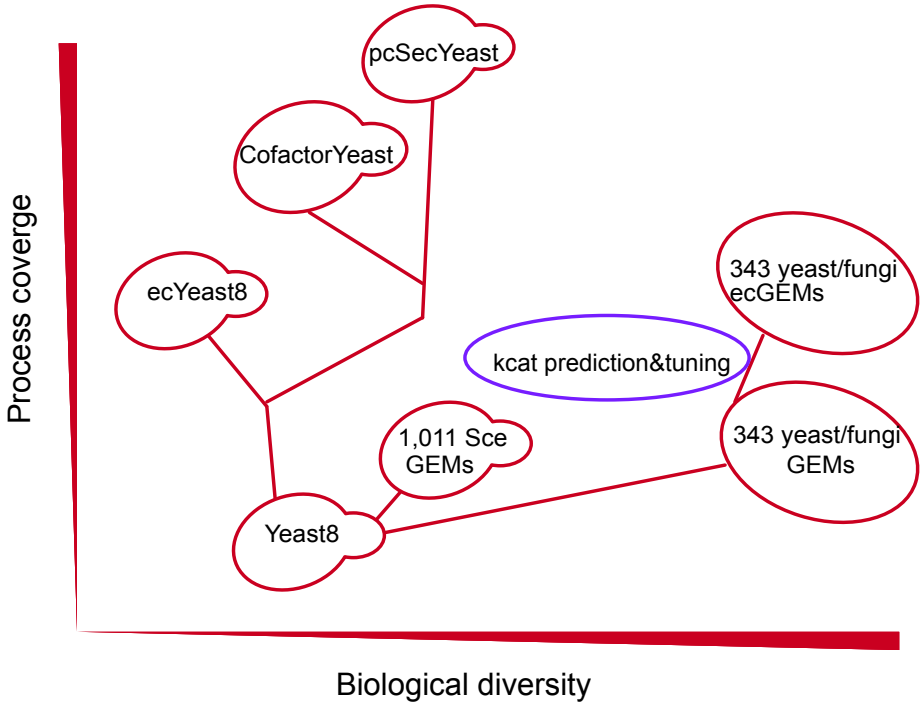

Figure 5 Graphic abstract for this thesis. The red circles mean models generated in this thesis while the purple circle represents the tool developed to facilitate model development.

2. Yeast8: the latest consensus GEM for *S. cerevisiae*

As mentioned in the background part, the GEM is the foundation stone for other model development. Particularly, this thesis starts with the *S. cerevisiae* GEM, which serves as the template model in the first-dimension expansion to generate GEMs for more yeast strains/species and as a crucial part of yeast complex models in the second dimension. Thus, the quality of the *S. cerevisiae* GEM would directly influence all yeast models developed in this thesis. Therefore, in this chapter, I discuss how the current GEM for *S. cerevisiae* was curated and updated to the latest version, improving the metabolic pathway annotation completeness and accuracy (**Paper I**).

2.1 Sustainable development of model update

Model development is a continuous effort with numerous changes and versions, which raises a major challenge to properly tracking those changes. Previously, researchers documented changes in a log file or a table. However, this situation becomes harder when model development is collaborated by multiple researchers. The GEM for *S. cerevisiae* aims to be developed and maintained in a community, urgently demanding an approach to coordinate this problem. Here, I introduce the online tracking system Git repository, which was used to store and keep track of the model development.

A pipeline to record changes was developed based on the version control system Git and its hosting service GitHub. All files related to the model updates, including the dataset, scripts, and iterative releases, are recorded in the GitHub repository. In order to benefit the parallel development by multiple researchers, a branch system is adopted in the repository (**Figure 6**). Modifications are incorporated in the different feature branches. Once completed, the changes made in the feature branch can be merged back to the development branch and further to the main branch. A senior researcher reviews each feature branch before being merged to the development branch to ensure that the model has been modified as intended. Using this system, researchers can work on their feature branch without interfering, which promotes simultaneous development. Model users can download and use models from the main branch, where official updates are announced for new model versions.

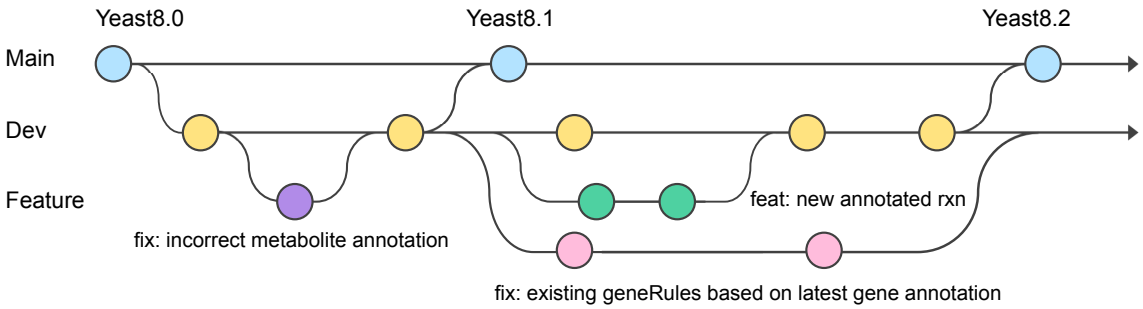

Figure 6 Diagram exemplifying the parallel development. Model users use the model in the main branch, while the model developers open different feature branches to curate the model. This branch system benefits the parallel development by multiple developers, which has been shown as a great platform in the *S. cerevisiae* GEM development.

This system has additional advantages for improving communications between model users and developers. Previously, when model users identified an error or bug in the model, users had to modify it themselves or communicate to the authors for modification, which caused inconvenience and could not automatically disseminate these curations to other users. Through the issue report system of GitHub, model users can directly report issues to model developers, promoting a quicker response and update. Besides that, since those issues are open, open discussions and even commits to fix the error can be submitted by model users or developers, promoting the community development. In the last two-year development, 12 researchers made commits to the updates of the *S. cerevisiae* GEM, 62 issues raised by the model developers and outside users were fixed and updated to the model, and in total 24 model versions were periodically released in GitHub.

This Git system has been applied to almost all model development in our group and it has proved to be an efficient system for transparent, reproducible and open research. I envision this system would become the state-of-art fashion and standard for the continuing model development.

2.2 Updates of *S. cerevisiae* GEM

To catch up with increased understanding, such as improved genome annotations and increased experimental phenotype data, GEMs need to be updated accordingly. Until we took on the task of hosting the *S. cerevisiae* community GEM, the latest model version, Yeast7 [69], accounted for 909 metabolic genes and was falling behind the latest genome annotation, presenting a bottleneck of usage of this GEM for systematic analysis.

The resulting updates for the *S. cerevisiae* GEM were divided into several rounds (**Figure 7**). Firstly, to improve gene coverage, additional gene annotations from the previous GEM iSce926 [70] were added after manual check.

Secondly, new annotations from five main databases (SGD [71], BioCyc [72], Reactome [73], KEGG [74], and UniProt [75]) were merged and incorporated into the model after manual check. In this round of model update, 48 original reactions in Yeast7 were updated by expanding existing GPRs with newly annotated genes. Meanwhile, 183 new reactions with 163 genes were added.

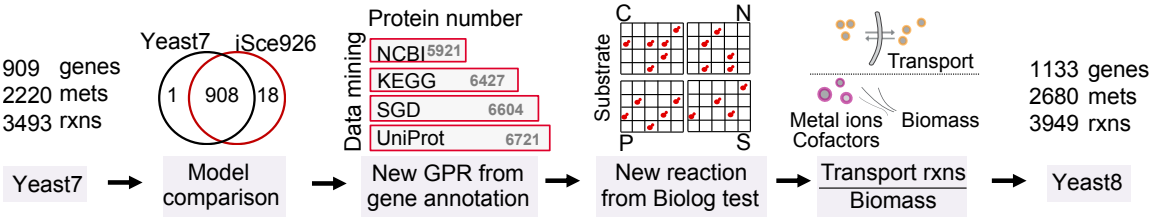

Figure 7 Major steps of development from Yeast7 to Yeast8. Starting from Yeast7, other GEMs for *S. cerevisiae* and new annotations in multiple databases were incorporated to improve the model scope. Substrate utilization experiments were used to curate the model especially the substrate degradation pathways. Besides that, transport reactions were reannotated to eliminate the reactions without gene annotation. Lastly, cofactors and metal ions were included in the biomass equation to activate more essential reactions in the growth simulation.

Thirdly, a gap-filling process was performed by investigating substrate utilization. The substrate utilization experiment was performed on Biolog Phenotype MicroArrays for 190 carbon, 95 nitrogen, 59 phosphorus, and 35 sulfur sources of two strains of *S. cerevisiae*, namely S288C and CEN.PK 113-7D. *S. cerevisiae* S288C can utilize 20 carbon, 40 nitrogen, 48 phosphorus and 19 sulfurous substrates, while CEN.PK 113-7D can utilize additional eight carbon and four nitrogen substrates. The experimental data were used to validate the *S. cerevisiae* GEM. Necessary reactions to fix the inconsistency between the model prediction and experimental data were added to improve the model quality. These reactions were extracted from the MetaNetX database [76] following the rule to introduce the least number of new metabolites and reactions. MetaNetX database was chosen due to the high coverage of metabolite and reaction ID association in Yeast7. In selected cases, the reversibility of existing reactions was modified in the curation process. In this step, a total of 225 new reactions and 148 new metabolites were added.

Fourthly, the transport reactions were updated. As a eukaryote, *S. cerevisiae* contains multiple organelles such as mitochondria and peroxisome. Therefore, to connect the metabolites between multiple compartments and generate a functional multi-compartment model, transport reactions were added previously [77], but many of these reactions did not contain associated gene annotations. To improve this, transporter annotations for *S. cerevisiae* in the TCDB transporter database [78] and the previous GEM issace [79] were collected, which is an automatically constructed GEM from the pan GEM of fungi species, containing more transporter annotation than Yeast7. In addition, the *S. cerevisiae* genome was reannotated through the EggNOG database [80] for transporter identification. After combining all those sources and careful manual check, 101 transport reactions were assigned with newly identified associated genes.

Lastly, the biomass equation was updated. While high-fidelity determination of yeast biomass composition has been reported [81]–[84], the biomass equation of *S. cerevisiae* GEM has not been updated in a long time. Thus, a new version of the biomass equation was formulated by accounting for cofactors and metal ions, which activate more reactions such as cofactor synthesis when simulating growth and thereby benefitting essential gene prediction.

Until here, I have reported major updates that were done to curate Yeast7.6 to Yeast8.4. In each round of model updates, standard quality-control tests for growth simulations, reaction mass balance check and NADH/ATP yield on glucose were performed to ensure model's functionality. All these updates, corresponding datasets and scripts to reproduce those changes are well documented in the GitHub repository. This study is not the end of the development for the *S. cerevisiae* GEM. Since **Paper I** was published, many new updates have been incorporated, including the improvement of the reaction stichometry balance and gap-filled reactions in aroma compounds production. There are remaining

issues by model users which will be continuously developed and integrated to the model in the future.

2.3 Evaluation of *S. cerevisiae* GEM

Through the above-described development, the updated GEM Yeast8 has a significantly increased model scope in terms of the numbers of genes (from 909 to 1133) and reactions (from 3493 to 3949) compared with Yeast7.6. Yeast8 was used to predict growth rates under different conditions to confirm whether the updates affected model performance. By comparing with experimental measurements, I found that Yeast8 performs well (**Figure 8**).

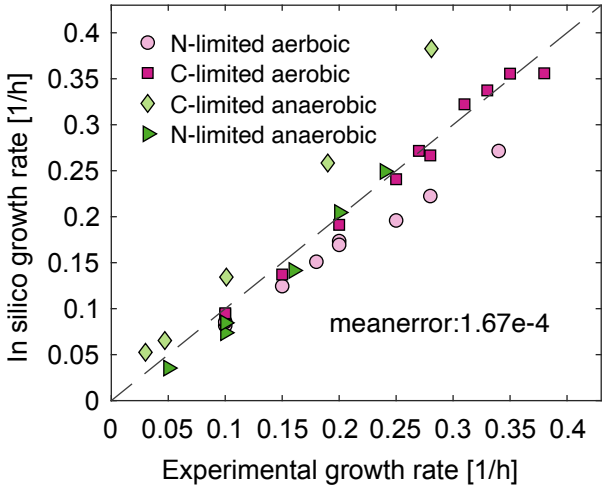

Figure 8 Growth simulation with Yeast8. Growth phenotype and related exchange rates of C-limited (carbon-limited) and N-limited (nitrogen-limited) chemostats were collected from literature. The model simulation was set up with minimal media with growth maximization as the objective function. Note that the biomass composition was adjusted for simulating N-limited conditions.

While Yeast8 has improved model scope compared with Yeast7.6, Yeast8 also outperforms Yeast7.6 in predicting some yeast phenotypes. Firstly, the accuracy of essential gene prediction is increased from 85.4% to 90.3% with an increase in the true positive (from 687 to 973) from Yeast7.6 to Yeast8. Next, Yeast8 outperforms Yeast7.6 in predicting substrate utilization with model prediction accuracy improving from 63.4% to 81.5% (**Figure 9a**). Furthermore, the memote scores [85], which demonstrate the model quality, also improved after the iterative updates, suggesting that the updated model has a more complete annotation for metabolites and reactions which would generalize the model usage and facilitate comparison with other models (**Figure 9b**). In conclusion, Yeast8 has shown its good performance and therefore can serve as the foundation for further expansion in two dimensions as described in the following chapters.

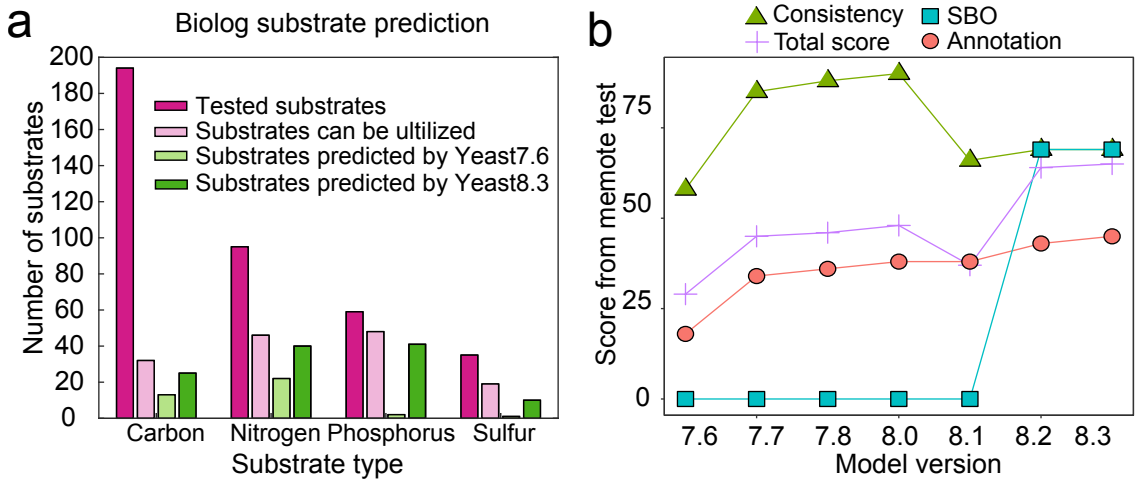

Figure 9 Comparison of Yeast7 and Yeast8 in simulation. a) Substrate usage comparison. Substrate utilization experiments were performed in the Biolog plates. The model simulation was set up with minimal media with the replacement of the corresponding tested substrates. b) Test scores based on the Memote test suite. The total score indicates the model annotation condition and whether the model reconstruction follows certain criteria. Consistency represents the score for the reactions as for the stochiometric consistency, mass and charge balance and unbounded flux carried by the reactions. Annotation represents the score for the completeness of universal ID annotation from standard databases. SBO stands for Systems Biology Ontology, which indicates the intended function of individual components, such as SBO:0000176 stands for biochemical reaction and SBO:0000247 stands for simple chemical metabolites. The decrease of score in Yeast8.1 is caused by pseudo metabolites introduced in the addition of SLIME reactions [86].

3. Yeast model development for more strains/species

There are two common approaches in GEM reconstruction: one is to reconstruct from scratch using the genome annotation from database as the scaffold; the other is to reconstruct the model based on homology search with a well-curated GEM as the template. Using the template to reconstruct GEMs allows to transfer the highly curated knowledge bases for one organism, in *S. cerevisiae* case, over 15 years development, quickly to another phylogenetically close organism based the assumption that homolog shares similar function. Due to the rapid homolog gene mapping tools for multiple organisms developed in these years, it has now become possible to utilize the second approach for large-scale GEM reconstruction. In this section, I introduce how the well-curated GEM for *S. cerevisiae* Yeast8 was used as basis to model more yeast species/strains. This chapter is divided into two parts: to model multiple *S. cerevisiae* isolates (**Paper I**) and to model multiple yeast species (**Paper II**).

3.1 Modeling 1,011 *S. cerevisiae* isolates

S. cerevisiae can be found worldwide in diverse habitats including natural biotopes and human-associated habitats. There is an increasing number of published genome sequences of *S. cerevisiae* isolates, which were demonstrated to have a high level of genetic diversity [87]–[89]. In order to systematically study the metabolic diversity, I reconstructed GEMs for 1,011 *S. cerevisiae* isolates from diverse origins [87] using Yeast8 as the template.

3.1.1 Reconstruction of GEMs for 1,011 *S. cerevisiae* isolates

Firstly, the pan-genome, which represents the complete representative gene sets (pan-genes) of 1,011 *S. cerevisiae*, was defined and annotated with different databases such as KEGG [74] and EggNOG [80] web services. The annotated KEGG Orthology (KO) for pan-genes were mapped to metabolic reactions, and then compared and merged with Yeast8 to form the panGEM containing comprehensive metabolic reactions for these 1,011 isolates.

The panGEM and the strain&pan-gene matrix containing the pan-gene existence information in each strain were used to generate GEMs for 1,011 isolates (**Figure 10**, **Figure 11a**). The number of reactions in those GEMs range from 3,396 to 4,013 (**Figure 11b**). Followingly, I generated coreGEM by extracting the reactions, metabolites, and pan-genes shared among all isolates, yielding 3,905 reactions, 2,666 metabolites and 892 pan-genes. Many reactions in Yeast8 are shared by most isolates, signifying that metabolism is well conserved among all those isolates. Notably, the number of accessory genes (478) that are not shared by all isolates is higher than the number of accessory reactions (147), which indicates many isozymes emerged during the evolution, demonstrating the robustness of the cellular function. Most of the accessory genes were associated with sugar and secondary metabolism, suggesting that isolates have diverse adaptive evolution towards habitats.

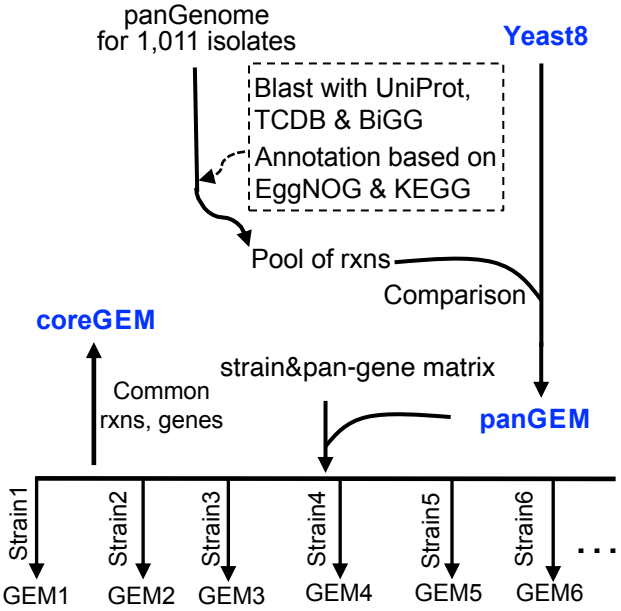

Figure 10 Reconstruction pipeline of GEMs for 1,011 isolates. The GEM for *S. cerevisiae* was used as the template to formulate the panGEM for all those *S. cerevisiae* isolates. The reconstructed panGEM combined with the strain&pan-gene matrix (the pan gene existence information in each isolate) was used to extract the strain-specific models for each isolate. coreGEM for those isolates were extracted by identification of the common reactions among GEMs of all isolates.

3.1.2 Evaluation of GEMs for 1,011 *S. cerevisiae* isolates

The 1,011 GEMs were used to estimate substrate usage and yield of biomass precursors (**Figure 11c-d**). The figure was plotted by grouping isolated from the same origin.The substrate usage analysis suggested that the isolates from the origin 'Industrial' and 'Lab strain' are typically able to utilize fewer substrates compared with isolates from other origins (**Figure 11c**), which may hint that those isolates have gone through reductive evolution during their domestication, resulting in loss of the ability to utilize certain substrates. Simulated biomass yield differs in isolates from diverse origins (**Figure 11d**). The isolates from the origin 'Human' have a relatively lower biomass yield, likely due to an adaptive evolution towards associated habits. Instead of *de novo* biosynthesis of biomass precursors, the 'Human' isolates can take up some of them from their environment as their human host is typically a rich source of such nutrients.

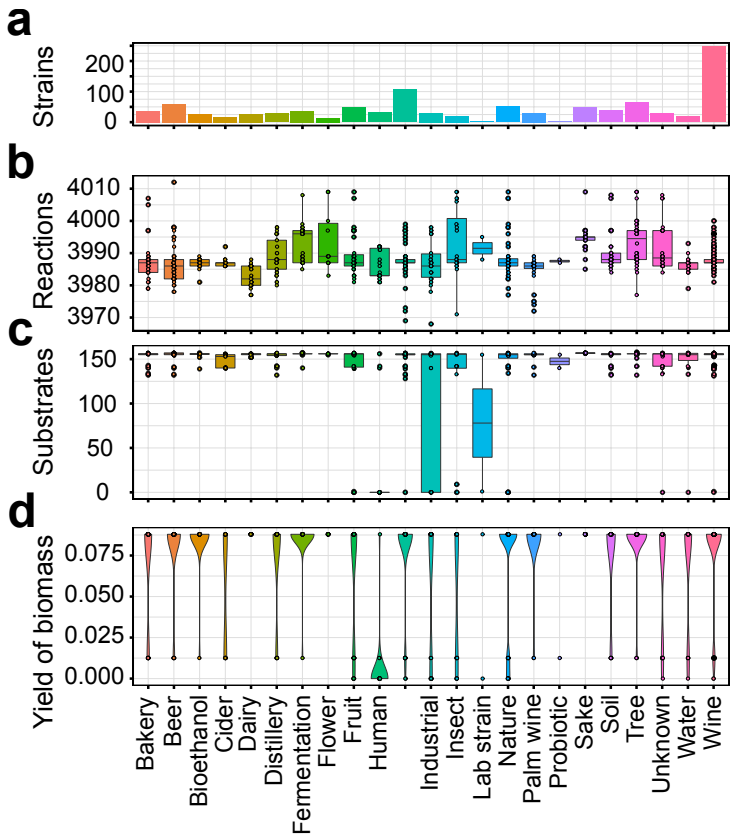

Figure 11 Evaluation of GEMs for 1,011 isolates. a) Number of isolates in each origin. b) Box plot for the reactions in the reconstructed GEMs for isolates from each origin. c) Simulated substrate usage profiles for isolates classified by origin. This test was performed using the same approach as the substrate utilization test for the *S. cerevisiae* as in Figure 9a. d) Box plot for the biomass yield on glucose with the minimal medium set up.

To further compare the metabolic potential of those isolates, the maximum yields of 20 amino acids plus six important biomass precursors were computed using 1,011 GEMs (**Figure 12a**). The results suggested that those yields differ among isolates. **Figure 12b** shows that the optimal yields of amino acids differ in isolates from the 'Industrial' origin. Combining the genotype information with the collected phenotype data, I identified that the variations are primarily caused by the energy production efficiency and the missing essential reactions in the amino acid biosynthetic pathways. For example, *S. cerevisiae* isolate AAH has a simulated low overall amino acid yield as it produces energy by fermentation rather than respiration. The experimental data showed that this strain has very slow growth on non-fermentable carbon sources such as ethanol and glycerol [87], suggesting that the respiratory pathway may be impaired. The consistency between the model prediction and the experimental report is indicative of the model quality.

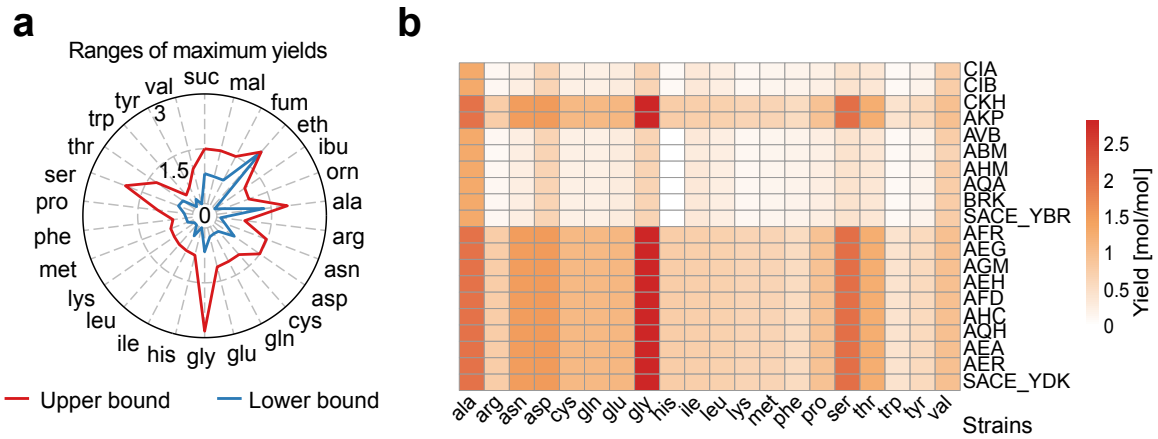

Figure 12 Simulated amino acid yields of GEMs for 1,011 isolates. a) Comparison of simulated range of maximum yields of 20 amino acids and six key chemicals for all 1,011 GEMs. b) Comparison of the yield of 20 amino acids from glucose for 20 strains from the group of "industrial strain" to exemplify the diversity of amino acid yields.

Subsequentially, the 1,011 GEMs were classified based on the *in silico* substrate usage, given that the ability to utilize different substrates may have a direct link towards the diverse origins (**Figure 13**). The isolates from the 'Human' origin can be separated from the isolates originating from 'Wine' with only two substrates (lactic acid and sorbitol, **Figure 13**), demonstrating that substrate usage profiles are indeed affected by the strain origins and can be used to for classification.

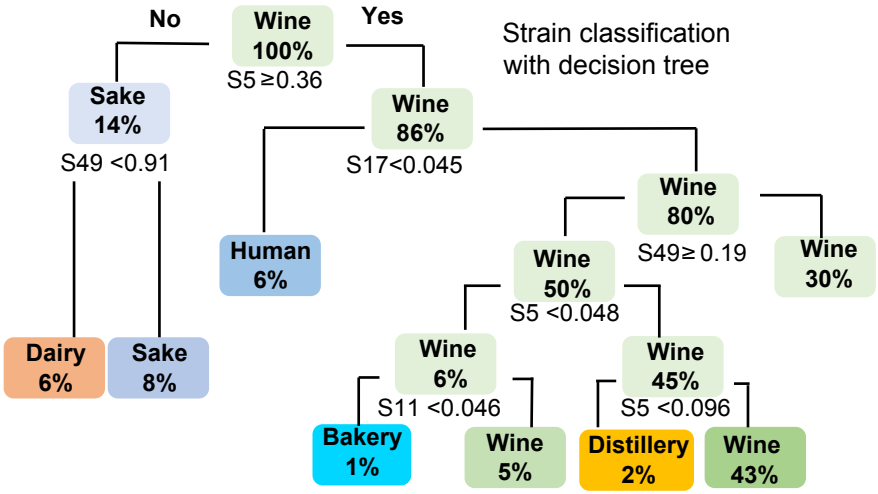

Figure 13 Decision tree classification of strains according to the simulated maximum growth rates on different carbon sources. The growth rate result was generated in the substrate utilization test as mentioned in the Figure 11c. S5: Latic acid; S11: Serine; S17: Sorbitol; S49: Trehalose.

3.2 Modeling 343 yeast/fungi species

Yeast species have been widely used as cell factories, while more non-conventional yeasts have drawn attention due to their fascinating and versatile phenotypes [90]. There are over 1,500 identified yeast species, evolved with over 400 million years [91]. It would be desired for systematic analysis of their metabolic diversity, and to this purpose, GEMs were reconstructed for diverse yeast species, including 332 yeast species spanning 12

phylogenetic clades plus 11 outgroup fungi species. **Figure 14** shows the species number from each phylogenetic clade and its genome size.

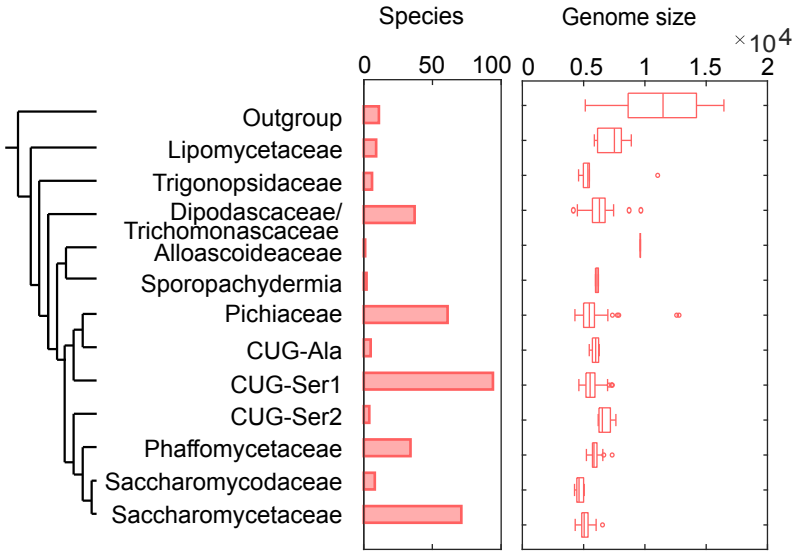

Figure 14 General information for 332 yeast species and 11 fungal outgroup species. The tip names in the phylogenetic tree represent 12 major clades in the subphylum classification for 332 yeast species plus 11 fungal species as the outgroup according to [91]. This classification is used in all following analyses.

3.2.1 Reconstruction of GEMs for 343 yeast/fungi species

The panGEM for all those 343 yeast/fungi species was reconstructed using Yeast8 as the template, in an approach similar to what was employed to reconstruct GEMs for 1,011 *S. cerevisiae* isolates (**Figure 15**). Compared with the 1,011 *S. cerevisiae* isolates, 343 yeast/fungi species have a more considerable phylogenetic distance towards *S. cerevisiae*, which suggests that more additional genes and reactions would be added into Yeast8 to form the panGEM. Therefore, to improve the model coverage for the accessory reactions, individual draft GEMs for 343 yeast/fungi species were generated through RAVEN2 toolbox [21] from KEGG and MetaCyc annotation. The reactions from the draft GEMs and detailed pan-gene annotations were incorporated together as the reaction pool. After several rounds of cross-referencing, filtering, and manual check, 562 new reactions and associated genes were added into Yeast8 to form the panGEM for 343 yeast/fungi species. The panGEM includes 3143 metabolites, 4587 reactions and 3751 pan-genes, which is a more complete metabolism representation than a prior fungal panGEM reported in the literature [79].

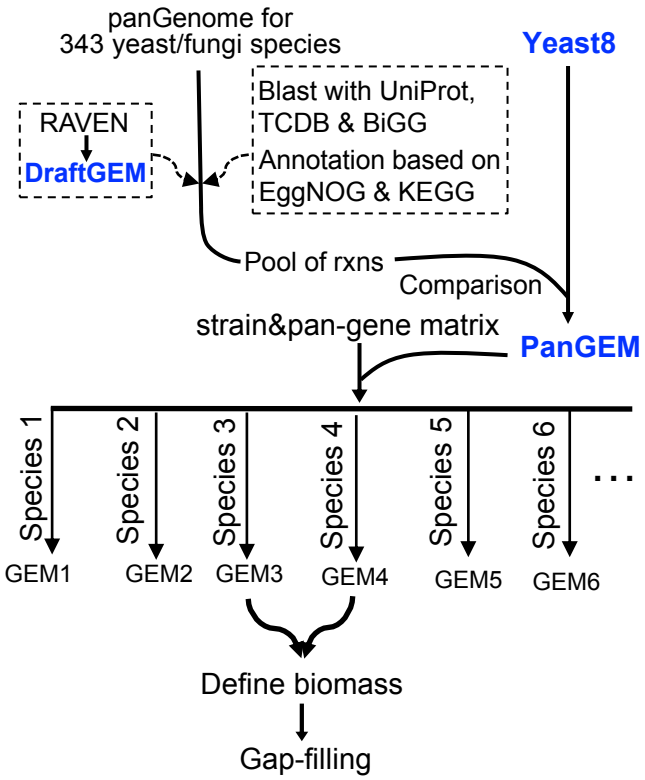

Figure 15 Reconstruction pipeline for GEMs of 343 yeast/fungi species. Yeast8 was used as the template together with reactions annotated from the pan-genome to formulate the panGEM for 343 yeast/fungi species. The reconstructed panGEM combined with the strain&pan-gene matrix (indicates the pan-gene existence in each species) were used to extract the species-specific GEMs. To represent the diverse phenotypes, biomass composition was defined in GEMs for 343 species based on their phenotypes. Lastly, to ensure the functionality of those GEMs, gap filling was performed, including a biomass precursor targeted pathway search check and an automatic gap-filling.

Then, the GEM for each yeast/fungi species was generated based on the species&pan-gene matrix, which contains the pan-gene homolog existence in each species. Unlike the reconstruction of GEMs of 1,011 *S. cerevisiae* isolates, a less strict strategy was adopted when defining the existence of enzyme complexes. If at least half of all subunits of an enzyme complex were present, the complex was considered functional, and the corresponding reaction would be kept in the species-specific GEMs. This step is essential when generating GEMs for phylogenetically distant species where enzyme complex configurations will be less conserved, ensuring that less gap-filling work will be required in the next step of species-specific GEM reconstruction.

Biomass equations were formulated for species with diverse phenotypes. According to the collected phenotypes, the 332 yeast species were split into four groups: 'Normal', 'Heat-tolerant', 'Oleaginous' and 'Pathogenic'. Species that could not be assigned to a certain phenotype were classified as 'Normal'. For each group, one biomass composition was defined according to previous published GEMs for the representative species.

Due to significant phylogenetic differences, GEMs for several yeast/fungi species contained gaps in the biomass precursor synthesis pathways and could not predict growth. Therefore, two rounds of gap-filling processes were performed. Firstly, the gaps were

identified by pathway search for dead ends in biomass precursor production, which were subsequently filled with reactions from the draft GEMs of the corresponding species. To ensure the quality of the models, I only added reactions with gene annotations in this step. Secondly, if gaps remained after the first step, an automatic gap-filling approach in the RAVEN2 toolbox was performed, where the panGEM was used as the reaction pool. The criteria were to add the least number of reactions into the GEM to achieve growth.

In order to improve the prediction of substrate utilization, additional steps for substrate metabolism annotation were performed using tblastn search. Template reactions and corresponding genes of substrate metabolism were collected from KEGG, MetaCyc and literature. This step significantly improved genome annotation for the substrate utilization pathways. For example, erythritol degradation has not been fully elucidated for most yeast species [92]. Through the tblastn search for enzymes involved in two erythritol degradation pathways from MetaCyc, the erythritol degradation pathway II is more likely to be involved in yeast species and the enzyme EC 5.3.1.34 was identified as the key enzyme for erythritol degradation (**Figure 16**). From the 85 yeast species with the genomic evidence for this enzyme, 68 yeast species were reported to have the erythritol phenotype, while 11 out of the remaining 17 species lacked experimental evidence. The consistency between enzyme existence and experimental evidence of the phenotype suggests that the enzyme is the key enzyme controlling the erythritol phenotype in yeast species.

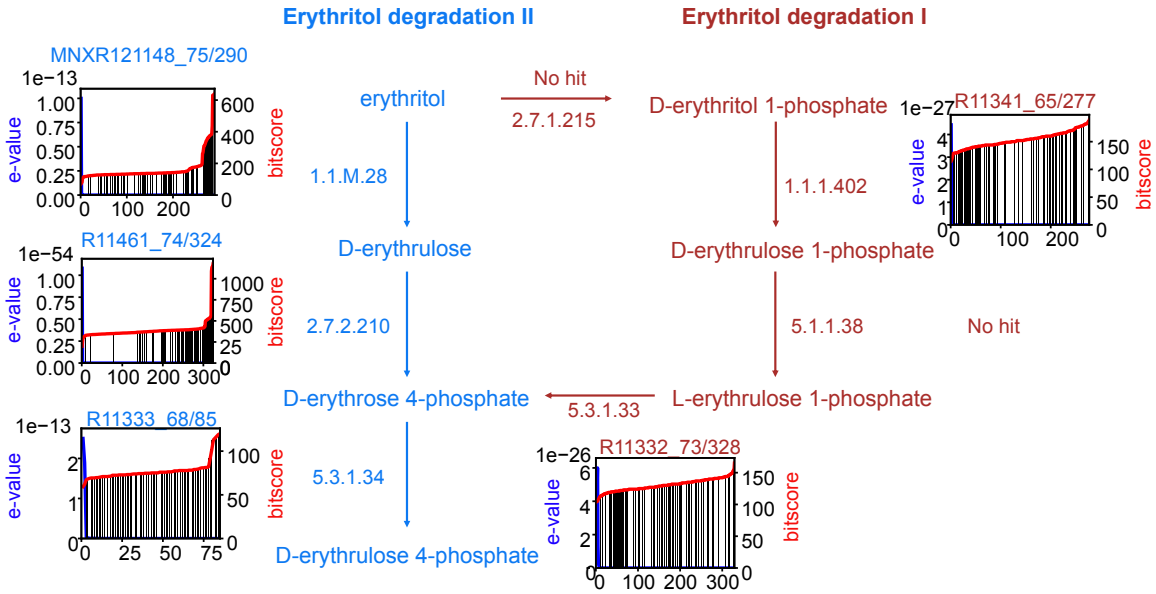

Figure 16 Gene mining result for the erythritol degradation pathways. The number above each plot indicates the ratio of the consistency species number and species number with the gene evidence from tblastn result.

After such curations, the reconstructed 343 yeast/fungi species GEMs contained 3,500-4,000 reactions and around 1,000 genes. There were 2,080 reactions shared by all yeast species, i.e., core reactions, and 2,519 accessory reactions (**Figure 17a-b**). The core reactions were involved in central metabolism, while the accessory reactions were mainly in the secondary metabolism. To check whether these GEMs reflect phylogenetic relations

among these species, I performed t-distributed stochastic neighbor embedding (t-SNE) analysis based on the reaction existence information, which shows that the GEMs can largely be clustered based on their phylogenetical clade, suggesting the conservation of network structure within clades. This clustering also suggests that the metabolic diversity could partially capture the evolutionary relationship (**Figure 17c**).

It should be noted that this template-based large-scale GEM reconstruction process improves the quality of Yeast8. For example, the degradation reaction from D-glucosamine 6-phosphate to fructose 6-phosphate catalyzed by glucosamine-6-phosphate deaminase is an essential reaction for utilization of N-acetyl-D-glucosamine. This reaction was present in a previous version of Yeast8 but has been removed as it was found to be wrongly annotated in the previous version. After that, the model simulated zero growth with N-acetyl-D-glucosamine as a carbon source, consistent with the experimental evidence of this phenotype [93]. In addition, 15 genes were annotated to 14 existing metabolic reactions of Yeast8, thereby improving the gene coverage.

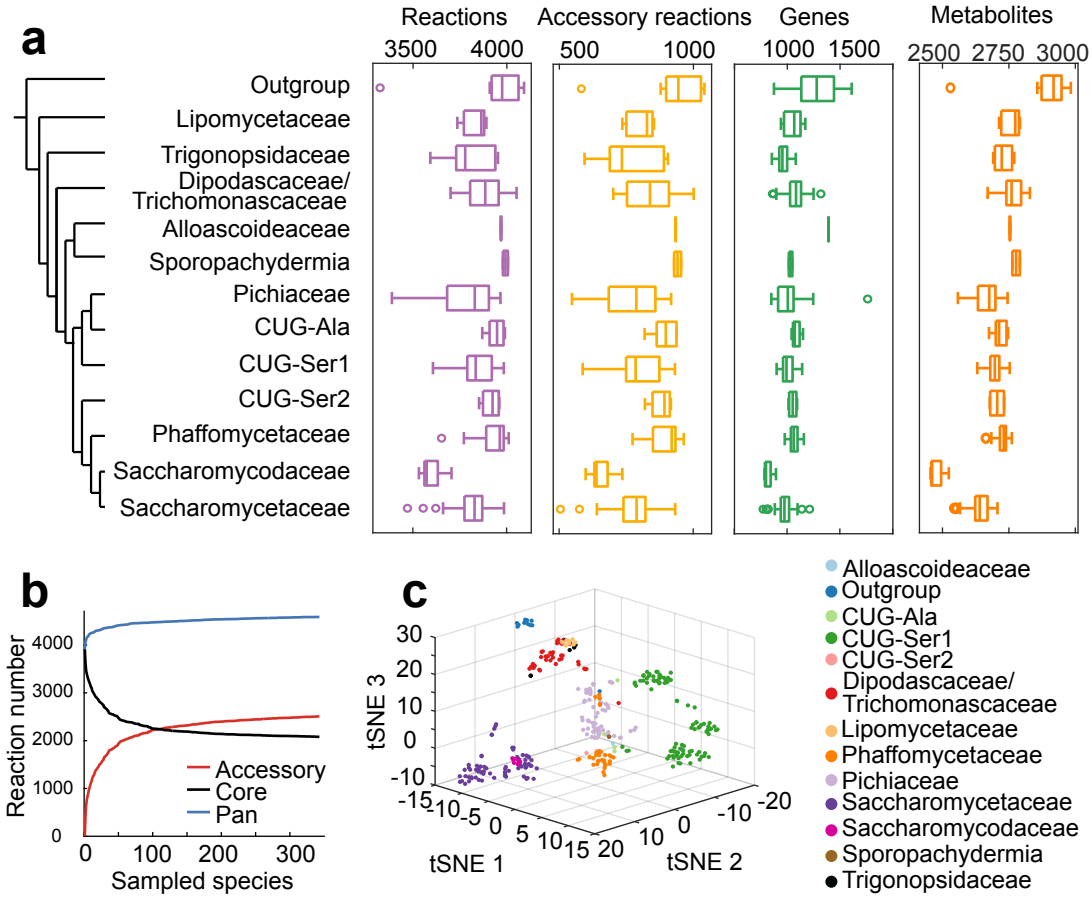

Figure 17 Comparison of reconstructed GEMs for 343 yeast/fungi species. a) Comparison of GEMs for species from 12 major clades regarding the numbers of reactions, accessory reactions, genes and metabolites. b) Reaction numbers for the core, pan and accessory reactions as sampling from 1 to 343 species. c) t-SNE analysis of yeast species based on the rection existence matrix of 343 yeast/fungi GEMs.

3.2.2 Evaluation of GEMs for 343 yeast/fungi species

The 343 GEMs were firstly evaluated by their substrate utilization profiles. As for collected substrate profiles for 322 yeast species, the average accuracy was above 0.75 (**Figure 18a**), suggesting the high quality of the reconstructed GEMs for substrate utilization prediction. The false positives could potentially primarily be the result of promiscuous enzymes in GEMs. For example, even though *S. cerevisiae* contains annotated enzymes for xylose degradation, it cannot utilize xylose as a sole carbon source due to the low kinetic activity of the promiscuous enzymes involved [94].

Secondly, essential gene prediction was performed using GEMs for five species with experimental essential gene profiles. The result showed an overall 0.78 accuracy for essential gene prediction, which is comparable with previously published well-curated GEMs for those species (**Figure 18b**). The high accuracy furthermore demonstrated the high quality of GEMs generated from this pipeline. The scope of the reconstructed GEMs was also compared with the previously published GEMs for the same species, demonstrating comparable numbers of reactions, metabolites and genes (**Figure 18c**).

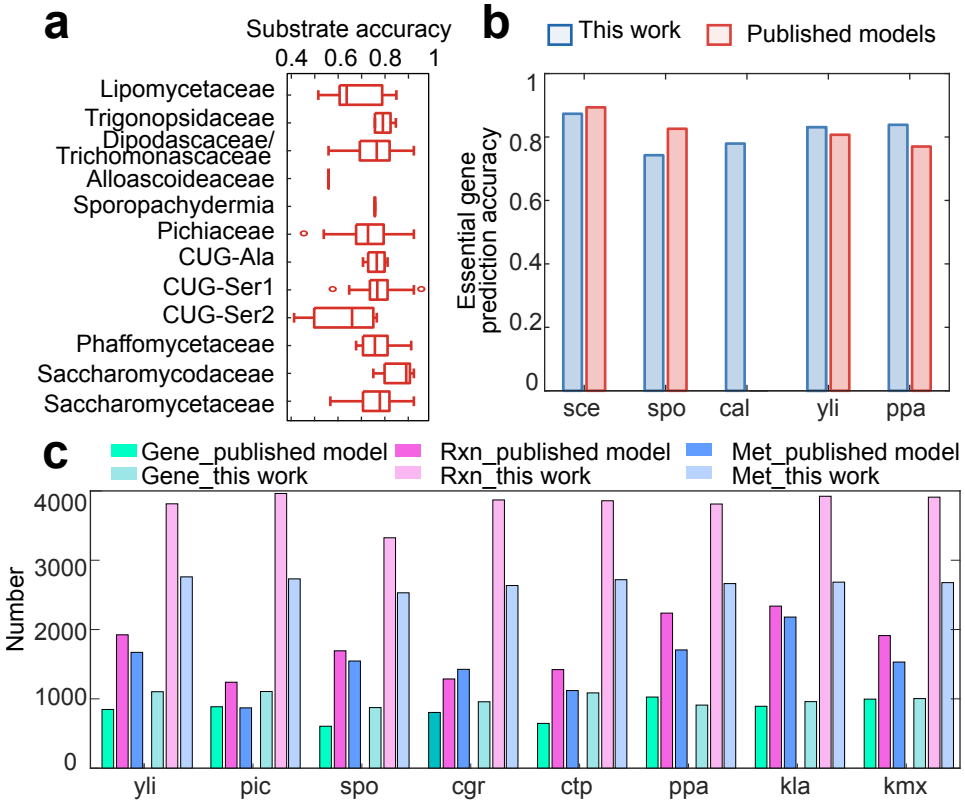

Figure 18 Evaluation of GEMs for yeast/fungi species. a) Comparison of substrate utilization prediction accuracy. Substrate utilization experimental phenotype was collected from literature [91], [95]. Model prediction was simulated using GEMs with the minimal media setup. Substrates were changed accordingly. The accuracy was calculated as (TP + TN)/ (TP + TN + FP + FN). b) Comparison of essential gene prediction for five species of models from this study and published models. The prediction accuracy for SpoMBEL1693 of *S. pombe* is from [96] and others were re-calculated in this work using the corresponding GEMs. c) Model scope comparison of GEMs in this study and previously published GEMs. Published GEMs used in this analysis: iYali4 for *Y. lipolytica* [30], iNX804 for *C. glabrata* [33], iSM996 for *K. marxianus* [32], iTL885 for *Scheffersomyces stipites* [97], SpoMBEL1693 for *Schizosaccharomyces pombe* [96], iCT646 for *Candida tropicalis* [98], iMT1026.v3 for *Komagataella pastoris* [99], iOD907 for *K. lactis* [31]. yli: *Y. lipolytica*; pic: *S. stipitis*; spo: *S. pombe*; cgr: *C. glabrata*; ctp: *C. tropicalis*; ppa: *K. pastoris*; kla: *K. lactis*; kmx: *K. marxianus*; sce: *S. cerevisiae*; cal: *Candida albicans*.

After demonstrating that the reconstructed GEMs were of overall high quality, the GEMs were used to evaluate the metabolic diversity among those species. The yields of ATP, amino acids and biomass on glucose were calculated using these 343 GEMs (**Figure 19a**). The result showed that species from the clade *Saccharomycodaceae* and *Saccharomycetaceae* have a lower yield of ATP and biomass (**Figure 19a**), which could be due to the absence of Complex I in the respiratory chain and thus a decrease in the proton motive force for energy production. This is consistent with experimental observations that yeast species with Complex I have a relatively lower biomass yield than those with Complex I [100]–[104]. The maximal theoretical amino acid yields were also calculated with those GEMs, which also differ among those species (**Figure 19b**).

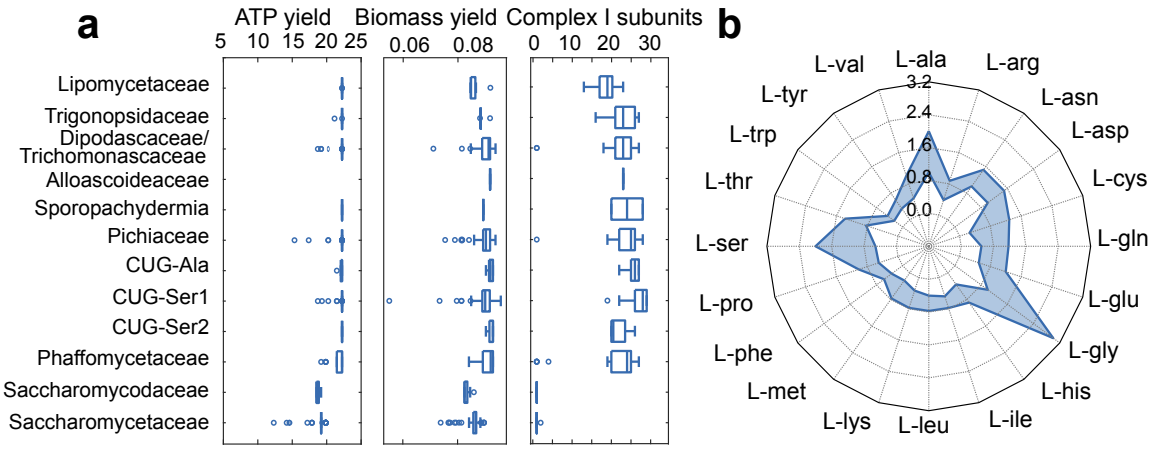

Figure 19 Metabolic diversity comparison of 343 yeast/fungi GEMs. a) Comparison of simulated ATP and biomass yields on glucose, and numbers of Complex I subunits for species from each clade. b) Range of simulated amino acid yields on glucose for all species. The result was calculated using GEMs with minimal media as the setup with glucose as the carbon source and amino acid exchange as the objective, respectively. The unit of yield is mol/mol.

3.2.3 Trait evolution analysis aided by GEMs for 332 yeast species

Since those 332 yeast species GEMs showed good predictions of substrate utilization (**Figure 18a)**, the mechanisms underlying substrate diversity could be investigated by complementing the GEMs with analysis on evolutionary events. The substrate utilization of each species was firstly compared to their common ancestor BYCA (Budding yeast common ancestor) [91], and the substrate utilization loss and gain for each species were denoted (**Figure 20a**). Then the reactions and genes responsible for the substrates utilization were identified from all 343 GEMs. After that, detailed gene-level evolution analyses were conducted, identifying events such as horizontal gene transfer (HGT) and gene family expansion and contraction, and the results were mapped to the substrate utilization changes for each species. Promiscuous enzymes that might have evolved different kinetic activities towards diverse substrates were also analyzed and related to substrate utilization loss and gain (**Figure 20a**).

In **Figure 20b**, many HGTs related to substrate utilization occurred with enzymes that interface with the environment, such as transporters and extracellular degradation enzymes, suggesting that those HGTs may result from adaption to their habitat. Besides

that, many of the horizontally transferred genes were derived from other fungi, suggesting that gene flow is more frequent among fungi than between bacteria and fungi (**Figure 20c**). HGTs related to substrate usages are more abundant in selected clades such as *Wickerhamiella/Starmerella* (W/S clade) and its close relatives, e.g., *Lipomycetaceae*, *Trigonopsidaceae* and *Dipodascaceae*/ *Trichomonascaceae*, which may be because the species from those clades ecologically cohabit with other fungi species. There are two clades (CUG-Ser1 and CUG-Ser2) with non-HGT events for the substrate utilization, indicating that the genetic code alteration may act as a barrier for HGT.

As for gain of substrate utilization, gene expansion and promiscuous enzymes might be the main driving force (**Figure 20d**). For example, species from the clade *Saccharomycetaceae* went through a whole-genome duplication, which contributed to the presence of multiple paralogs in those species. Duplicated promiscuous enzymes may gain activity towards new substrates through divergent evolution, contributing to the broader substrate profiles in those clades [105].

Loss of reactions or attenuated enzyme activities could cause the loss of substrate utilization. Firstly, I divided the substrate utilization reactions into two parts: highly correlated reactions and non-highly correlated reactions, which indicated the coexistence of a reaction towards a particular substrate utilization phenotype (**Figure 20e**). Highly correlated reactions are defined as always present reaction when a particular phenotype is observed (consistency > 0.83, sensitivity > 0.92). Highly correlated reactions were identified in the corresponding utilization pathway for 14 out of 32 analyzed substrates, suggesting that losses of 14 substrate utilization phenotype among different species are mainly caused by the loss of same reactions. However, the loss of non-highly correlated reactions could also contribute towards the loss of substrate utilization ability (**Figure 20f**), demonstrating that the diverse reductive evolution routes exist for the same phenotype. Besides these two cases, occasionally, the substrate utilization phenotype is lost even while all enzymes and reactions linking the substrate to central carbon metabolism are present, which is caused by the loss of distant (or downstream) reactions. For example, several species from the genus *Hanseniaspora* lose the ability to utilize ethanol, which may be caused by the loss of an essential reaction in gluconeogenesis, where oxaloacetate carboxylase (EC4.1.1.49) catalyzes the conversion from oxaloacetate to phosphoenolpyruvate (**Figure 20g**). For the same reason, those species cannot utilize citrate or succinate, while glycerol utilization is not affected. These four substrate utilization phenotypes are consistent in the model simulation and experimental data. This scenario highlights the potential of GEMs in systematically analyzing cell metabolism.

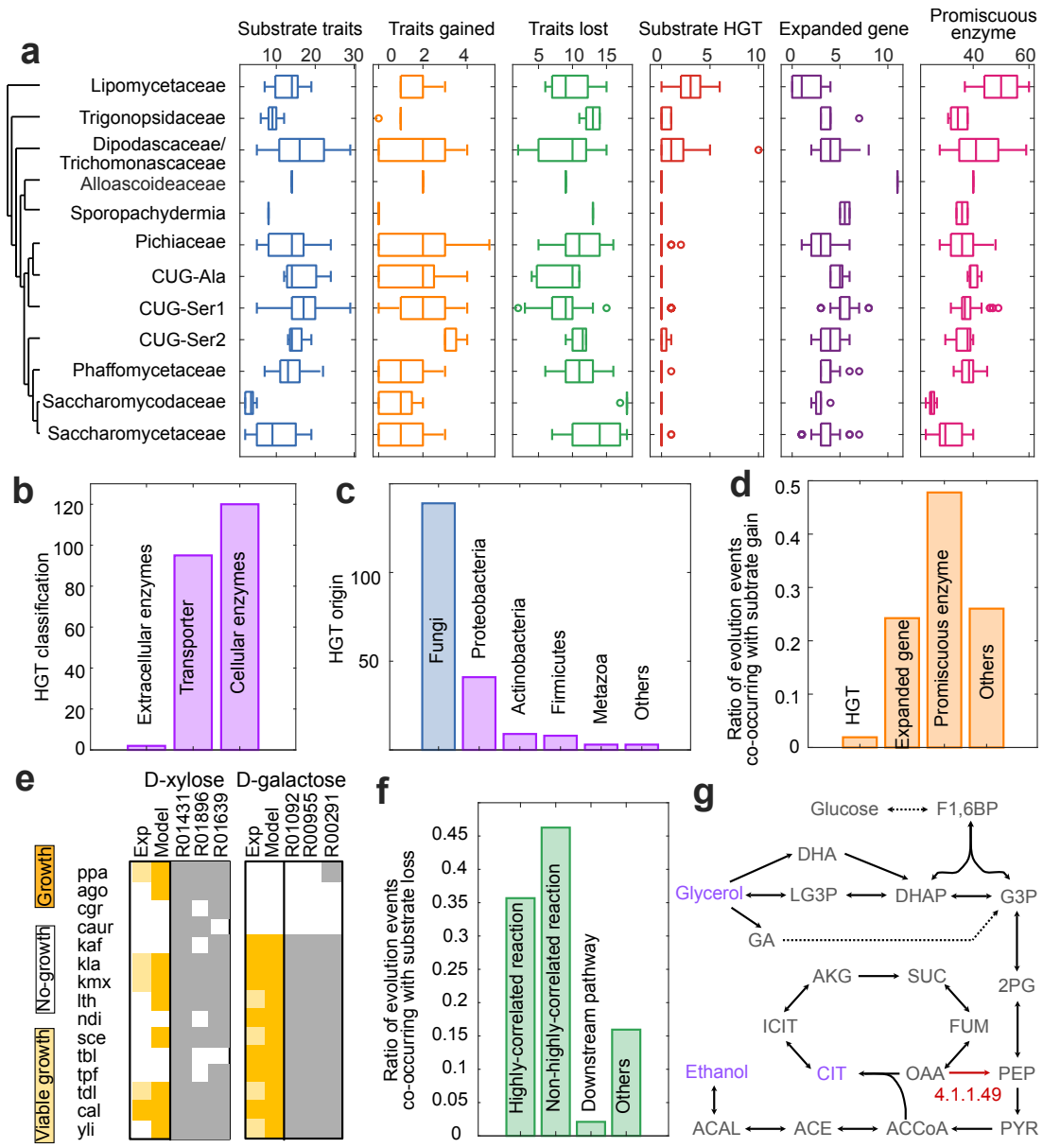

Figure 20 Investigation of substrate utilization diversity by GEM simulation and evolution analysis. a) Comparison of the numbers of utilized substrates, gain- and loss- of the substrate utilization compared with the ancestor BYCA, HGTs related to substrate utilization, expanded genes related to substrate utilization, and promiscuous enzymes in the substrate utilization pathway. b) Classification of the genes with HGT events in substrate utilization. c) Classification of doners for those HGT events in substrate utilization. d) Ratios of evolution events co-occurring with substrate gain. e) Example of highly correlated reactions and non-highly correlated reactions. Reactions for xylose degradation in the model correlate poorly with the existence of the xylose utilization phenotype since those enzymes are also responsible for other substrate degradations. Thus, those reactions were defined as non-highly correlated reactions. In contrast, the reactions for galactose utilization correlate well with the galactose utilization phenotype, which were classified as the highly correlated reactions (Highly correlated reactions are defined as consistency > 0.83, sensitivity > 0.92.). Exp: experimental phenotype. Model: model simulated phenotype. ppa: *K. pastoris*; ago: *Eremothecium gossypii*; cgr: *C. glabrata*; caur: *Candida auris*; kaf: *Kazachstania africana*; kla: *K. lactis*; kmx: *K. marxianus*; lth: *L. thermotolerans*; ndi: *Naumovozyma dairenensis*; sce: *S. cerevisiae*; tbl: *T. blattae*; tpf: *T. phaffii*; tdl: *Torulaspora delbrueckii*; cal: *C. albicans*; yli: *Y. lipolytica*. f) Ratios of evolution event co-occurring with substrate loss. Highly correlated and non-highly correlated reactions are defined by the consistency of reaction existence with the substrate utilization phenotype existence. Downstream pathway is defined as when all enzymes and reactions are included in the original substrate utilization pathway, but specific enzymes are absent in a distantly related pathway. g) An example of how GEM reconstruction could aid in the identification of the reason for the loss of traits caused by the gap from the downstream pathway. The loss of substrate utilization such as ethanol and citrate are caused by the downstream reaction's loss (catalyzed by the EC 4.1.1.49, marked by red). All substrate traits were marked in purple.

4. Yeast model development to more constraints/biological processes

GEMs enable the prediction of cellular metabolic states under certain external and internal constraints, thereby serving as an evitable tool for systems biology. However, biological processes in the living cell are highly interacted. Thus, to study more complex and delicate phenotypes, more biological processes and constraints should be incorporated into GEMs. In this section, I describe the work related to the integration of extra constraints and/or biological processes to the basic GEMs, which resulted in three types of models: ecGEMs (**Paper I** & **Paper II**), CofactorYeast (**Paper III**) and pcSecYeast (**Paper IV**).

4.1 ecGEMs for yeast species

Previously, GEMs for *S. cerevisiae* and 343 yeast/fungi species were reconstructed. To enhance their predictive power, I incorporate enzymatic constraints into those GEMs using the coarse-grained approach GECKO [46], resulting in ecGEMs for *S. cerevisiae* and diverse yeast/fungi species. Due to the limitation of available kinetic information, only the ecGEMs for *S. cerevisiae* and 14 relatively well-studied yeast/fungi species were generated. The development of ecGEMs for the remaining 300 yeast/fungi species are introduced in the next chapter (**Chapter 5**).

As for the generated ecGEM for *S. cerevisiae* ecYeast8, maximum growth rates on 332 different combinations of carbon and nitrogen sources were predicted and compared with the experimental data from microtiter platers (**Figure 21a**). The result shows that ecYeast8 can significantly reduce the error in prediction when compared with the basic GEM Yeast8. Besides that, the flux control coefficient (FCC) analysis, where the importance of specific enzyme towards the growth rate were defined, showed that glyceraldehyde-3-phosphate dehydrogenase (Tdh1p) is the major flux controlling enzyme for growth on medium with glucose as the carbon source, while the F0-ATP synthase subunit c (Oli1p) and isocitrate lyase (Icl1p) are major controlling enzymes for growth on ethanol and acetate (**Figure 21b**).

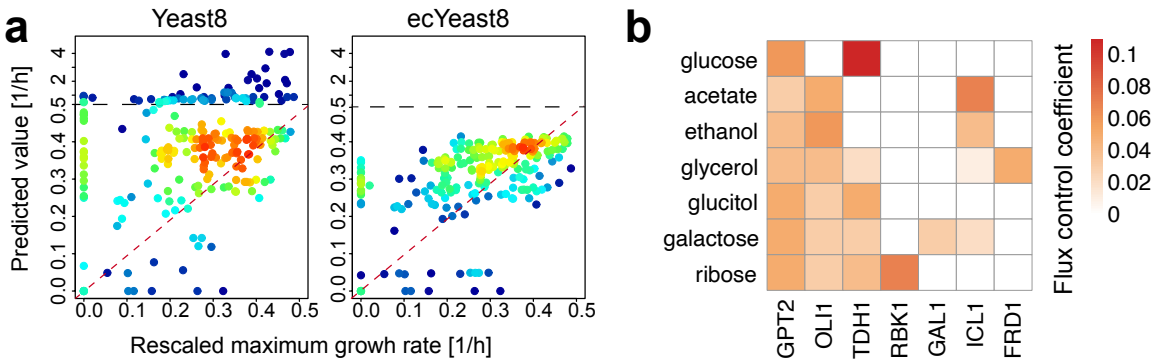

Figure 21 Growth simulation and analysis under different substrate sources. a) Prediction of maximum specific growth rates under 332 combinations of carbon and nitrogen sources using Yeast8 and ecYeast8. Red color denotes the high density of overlapping points. b) FCC analysis for growth under different carbon sources. FCC quantifies the effect of 0.1% change in k_{cat} of each enzyme on growth.

As for the generated ecGEMs of 14 yeast/fungi species, the coverage for enzymatic constraints was relatively low compared to the total metabolic enzymes that were specified in the GEMs (~1,000) (**Figure 22a**), which could introduce bias in simulations. I will discuss this uncertainty in detail in **Chapter 5**. The FCC analysis was also performed for growth on minimal media with glucose as the carbon source. The result suggests that the enzymes with high FCCs are largely consistent among those 14 yeast species, demonstrating that the central carbon metabolism is largely conserved (**Figure 22b**).

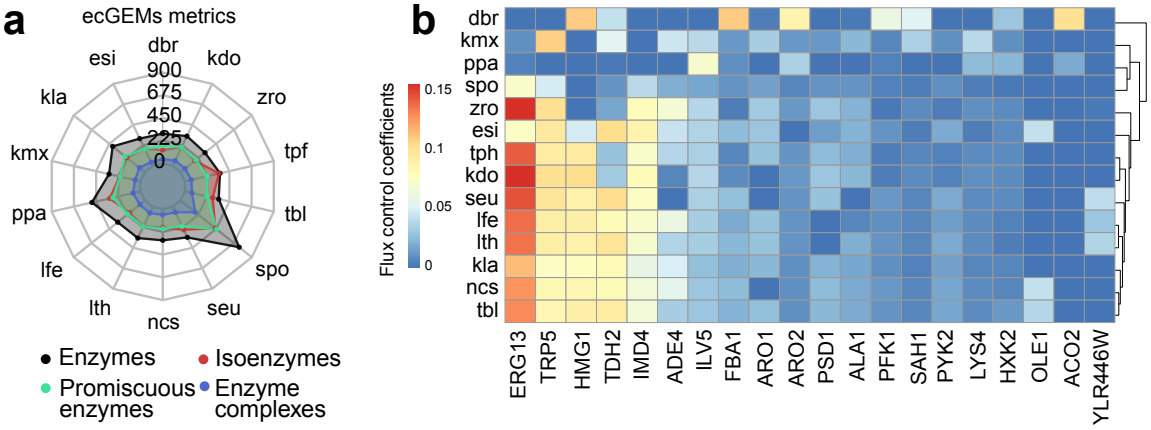

Figure 22 Analysis of ecGEMs for 14 yeast species. a) Comparison of enzymatic constraints of 14 ecGEMs. b) Heatmap of FCCs calculated by ecGEMs. The x tick label represents the gene names. dbr: *Dekkera bruxellensis*; esi: *Eremothecium sinecaudum*; kla: *K. lactis*; kmx: *K. m*arxianus; ppa: *K. pastoris*; lfe: *Lachancea fermentati*; lth: *L. thermotolerans*; ncs: *Naumovozyma castellii*; seu: *Saccharomyces eubayanus*; spo: *S. pombe*; tbl: *T. b*lattae; tpf: *T. phaffii*; zro: *Z. rouxii*; kdo: *Kluyveromyces dobzhanskii*.

In this part, I introduced that basic GEMs can be extended to ecGEMs by adding enzymatic constraints, but these extended models cover the same scope of biological processes as the basic GEM, solely describing metabolism. I also showed that there is a limitation of generalization of ecGEM reconstruction towards less-studied species.

4.2 Modeling enzyme cofactor: CofactorYeast

Metal ions are key cofactors for enzymes to ensure proper functioning. While numerous enzymes have been identified to interact with metal ions, there is a lack of knowledge on quantitative relationships between metal ions and metabolism. Fine-grained pcGEMs can integrate protein synthesis in genome-scale models and could be further extended to account for enzyme cofactors. In this part, I introduce the development and application of the model CofactorYeast.

4.2.1 Development of CofactorYeast

The basic GEM for *S. cerevisiae* Yeast8 was expanded by a fine-grained approach to incorporate protein translation and cofactor binding reactions, resulting in the model named CofactorYeast. The model assumed that metabolic enzymes are only fully activated

when bound to all their respective cofactors, and metabolism can thus be affected by cofactor availability (**Figure 23**). In this model, eight metal ions and iron-containing compounds, i.e., heme and iron-sulfur clusters (ISCs) were studied. Among those, the top three metal ions that bind to proteins in *S. cerevisiae* are zinc (11% of proteins contain zinc), magnesium (9%) and iron (2%).

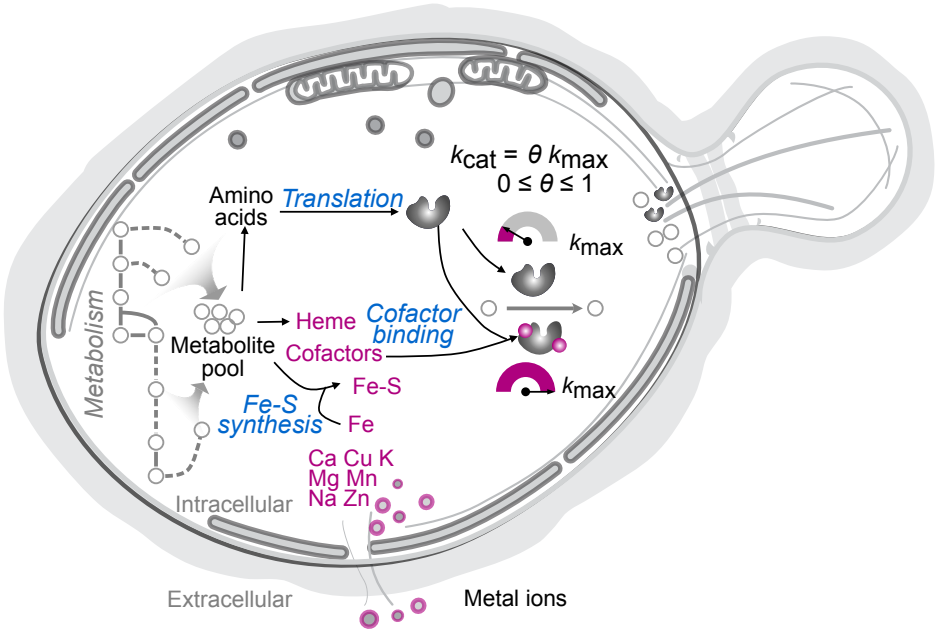

Figure 23 Schematic figure for CofactorYeast. This model expands Yeast8 by including protein translation and cofactor synthesis and binding reactions. Yeast compartmentalized figure source: SwissBioPics under CC BY4.0 license.

4.2.2 Simulation of metal ion abundance

CofactorYeast was used to simulate growth on different substrates under aerobic conditions using Biolog plates as depicted in **Figure 9a**. Among these conditions, there were 116 conditions where the CofactorYeast simulations were consistent with observed phenotypes. Thus, simulations of those conditions were selected to estimate the abundances of the metal ions binding on enzymes. While basic GEMs might predict constant metal ion content across all conditions if metal ions are presented with fixed coefficients in the biomass equation, CofactorYeast can simulate the changed abundances of metal ions. The simulated abundances of metal ions by CofactorYeast were compared with experimentally measured metal ion composition in the biomass under various culture conditions, which overall showed good consistency for several metal ions, including copper, manganese, iron and zinc (**Figure 24**). Since CofactorYeast only considers the bounded metal ions, rather than the free forms in the cell, the simulated abundances can be expected to be lower than measured concentrations, as was observed. Meanwhile, sodium, potassium and calcium showed large deviations between simulated and measured data, which could be explained by the fact that they are also involved in other biological functions such as maintaining membrane potential, which are not included in the model.

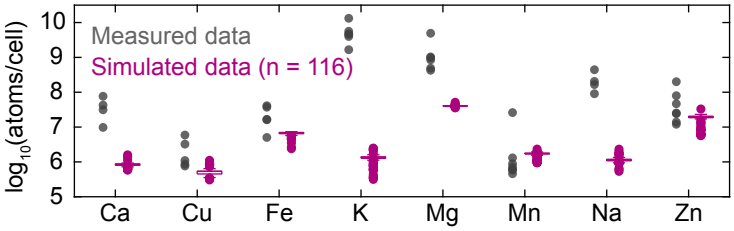

Figure 24 Simulated abundances of metal ions compared with experimentally measured data. Simulations were performed for 116 conditions, while the reported metal ion contents were measured under diverse culture conditions as reported in the literature.

4.2.3 Simulation of iron deficiency

Iron is one of the most interesting metal ions as it (i) can be present in various forms, including heme and ISCs, (ii) is one of the most widely used and most studied metal ions in the cell, and (iii) can serve as electron carriers in energy metabolism. Thus, iron was used to showcase the predictive potential of CofactorYeast.

Iron deficiency was simulated with CofactorYeast by reducing the iron uptake to 50% of the reference state as it resulted a 20% reduction in simulated growth rate, which is close to the experimentally measured growth reduction of yeast cells in response to iron deficiency [106]. Furthermore, the parameter θ was used to present the enzyme activity ratio of enzymes without binding cofactors compared with the enzymes bound with cofactors. This parameter was sampled from 0 to 0.9 to evaluate the impact on simulations.

CofactorYeast predicted that iron deficiency accompanies reduced growth with increased glucose uptake and ethanol and glycerol production (**Figure 25a**), suggesting that iron deficiency increases the glycolytic flux and leads to redox imbalance. The increased flux of glycolysis [106] and glycerol production [107] has previously been reported under iron starvation conditions. Note that the parameter θ did not have significant impact on the simulated growth and exchange rates.

Furthermore, the changes in the expression levels of individual enzymes in response to iron deficiency were simulated (**Figure 25b**). By comparing with transcriptomics data, CofactorYeast was able to capture the key changes, such as the downregulation of enzymes in tricarboxylic acid cycle (TCA cycle), electron transport chain, heme synthesis, amino acids metabolism and ergosterol biosynthesis [108], [109], indicating that yeast under iron deficiency could optimize iron usage through reducing the levels of unnecessary iron-containing enzymes. Most of enzyme expression level changes were not affected by the parameter θ, except C-4 methyl sterol oxidase (Erg25p) and delta 9 fatty acid desaturase (Ole1p). For those two proteins, the change of θ would cause opposite simulated expression changes, which can capture their experimentally measured upregulation in response to iron deficiency [108]–[110].

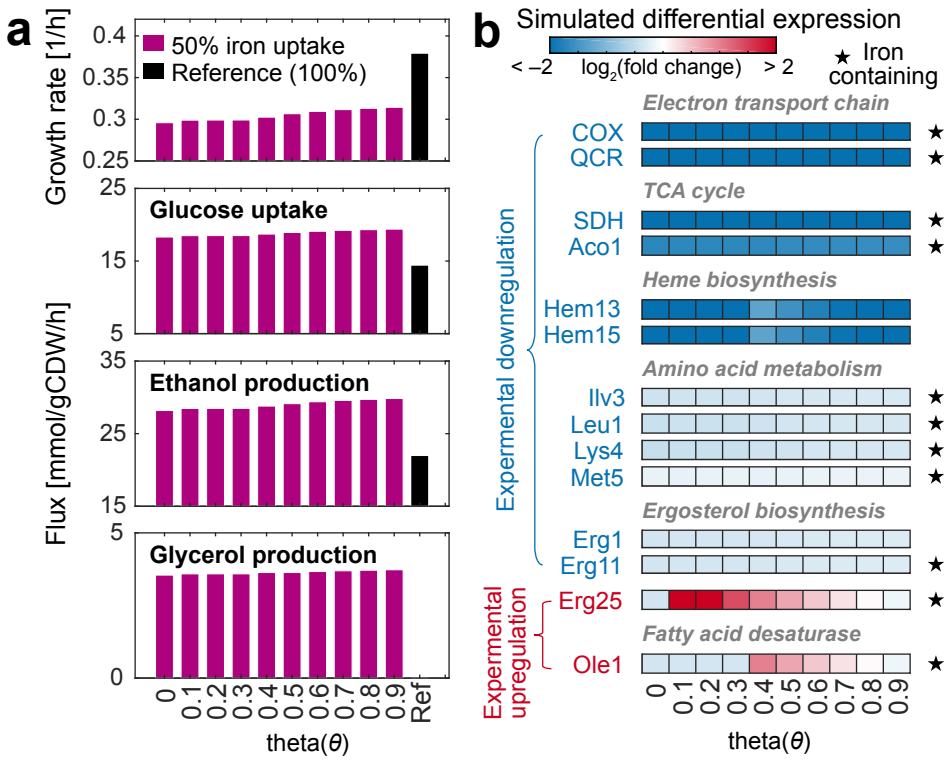

Figure 25 Simulations of iron deficiency. a) Simulated growth rates and exchange fluxes with various θ values for iron deficiency and reference condition. 50% iron uptake represents the iron deficiency condition. b) Comparison of experimentally measured differential expression with simulations upon iron deficiency. The color of the enzyme name represents experimentally measured differential expression upon iron deficiency (i.e., blue means measured down-regulation and red up-regulation). Heatmap represents simulated differential expression. Complex: COX, copper-dependent cytochrome c oxidase; QCR, ubiquinol cytochrome-c reductase; SDH, succinate dehydrogenase.

CofactorYeast adopts the fine-grained proteome-constrained approaches and can thus relates enzyme cofactors with metabolism. This model showcases how extending the basic GEM with enzyme cofactors enables the study of complex phenotype such as iron deficiency.

4.3 Modeling protein secretion: pcSecYeast

Yeast cells are used as cell factories to produce 15% of microbe-derived recombinant proteins [111]. Besides that, around 30% of native proteins are processed in the secretory pathway [65], which would compete with the recombinant secretory proteins for limited energy and proteome resource. The protein secretory pathway in yeast cells involves many processes such as PTMs, folding and vesicle sorting. The complexity of the secretory pathway hinders the system design from improving the recombinant protein production, thus there is an urgent need for developing a detailed protein secretory model. In this part, I describe the development of the proteome-constrained secretory model of *S. cerevisiae*, named pcSecYeast (**Figure 26**), and its applications.

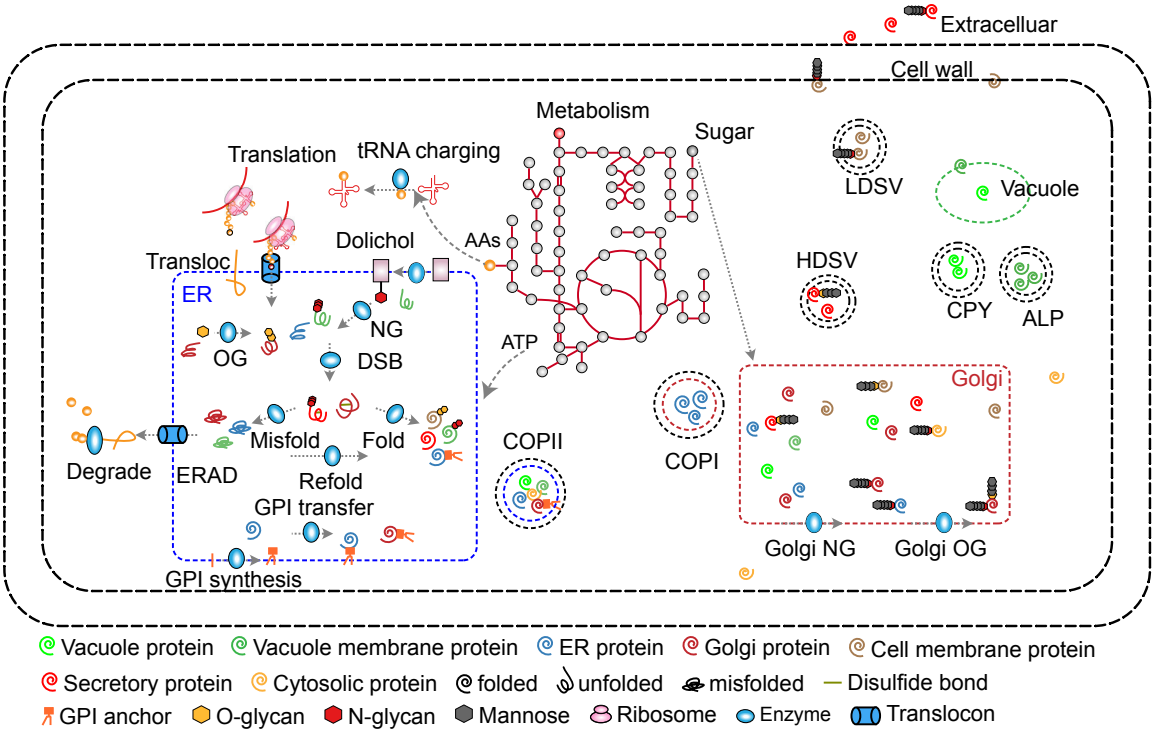

Figure 26 Schematic figure for pcSecYeast. pcSecYeast includes protein translation, translocation, glycosylate, GPI transfer, ERAD and sorting processes. Transloc: translocation; NG: N-glycosylation; OG: O-glycosylation; DSB: disulfide bond formation; GPI: glycosylphosphatidylinositol; ER: endoplasmic reticulum; ERAD: ER-associated degradation; LDSV: low-density secretory vesicles; HDSV: high-density secretory vesicles; ALP: alkaline phosphatase pathway; CPY: carboxypeptidase Y pathway.

4.3.1 Development of pcSecYeast

Firstly, I will define the rationale behind the selection of a fine-grained pcGEM rather than a basic GEM as a base for extension with secretion. Previously, there were published frameworks or models as the simple extension for recombinant protein secretion for yeast and mammalian cells [41], [112], [113]. However, even with constrained exchange rates, the predicted secretion rates of recombinant proteins would be 1,000-fold higher than the experimental data [114], suggesting that metabolism may not be the bottleneck for recombinant protein secretion. Moreover, fine-grained pcGEMs are expected to simulate the competition of recombinant proteins with native secretory proteins. Thus, a secretory model that would adopt a fine-grained proteome-constrained approach should be able to simulate how the cell would allocate its proteome and secretory capacity under various environmental conditions.

pcSecYeast was based on Yeast8 and expanded by protein expression, translation, folding and degradation reactions. For each protein processed in the secretory pathways, detailed protein processing was added into the model, including translocation from cytosol to ER, PTMs such as glycosylation, disulfide bond formation and GPI anchor transfer, misfolding and misfolded protein degradation, and vesicle sorting, detailly describing the processing of nascent peptide to its mature form. pcSecYeast differs from published models since those models only included the protein processing and secretion for the recombinant

protein while excluding all native secretory proteins. pcSecYeast contains protein synthesis processes for 1,639 proteins, which accounts for 70% of total proteome mass estimated from the PaxDb database [115]. Among those proteins, 492 proteins were responsible for the protein synthesis and secretion. The model contains 38,020 reactions, of which 31,824 are related to the protein part, indicating the complexity of the protein secretory pathway and the large scope of proteins in the model.

4.3.2 Simulation of growth upon different extracellular glucose concentrations

S. cerevisiae contains more than ten hexose transporters that are differently expressed in response to extracellular glucose concentrations [116]. Since those transporters are processed and secreted by the secretory pathway, pcSecYeast was used to simulate growth at different glucose concentrations to investigate the selection of glucose transporters. As previously mentioned, the simulation of growth relies on internal constraints such as the total protein abundance and the enzyme activities, which do not require the external exchange rates as constraints that used in the basic GEM simulation. Therefore, in the simulation of growth under different extracellular glucose concentrations, only the glucose concentration was used as the input and the maximum growth rate was searched. The simulation result showed that pcSecYeast can capture the Crabtree effect (**Figure 27a**). Besides that, the model predicted the switch from high-affinity glucose transporter Hxt7p to low-affinity glucose transporter Hxt3p and Hxt1p, which is consistent with reported transcriptome data [116]. In order to illustrate the mechanism behind this switch, I calculated and compared the energy requirements of these glucose transporters. The energy cost for glucose transporter was calculated based on:

$$Energy\ cost_i = unit\ \text{energy}\ cost_i * [\mathrm{E}_i] = unit\ \text{energy}\ cost_i * \frac{V_{glc}}{kcat_i * \frac{[S]}{[S]+K_{M,i}}}$$

which denotes the energy requirement for sustain a certain glucose uptake rate.

The unit energy cost for each glucose transporter was predicted from model simulations as the cost of synthesis, modification and secretion of one mmol of each of the glucose transporter. I predicted the unit energy cost for all native secretory proteins in *S. cerevisiae*, and the results showed that Hxt1p and Hxt3p had a smaller unit energy cost compared with Hxt7p, suggesting that synthesizing one mmol Hxt1p or Hxt3p would pose less energy burden on the cell. Hxt1p has a lower energy cost than Hxt7p, partly due to the fact that it has fewer N-glycosylation modification sites. The energy cost for each glucose transporter was calculated by combining the protein abundance of the glucose transporter $[\mathrm{E}_i]$, which was computed from glucose uptake rates sum at a specific growth rate (**Figure 27b**); extracellular glucose concertation; k_{cat} values; and K_{M} (the concentration of substrate which permits the enzyme to achieve half Vmax). The energy cost switch in **Figure 27c** explains the glucose transporter switch as utilization of Hxt1p and Hxt3p gradually gains the advantage due to the lower energy cost at high extracellular glucose concentrations. This

simulation demonstrates that pcSecYeast can capture the delicate changes in protein usage and could explain the phenotype.

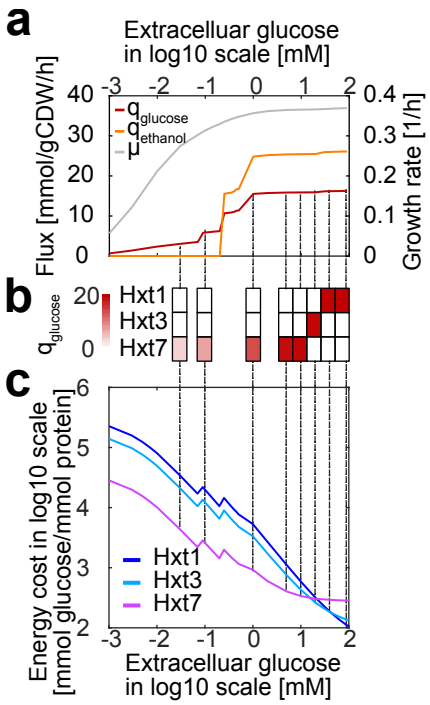

Figure 27 Simulations of changing extracellular glucose concentrations. a) Simulated glucose uptake, ethanol production, and specific growth rates upon different extracellular glucose concentrations. The simulations were performed with minimal media and free uptake of all components in the culture. Specific growth rates were obtained through a binary search. $q_{glucose}$: specific glucose uptake rate; $q_{ethanol}$: the specific ethanol production rate (units given as fluxes on the left axis); μ: specific growth rate (unit on right axis). b) Specific glucose uptake rate carried by each glucose transporter. Hxt1 and Hxt3: low-affinity glucose transporters; Hxt7: high-affinity glucose transporter. c) Calculation of energy costs of different glucose transporters with the input of specific glucose uptake rates on various extracellular glucose concentrations, unit glucose cost, K_M and k_{cat}. The calculation is based on equation in the text.

4.3.3 Simulation of protein misfolding

Protein folding is an error-prone process. The accumulation of misfolded proteins would cause severe cell disorder. Thus, pcSecYeast was expanded for vacuolar carboxypeptidase Y (YMR297W, CPY) production to study protein misfolding, since CPY and its derived misfolded form CPY* have been widely used to study cell disorder caused by the accumulation of misfolded proteins [117]. I used the expanded model to simulate different scenarios for protein misfolding and the impacts on growth (**Figure 28**). Simulated scenarios for CPY are exemplified in **Figure 28a**. The result suggested that expression of the correctly folded CPY would have the smallest growth reduction, while misfolded CPY would cause more fitness cost and more growth reduction (**Figure 28b**). If misfolded protein is retained in ER, growth would be further decreased, and the decrease level correlates with the retention time (**Figure 28b**). When misfolded protein is degraded, the amino acids and glycans would be recycled, while retained in the ER, those proteins would compete for the secretory machinery resources with the native secretory proteins.

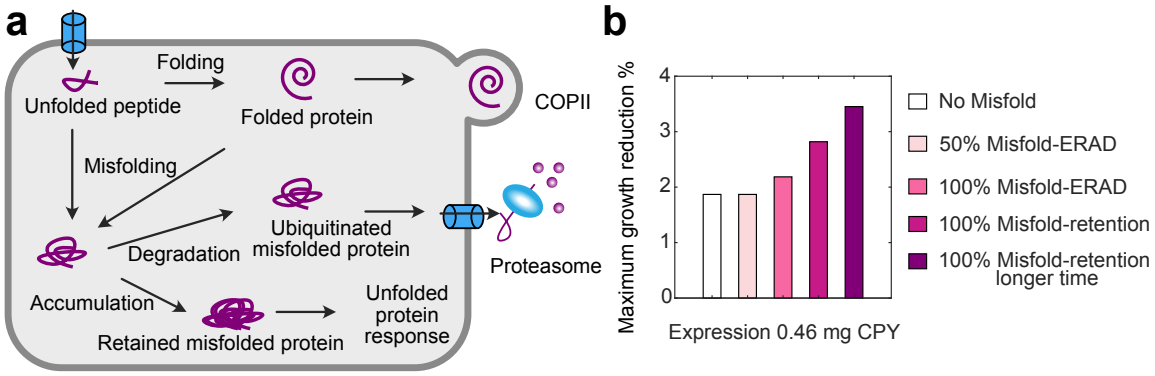

Figure 28 Simulation of CPY overexpression. a) Schematic view of different routes of expressed CPY. b) Reduction in simulated maximum specific growth rate due to expression at certain levels of CPY following different routes.

Furthermore, I identified when CPY was misfolded at high levels, then parts of it would be retained in ER caused by the imbalance of misfolding and the ERAD pathway. The retention would result in a steeper decrease of the maximum specific growth rate. I also noticed that there is a plateau representing the maximum CPY degradation rate, which indicates the maximum limit of the retro-translocation and ERAD pathway (**Figure 29**).

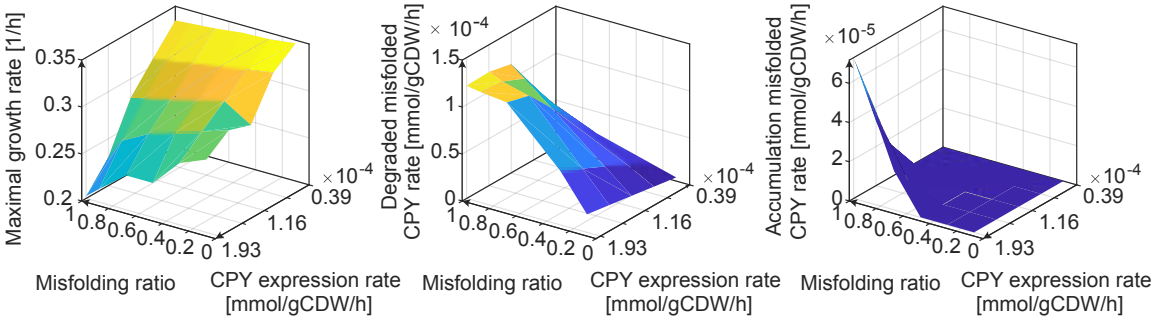

Figure 29 Simulation for various CPY expression levels and misfolding ratios.

4.3.4 Simulation of recombinant protein production

S. cerevisiae has been used to produce various recombinant proteins; therefore, I used the model to simulate recombinant protein production to identify factors influencing production. In total, eight recombinant proteins produced by *S. cerevisiae* were collected and added in the pcSecYeast one at a time (**Figure 30**), which was respectively used to simulate recombinant protein production at minimal media under different specific growth rates. The results showed that all those recombinant proteins production follow bell shape kinetics (**Figure 30b**), which are consistent with reports for several recombinant proteins production in *S. cerevisiae* [118]–[120]. On the contrary, the simulations of basic GEMs expanded with recombinant protein production showed negative linear correlations between protein production and growth (**Figure 30c**). Furthermore, while the simulated α-amylase production by basic GEM is around 1,000 times higher than the experimental values with the measured glucose uptake rate as the constraint [121], the simulation by

pcSecYeast is only five times higher, correcting the theoretical production rate for α-amylase.

a

Protein	abbr.	DSB	NG	OG	GPI	Length
Insulin precursor	IP	3	0	0	0	144
HGCSF	HGCSF	2	0	1	0	230
Hemoglobin	Hemoglobin	0	0	0	0	390
β-glucosidase	BGL	0	0	0	0	462
α-amylase	α-amylase	4	1	0	0	569
Acid phosphatase	PHO	8	9	0	0	526
Human serum albumin	HSA	17	1	0	0	609
HumanTransferin	HTF	19	0	1	0	703

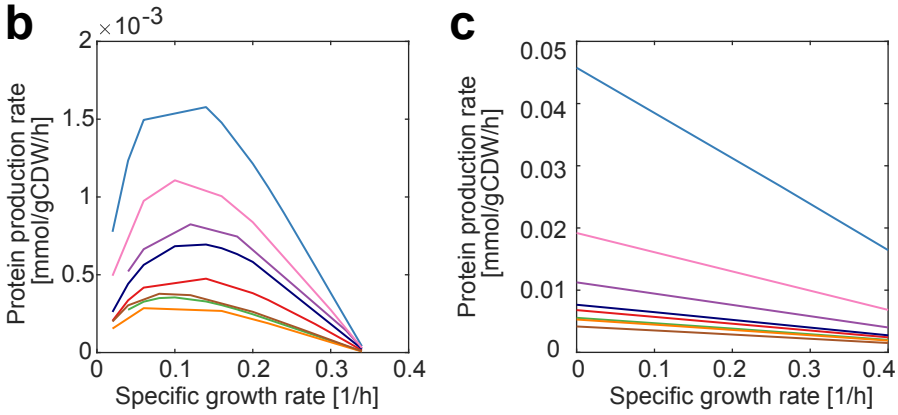

Figure 30 Simulation of recombinant protein production. Features of eight recombinant proteins produced by *S. cerevisiae*. Protein length includes the leader sequence. b) Simulation of maximum specific recombinant protein production rate as a function of specific growth rate using expanded pcSecYeast. c) Simulation of maximum specific recombinant protein production rate as a function of specific growth rate using basic GEM expanded with recombinant protein production.

4.3.5 Identification of overexpression targets for recombinant protein overproduction

Since there is a requirement for a systematic and rational design for recombinant protein production, the generated specific recombinant protein production models were used to predict overexpression targets for improving recombinant protein production. The prediction was enabled by an adapted Flux Scanning based on Enforced Objective Function (FSEOF) (**Figure 31**), which was initially designed for overexpression target identification for product production using basic GEMs. The original method identifies those amplified reaction fluxes with the enforcement of the product production, and corresponding enzymes and reactions are thus the overexpression targets. Since pcSecYeast can calculate the protein abundances, the upregulated proteins with the enforcement of the recombinant protein production can be directly selected as the raw overexpression targets. This step selects proteins with increased levels as targets. In order to improve the predictive accuracy, a ranking and filtering system would be used for giving priority to the predicted targets. In this step, three factors would be considered: whether the target is a subunit in a complex or has homologs, whether the protein has a large

fraction of reservation in proteome data. By doing so, the most effective targets would be given the highest priority (**Figure 31**).

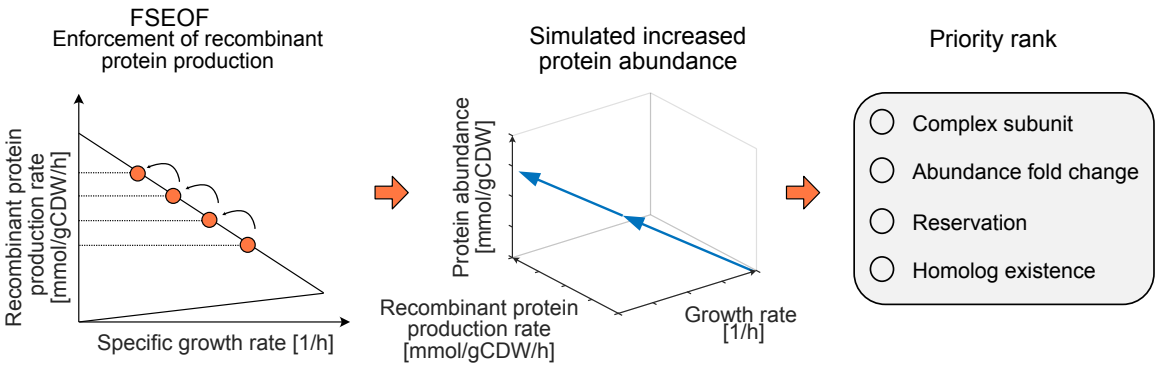

Figure 31 Schematic method for predicting overexpression targets to improve recombinant protein production. Adapted FSEOF method was used for target identification.

Even though much more proteins in the model are involved in the metabolism than protein secretion, 70% of those predicted targets are the secretory machinery proteins, while the remainder is related to metabolism. By comparing the predicted targets for those eight recombinant proteins, I found that there are 12 targets shared by all those eight recombinant proteins, which are mostly related to vesicle transport (**Figure 32**). Besides that, recombinant proteins which have shared PTMs always share similar targets. For example, the O-glycosylation-related machinery proteins are shared targets for O-glycosylated human-transferrin (HTF) and human granulocyte colony stimulating factor (hGCSF), suggesting the generalization of the targets for the recombinant proteins with the similar PTMs.

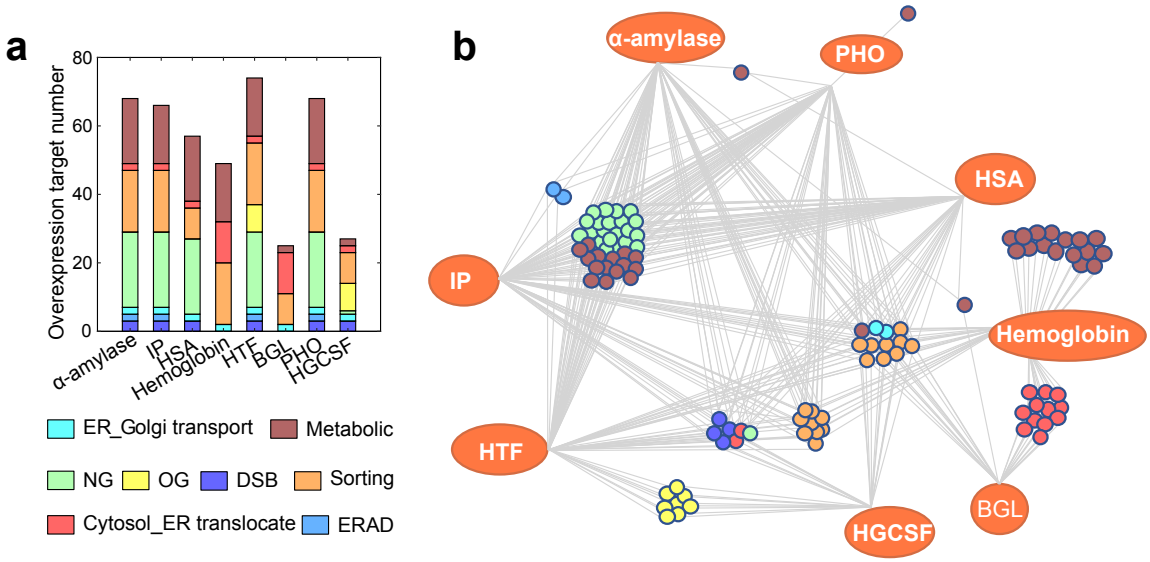

Figure 32 Comparison of predicted overexpression targets of eight recombinant proteins. a) The predicted overexpression targets of eight recombinant proteins grouped by pathways. b) Comparison of predicted targets of the eight recombinant proteins.

Next, we experimentally validated the predicted targets for α-amylase production. I selected 17 targets from the predicted targets based on their functions: covering the metabolism and different secretory processes (**Figure 33a**). Among those targets, the glucosidase Cwh41p, COPII-coated vesicles proteins Erv29p and Sec16p and protein disulfide isomerase Pdi1p were proved to be overexpression targets previously. Among the other secretory machinery targets, individual gene overexpression experiments were performed. From the **Figure 33b**, overexpression of the five genes led to significant increase of amylase yield: *SEC65* (2.2-fold), *ERO1* (2.0-fold), *SWA2* (1.3-fold), *ERV2* (1.4-fold) and *MNS1* (1.5-fold).

Sec65p is a subunit of the signal recognition particle, which has an important role in co-translational translocation. Nascent α-amylase peptide translocation from cytosol to ER would benefit from the *SEC65* overexpression **(Figure 33b**). Ero1p is an essential protein for maintaining ER redox balance. Overexpressing of *ERO1* has been shown to have a positive effect on production of several disulfide bonded recombinant proteins in *S. cerevisiae, P. pastoris* and *K. lactis* [122]–[124]. Based on the simulations, overexpression of *ERO1* also showed a positive effect on α-amylase production, suggesting that there is a general redox imbalance in the recombinant protein production. However, excessive overexpression of *ERO1* has a negative impact on the Human serum albumin (HSA) and human growth hormone (HSA/GH) fusion protein production in *P. pastoris* [124], which means that the sensitivity of the redox balance and the delicate protein regulation may be required for the optimal condition. *ERV2* is also involved in the disulfide bond formation in parallel with *ERO1*, which also increases the α-amylase production when it is overexpressed, as shown in the result **(Figure 33b**). Surprisingly, the three selected genes related to disulfide bond formation were all validated as positive targets, which suggested the high pressure of the redox balance in α-amylase production. Clathrin uncoating factor Swa2p is a cofactor for the coat disassembly, which is important for the proper fusion of the vesicle with the target membrane. Overexpression of *SWA2* also increased the α-amylase production. Mns1p is responsible for removing mannose residue from a glycosylated protein, essential for folding and ERAD. Since α-amylase contains multiple N-glycosylation sites, overexpressing *MNS1* would improve the N-glycosylation modification and the experiment result of increased α-amylase production indicates the importance of this step.

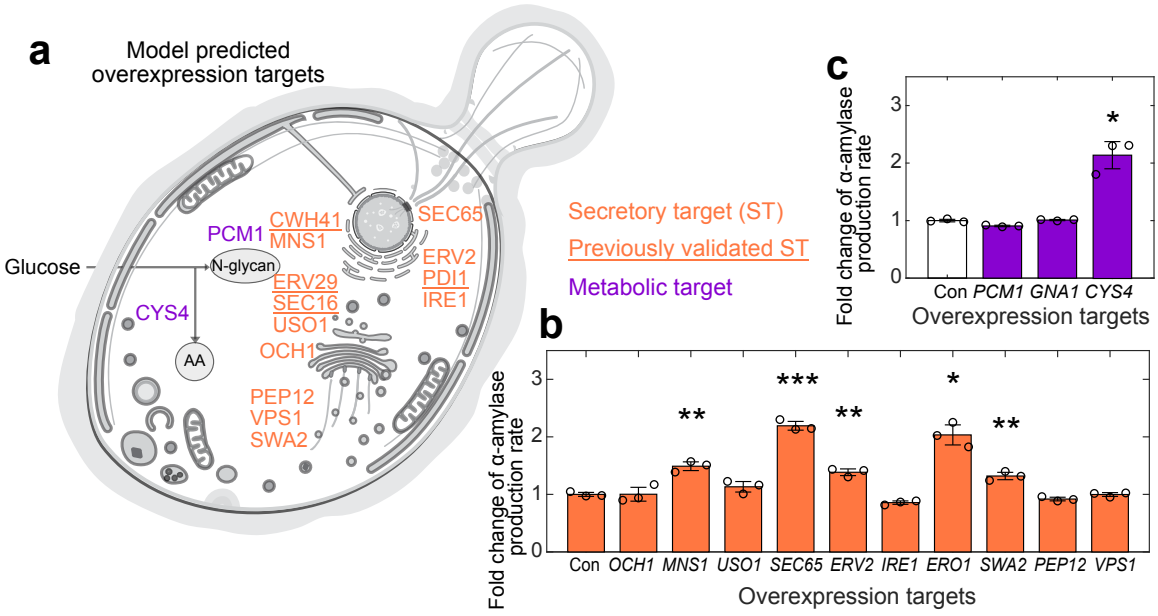

Figure 33 Experimental validation of predicted overexpression targets for α-amylase overproduction. a) Localization of the predicted targets. b) Validation result of predicted secretory targets. c) Validation result of predicted metabolic targets. *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$. Yeast compartmentalized figure source: SwissBioPics under CC BY4.0 license.

Among all three metabolic gene targets selected for α-amylase production, only *CYS4* had a positive effect on the α-amylase production when being overexpressed (**Figure 33c**). Cys4p is responsible for the synthesis of L-cysteine from L-homocysteine and L-serine. The cysteine composition in α-amylase is around nine times than that of average *S. cerevisiae* protein, suggesting the large drain of L-cysteine towards the α-amylase production. The other two metabolic genes (*PCM1* and *GNA1*) are related to the synthesis of N-glycosylation precursor. Overexpression of those genes did not have a positive impact on the α-amylase production.

In summary, the secretory pathway-related targets have higher accuracies compared with the metabolism-related targets. Combined with the higher ratio of predicted targets related to the secretory part, the secretory pathway may endure more burden under the recombinant protein production.

So far, pcSecYeast can capture good performance in capturing the delicate change for the glucose transporters with changing environmental conditions, simulate the phenotype of protein misfolding and ER retention, simulate the recombinant protein production and identifying related engineering targets, and therefore shows high quality and wide applications.

In this chapter, I described the second dimension of yeast model development, which is to incorporate more processes and/or constraints into GEMs. Each of those models is developed for specific purpose, and I have shown that they all have good performance. Reconstruction of pcGEMs require more parameters compared with the basic GEMs. To fill the gap for the lack of systematically measured data, assumption and estimation were

used to parameterize model, which would cause considerable uncertainties. Thus, I will discuss how to resolve these uncertainties in the next chapter.

5. Using machine learning to reduce uncertainties of k_{cat} values

Critical parameters used in all mentioned ecGEMs and pcGEMs are genome-scale k_{cat} values. The availability of measured k_{cat} values is scarce, as there are no high-throughput methods for k_{cat} measurement [63], [125]. When k_{cat} measurement is unavailable, modelers usually turn to arbitrary values or estimate k_{cat} values through fuzzy matching. Besides that, *in vitro* k_{cat} measurements can be considerably different from their *in vivo* counterparts, which could cause considerable uncertainties in model simulations. Machine learning have successfully been applied for reducing uncertainties in various models [60], [126], [127]. Therefore, in this chapter, I first compared two Bayesian statistic learning methods (traditional approximate Bayesian computation method and sequential Monte Carlo based approximate Bayesian computation method) by developing the ecGEM for *S. cerevisiae*, which contains ~ 4,000 enzymatic reactions and k_{cat} values (**Figure 34**) and then used the latter method to parameterize the 343 yeast/fungi ecGEMs, which is included in **Paper V**.

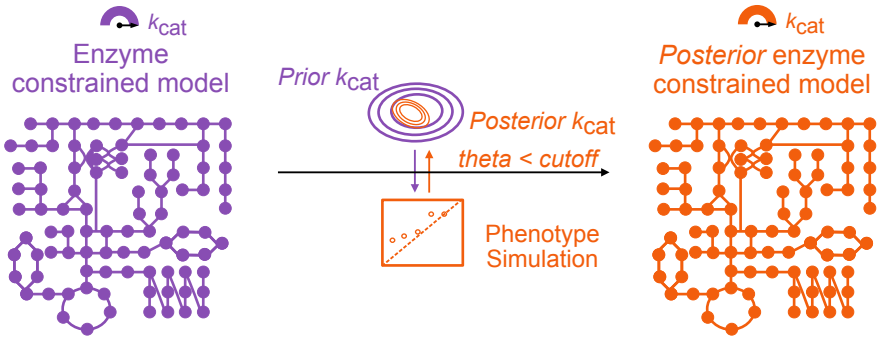

Figure 34 Schematic figure for the Bayesian method.

5.1 Method evaluation

I evaluated two Bayesian methods:

1) **traditional approximate Bayesian computation (traditional-ABC) method**, which samples k_{cat} values from *Prior* distributions until enough accepted samples are collected. For each enzymatic reaction in the GEM, I assumed that *Prior* distributions follows log10-transformed normal distributions where the collected k_{cat} values for each enzyme were set as mean value and a variance of one magnitude. One sample here means the combination of one round of random sampling of k_{cat} values for all enzymes. The sampled k_{cat} dataset was used to parametrize the *S. cerevisiae* ecGEM to simulate growth, and the simulated growth and exchange rates were compared with experimental data to calculate root mean squared error (RMSE). Only when the RMSE was lower than the cutoff, the sample of k_{cat} values was accepted. To monitor the sampling process, I reported the lowest RMSE after each generation (one generation equals 128 samples).

2) **sequential Monte Carlo based approximate Bayesian computation (SMC-ABC) method**, which samples the *Posterior* k_{cat} dataset from gradually updated *Prior* k_{cat} distributions. In that case, instead of sampling *Posterior* dataset from constant *Prior* in all generations as in the traditional-ABC method, the SMC-ABC method samples from the changing *Prior* distribution. I sampled 128 times within the *Prior* distribution for each generation, and 100 among those 128 datasets with lower RMSE between measurements and predictions were filtered out to make the distribution for the next generation *Prior* dataset.

Both Bayesian methods (traditional-ABC and SMC-ABC) rely on experimental observations. The experimental data, i.e., fermentation rates from batch and chemostat cultivations of *S. cerevisiae* were collected from literature, combining 40 entries for either aerobic or anaerobic conditions with 6 carbon sources.

Measured k_{cat} values were used as the mean values for the *Prior* distribution. I evaluated three different k_{cat} collection methods to assign k_{cat} values for enzymatic reactions in the GEM:
1) **Global**: a global k_{cat} assumption method that adopts the same arbitrary k_{cat} values for all enzymes as 7.9/s, the median k_{cat} value for enzymes in central carbon metabolism [64].

2) **Classical**: *in vitro* k_{cat} values collection from enzyme database. The k_{cat} values were queried from BRENDA database [128] by matching the EC numbers annotated in the UniProt database for enzymes in the model.

3) **DL**: deep-learning model predicted k_{cat} values. A deep learning model was pretrained using a dataset containing over 15,000 *in vitro* k_{cat} entries extracted from BRENDA and SABIO-RK databases. The RMSE for the k_{cat} prediction was benchmarked with the measurement to be 0.99 in the log10 scale. The training and evaluation of the deep learning model itself were introduced and discussed in detail in **Paper V**. Here, we used the deep learning model to predict the genome-scale k_{cat} data for 343 yeast/fungi species.

The coverage of collected k_{cat} data is visualized in **Figure 35**. Compared with the 'DL' method, the 'Classical' method had a relatively lower enzyme coverage and lower enzymatic reaction coverage since this method heavily relied on protein annotation for certain species to find EC numbers [46]. Besides the EC number annotation limitation, k_{cat} values measured for even well-studied species are also far from completeness. For *S. cerevisiae*, only 47 k_{cat} values are fully matched with proteins and substrates in the GEM, while other k_{cat} values are mostly from fuzzy matching with other substrates and organisms, or even introducing wild cards in the EC number. The missing part in the 'DL' method is caused by uncertain SMILES structures for generic metabolites.

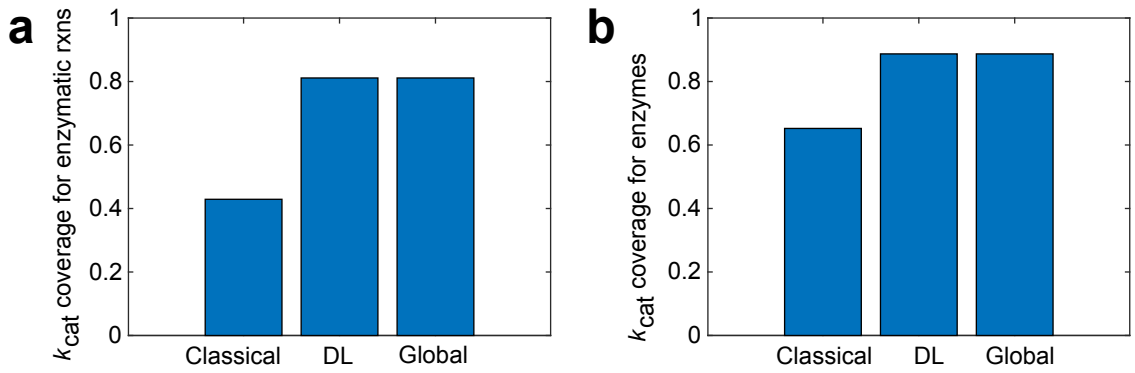

Figure 35 Coverage of enzymatic constraints a) for enzymatic reactions and b) for enzymes using three k_{cat} collection methods.

k_{cat} values collected from the three methods were used as the mean values for *Prior* distributions, and I assumed the *Prior* follows a log10 transformed normal distribution with a variance of one magnitude. Using the same cutoff (RMSE $\leq$ 0.5) for these two ABC methods, I performed sampling until there were 100 acceptable *Posterior* k_{cat} datasets for each method and each *Prior* dataset. **Figure 36** shows the sampling process for these two ABC methods. The SMC-ABC method requires less sampling than the traditional-ABC method to reach the same result, demonstrating that the SMC-ABC method is computationally more efficient (**Figure 36**). Due to the large number of parameters in the k_{cat} dataset and some k_{cat} distributions being too far from reality, it is almost impossible to collect 100 *Posterior* k_{cat} datasets from the traditional-ABC method. In **Figure 36a**, even after 300 generations (30,000 samples) which is ten times more than the samples as in the SMC-ABC method, the RMSE is still far away from the cutoff without even an acceptable *Posterior* dataset. Therefore, in the latter part, I would only focus on evaluating *Posterior* k_{cat} datasets sampled from the SMC-ABC method.

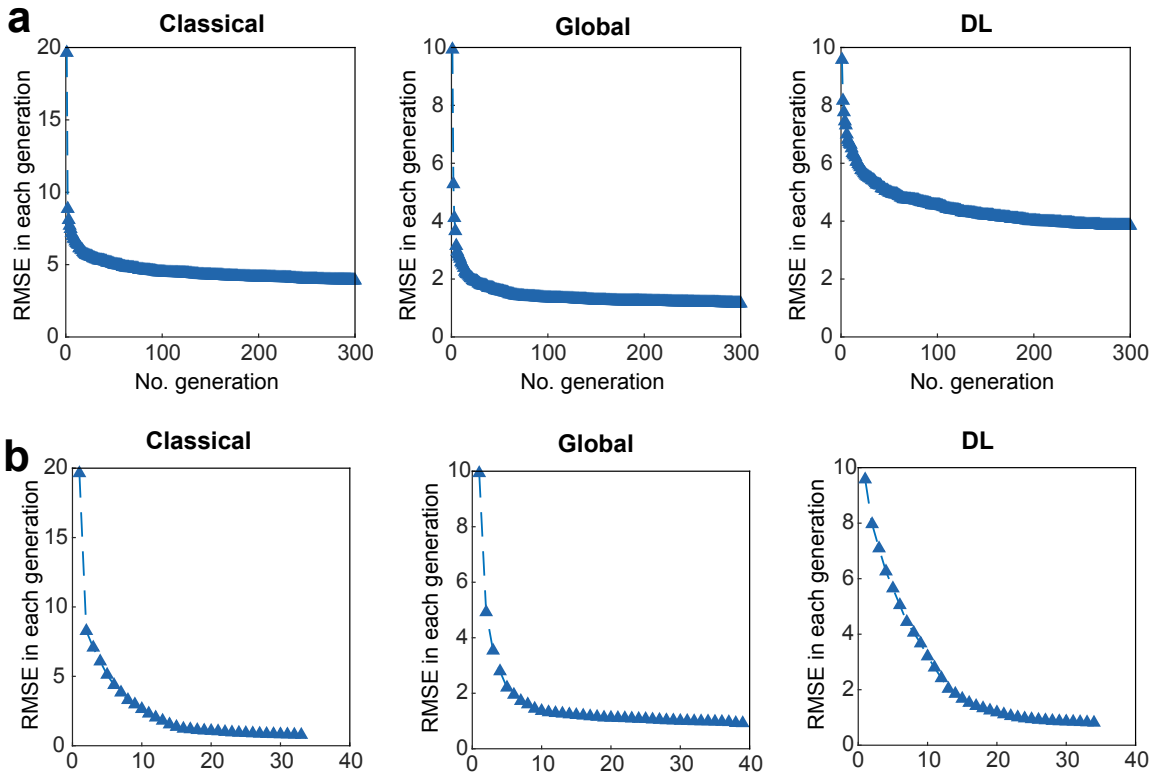

Figure 36 Training performance of a) Traditional-ABC Bayesian method and b) SMC-ABC method.

The ecGEMs parameterized by mean values of *Posterior* k_{cat} datasets from SMC-ABC method (*Posterior*-mean-ecGEM) were able to describe the observed measurements, and all captured the Crabtree effect accurately (**Figure 37a**). I then explored which parameter had been updated during the process. Through a principal component analysis (PCA) for all generated k_{cat} datasets, I found that in the SMC-ABC method, *Prior* k_{cat} datasets had gradually been updated towards the distinct *Posterior* k_{cat} datasets (**Figure 37b**). By comparing the variance for the *Prior* and *Posterior* datasets, I found that most variances were reduced, while mean values remained relatively unchanged during the process (**Figure 37c**). I found that 1,500~2,000 enzymes significantly reduced the variance for k_{cat}, while only less than 300 enzymes had significantly changed mean values of k_{cat}. Using *Prior* k_{cat} from 'DL' method would have the least number of enzymes with changed mean values. *Posterior* k_{cat} dataset strongly correlates with the *Prior* k_{cat} dataset from 'DL' and 'Classical' method (**Figure 37d**).

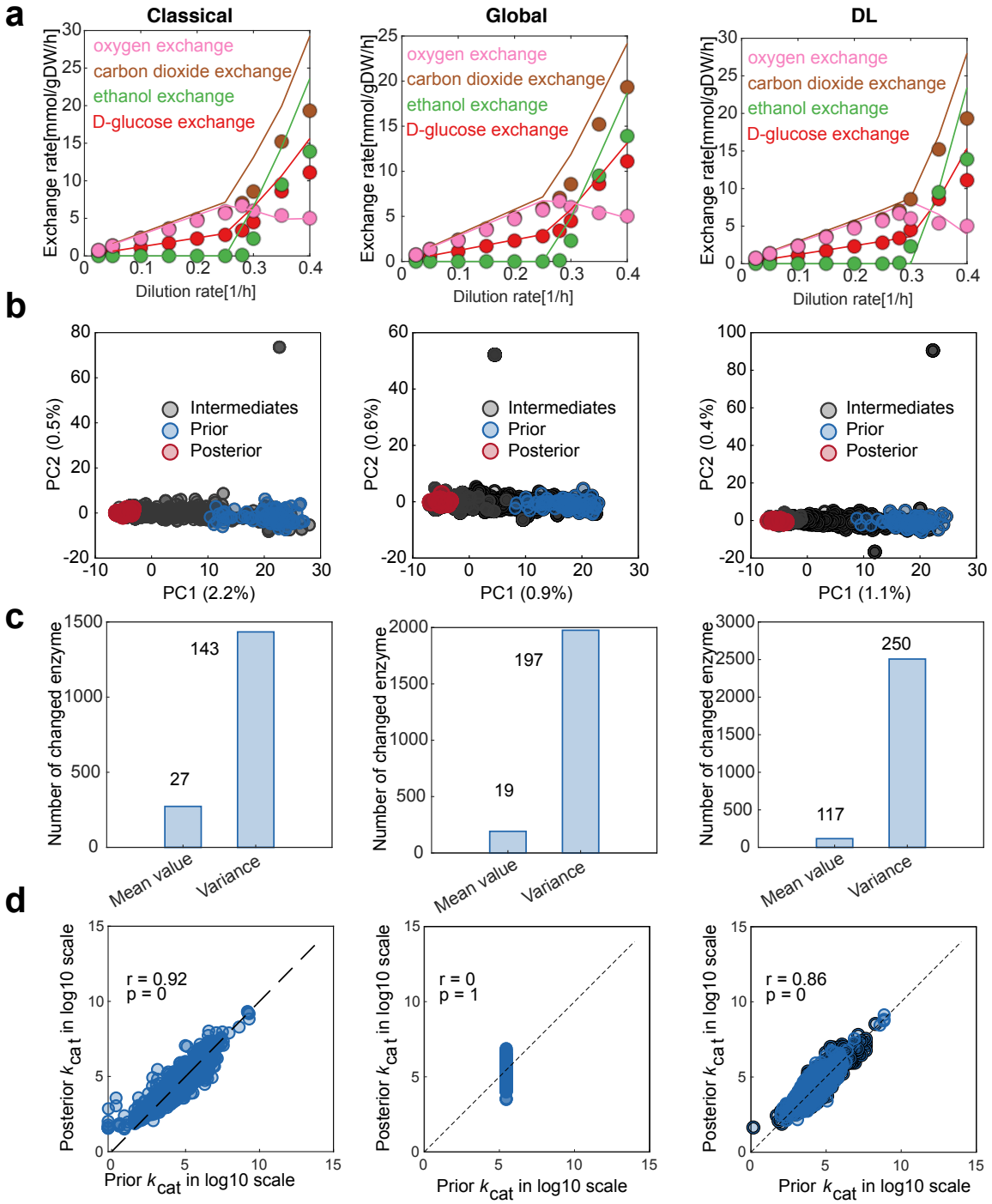

Figure 37 Evaluation of SMC-ABC Bayesian method with three k_{cat} collection methods. a) *Posterior*-mean-ecGEM simulations compared with experimental data. k_{cat} values in the ecGEMs here are mean values of *Posterior* datasets after the Bayesian training process. b) PCA for k_{cat} datasets sampled in the Bayesian method. Each parameter in the set was standardized by subtracting the mean and then dividing by the standard deviation before PCA. *Prior* datasets are in blue, while *Posterior* datasets are in red. All other datasets were termed as "intermediate" and marked in gray. c) The number of enzymes with a significantly changed mean value (Šidák adj. Welch's t-test p value < 0.01, two-sided) and variance (Šidák adj. one-tailed F-test p value < 0.01) between *Prior* and *Posterior* k_{cat} datasets. d) Analysis of *Prior* k_{cat} values and *Posterior* k_{cat} mean values.

To test the generality of the SMC-ABC method and monitor the training process, experimental growth datasets were split into training (50%) and test (50%) datasets. I used the training dataset to update the *Prior*, while the result was tested against the test dataset. RMSE between the experimental measurement and prediction for the test dataset was reduced proportionally with the training dataset. After 30 generations, RMSE for the training dataset was 0.5 and for the test dataset was 1, demonstrating the generalization of SMC-ABC method (**Figure 38**).

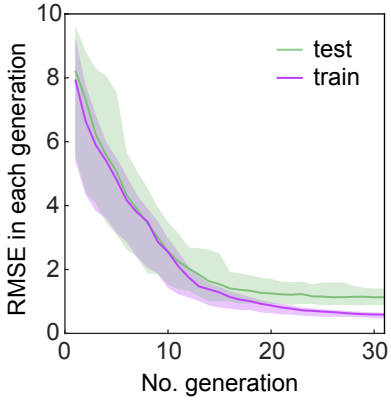

Figure 38 Validation of generalization of SMC-ABC method. In this validation method, 50% of experimental datapoints were used to update the *Prior* and then tested on the remaining 50%. Lines indicate median values and shaded areas indicate regions between the 5-th and 95-th percentiles (n=100).

So far, I have demonstrated that the SMC-ABC method is better than the traditional ABC method in the ecGEM reconstruction. Thus, I applied SMC-ABC method to the large-scale ecGEM reconstruction for 343 yeast/fungi species in the next section.

5.2 Large-scale ecGEM reconstruction for 343 yeast/fungi species

In **Chapter 3**, I described the GEM reconstruction for 343 for yeast/fungi species. We attempted to generate ecGEMs for all 343 species by adding enzymatic constraints in **Paper II**. However, the scarcity of experimentally measured k_{cat} and manual work required in the classical ecGEM reconstruction hinders large-scale ecGEM reconstruction for all studied yeast/fungi species. I introduce an automatic pipeline for ecGEM reconstruction using the deep-learning predicted k_{cat} as input (detailed described in **Paper V**), aided by a SMC-ABC method to bridge the manual work. The predicted k_{cat} values were used as mean values for *Prior* distribution, which were updated to *Posterior* using experimentally measured phenotypes. 445 entries of growth phenotype data for 76 yeast/fungi species with 16 carbon sources were collected from the literature. Using all these data, I automatically generated 343 functional ecGEMs for yeast/fungi species using the SMC-ABC method for DL-ecGEM and *Posterior*-mean-ecGEM reconstruction of *S. cerevisiae* as in **Section 5.1**.

In order to compare our work with the classical methods, I also built Classical-ecGEMs for the same species. The classical method is how ecGEMs are routinely parameterized with k_{cat} values extracted from enzyme databases. As for those newly sequenced yeast

species that do not contain EC number annotation, the corresponding EC number for homologs in *S. cerevisiae* were used to search k_{cat} values. Similar to the result for *S. cerevisiae* as in **Figure 35**, the Classical-ecGEMs managed to extract k_{cat} values for around 40% enzymes in the model and generated enzymatic constraints for around 60% enzymatic reactions, which is much lower than the DL-ecGEMs and its derived *Posterior*-mean-ecGEMs with around 80% of enzymes and 90% enzymatic reactions for 343 yeast/fungi species (**Figure 39**).

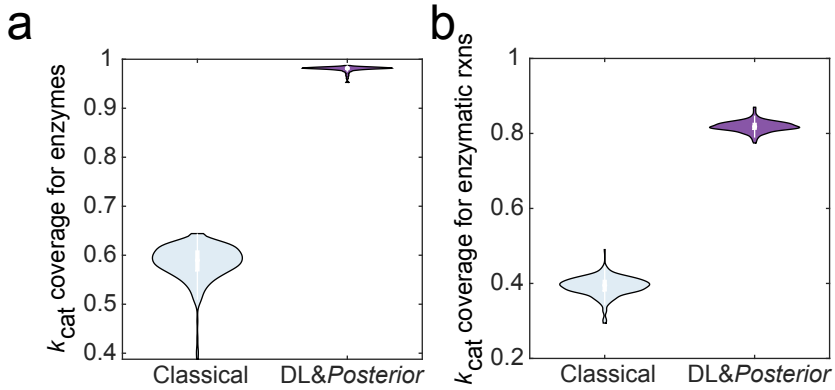

Figure 39 Coverage of enzymatic constraints a) for enzymes and b) for enzymatic reactions in 343 yeast/fungi species using three types of k_{cat} collection methods.

Then, I tested the phenotype prediction of three types of ecGEMs, which also suggested that the *Posterior*-mean-ecGEMs and DL-ecGEMs outperform the Classical-ecGEMs in prediction of exchange rates and maximum growth rates under diverse conditions for collected 445 experimental growth datasets (**Figure 40a-b**). Furthermore, as ecGEM can estimate the protein abundances, I also compared the simulated protein abundances from the three types of ecGEMs with available quantitative proteome data under diverse culture conditions such as different carbon sources, culture modes and different oxygen levels. The comparison suggested that the proteome prediction by *Posterior*-mean-ecGEMs had the lowest RMSE, while DL-ecGEMs reduced around 30% RMSE compared with that of Classical-ecGEMs (**Figure 40c**). In total, *Posterior*-mean-ecGEMs are the best representatives of those 343 yeast/fungi species.

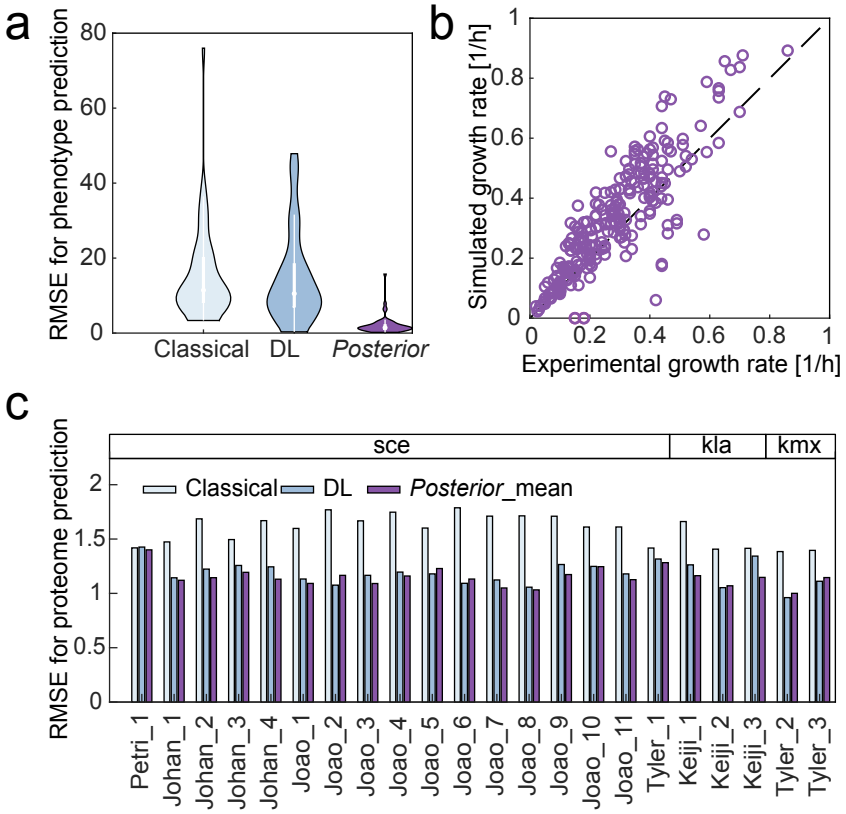

Figure 40 Predictions of different ecGEM modeling frameworks using SMC-ABC method of a) phenotype, b) growth rate and c) quantitative proteome data. RMSE is shown on log10 scale. Petri_1 proteome is from the reference [129], Johan_1-4 are from the reference [130], Joao_1-11 are from the reference [131], Tyler_1-3 is from the reference [132] and Keiji_1-3 is from the reference [133]. Petri_1: chemostat_D0.1, minimal media, D-glucose;Johan_1, batch, minimal media, D-glucose; Johan_2: batch, minimal media + amino acids, D-glucose; Johan_3: batch, minimal media, D-glucose; Johan_4: batch, minimal media + amino acids, D-glucose; Joao_1: batch, minimal media + amino acids, maltose; Joao_2: batch, minimal media + amino acids, oleate; Joao_3: batch, minimal media + amino acids, D-fructose; Joao_4: batch, minimal media + amino acids, sucrose; Joao_5: batch, minimal media + amino acids, trehalose; Joao_6: batch, minimal media + amino acids, (S)-lactate; Joao_7: batch, minimal media + amino acids, acetate; Joao_8: batch, minimal media + amino acids, pyruvate; Joao_9: batch, minimal media + amino acids, glycerol; Joao_10: batch, minimal media + amino acids, D-galactose; Joao_11: batch, minimal media + amino acids, raffinose; Tyler_1: chemostat_D0.1, minimal media, D-glucose; Keiji_1: batch, YPD, glycerol; Keiji_2: batch, YPD, D-glucose; Keiji_3: batch, YPG, glycerol; Tyler_2: chemostat_D0.1, minimal media, D-glucose; Tyler_3: chemostat_D0.1, minimal media, D-glucose.

In this chapter, I first compared two Bayesian methods for parameterizing the ecGEM of *S. cerevisiae* and found that the SMC-ABC is a better choice. Then the SMC-ABC method was used to parameterize the ecGEMs for 343 yeast/fungi species. Those automatically reconstructed ecGEMs are better representatives for yeast/fungi species considering the k_{cat} coverage, phenotype prediction and protein abundance prediction compared with the classical method. The reconstructed pipeline for ecGEM reconstruction would benefit large-scale ecGEMs reconstruction for other species in the future.

6. Conclusions

In this thesis, I explored yeast model development. Firstly, I started with reviewing the development of the fundamental base model: GEM for *S. cerevisiae*. The reconstructed GEM for *S. cerevisiae* Yeast8 is the currently most comprehensive reconstruction of yeast metabolism, which would contribute to systems biology analysis of yeast, including the use for *in silico* strain design and multi-omics integration and analysis. The platform provided through the GitHub repository enables the addition of new knowledge when it is acquired as well as for further improving the model for simulations.

Then I described the model reconstruction process in two dimensions: 1) how Yeast8 was used as basis for GEM development of multiple species/isolates and 2) how Yeast8 was expanded to include more biological constraints and processes. As for the first dimension, I reviewed the main content of **Paper I** and **Paper II**, and introduced the template model-based pipeline for large-scale GEM reconstruction and how this pipeline was used for the reconstruction of GEMs for 1,011 *S. cerevisiae* isolates and 343 yeast/fungi species. I showed that the GEMs reconstructed using this pipeline are comparable with published well-curated models in model scope and essential gene prediction. I also showed that those models can be used to identify metabolic diversity among those species/isolates. As for the second dimension, I reviewed minor parts of **Paper I** and **Paper II**, and the main content of **Paper III** and **Paper IV**, and introduced the different processes and constraints that were expanded to the basic GEMs. CofactorYeast was developed by linking metal ions to the metabolic enzymes, which was applied for quantitative and systematic investigation of metabolism and metal ions. I showed that the model simulation adopts the optimization strategy upon iron deficiency. To model protein secretion, I implemented the protein secretory pathway to the basic GEM with the fine-grained proteome-constrained concept to generate pcSecYeast, which was used to simulate the interaction of metabolism and gene expression in yeast, serving as a platform for elucidating protein secretion and identifying engineering targets for recombinant protein production.

During the second-dimension model development process, I realized that there are considerable uncertainties in model parameters, i.e., k_{cat} values. I showed that the Bayesian learning method could reduce those uncertainties and aid the automatic ecGEM reconstruction. Based on that, I developed an automatic ecGEM reconstruction pipeline, which reduces the manual work and improves the ecGEM quality to better represent the phenotype (**Paper V**).

In total, there are 1,699 models generated and described in this thesis, including 1,354 basic GEMs, 343 ecGEMs, CofactorYeast and pcSecYeast. Those models largely expand the number of yeast species with reconstructed models, span diverse complexity levels and species/strains, and would be a valuable collection for community usage.

7. Future perspectives

The model development can be used to reflect what is known and how the organism functions. On the other hand, the difference between model predictions and experiments can guide the discovery of missing parts in our understanding. Thus, there is not a foresee endpoint for the model development, but it is more like a continuous process that improves with our understanding. As for model development of *S. cerevisiae*, it follows the continuous development which can be reflected from the multiple times of updates for *S. cerevisiae* GEM, with each time either expanding the model scope by adding more reactions or curating the existing networks to improve the quality. Other yeast models are somehow behind the development of GEM for *S. cerevisiae*, but we can expect that model curation and updates would follow up for improving the quality. Since the continuous update is essential in the model development, how to ensure the reproducibility and traceability of each update would be major issues. So far, to achieve this goal, only model developments for several yeast species have used with version control tools such as Git. Even though there are other alternative approaches to host GEMs such as database-based systems, e.g., BiGG, but they would make for a very inflexible and less-open system in comparison with Git-versioned repositories. Therefore, this traceable Git system should become the standard for the GEM development. Efforts are underway to standardize this process by defining a standardized structure for Git repository of GEM (standard-GEM). Besides that, the sustainable development should also be considered for the complex models. Since most of complex models are built on the top of basic GEMs, the extension should be developed as add-on modules which could be easily transferred to the continuously updated GEM or to other phylogenetically close GEMs. Then, the workload for complex model reconstruction would reduce significantly. Thus, more complex models would be developed and employed for systems biology.

The continuous development poses another issue, i.e., what is the scope of a GEM. As I noticed, the initial GEM focuses on the central carbon metabolism and the biomass production. Then each version of updates would add new reactions into the model to expand the model scope, sometimes, this could lead to the reduction of flux prediction. For example, enzyme X is a phosphatase that very efficiently works on metabolite A, but also has significant moonlighting activity against metabolites B and C that contain a high-energy phosphate bond. In the model simulation, B and C would be favored and carry high flux, which would deviate from the real flux distribution. The dilemma for whether GEM should cover all possibilities of reactions or should ensure the precision of flux simulation requires modelers to clearly define the purpose of GEM.

Currently, the development of complex models of yeast species is still at its infant stage, mainly developed for *S. cerevisiae* as it has comprehensive reported parameters coming out of the extensive research. With the development of the understanding for other yeast species, the development of complex models for other yeast species should come at hand. Another factor hinders large-scale pcGEM or whole-cell model reconstruction is the

massive undertaking during the model reconstruction because most information in the model is fragmentally collected from literature, such as ribosome and proteasome composition. This situation is the same as how a basic GEM was reconstructed at the very beginning stage, which was boosted by the development of unified and combined metabolic reaction database such as KEGG and MetaCyc. With a database harboring standard formulation of protein-related information as the community did for metabolism, the blooming era of complex models would come. I have shown that template-based modeling approach could benefit large-scale GEM reconstruction, the complex model reconstruction for other yeast species could also benefited from a similar approach, given that other cellular processes than metabolism could also be widely conserved across species and can therefore be repurposed for each model. Then, through necessary manual curation, those complex models can turn into high-quality models.

The same question of model scope for the GEM development also goes for the complex model development: should we always aim to make the model more accurate and more complex reflections of reality, or should it be more like an abstraction of reality? Both these two aspects can be seen as the destination of the model development, with the first to represent a comprehensive platform to identify knowledge boundaries for the organism and the second to detangle the complex problem and to uncover the mechanism. Thus, models with different complexity should be developed regarding the purpose and scientific question.

Ultimately, I believe that this thesis showcases the predictive power of models with different complexity and could inspire model development and application in the future.

8. References

[1] P. E. McGovern *et al.*, “Fermented beverages of pre- and proto-historic China.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 51, pp. 17593–17598, 2004, doi: 10.1073/pnas.0407921102.

[2] X. Chen *et al.*, “FMN reduces Amyloid-β toxicity in yeast by regulating redox status and cellular metabolism.,” *Nat. Commun.*, vol. 11, no. 1, p. 867, 2020, doi: 10.1038/s41467-020-14525-4.

[3] M. G. Smith and M. Snyder, “Yeast as a model for human disease.,” *Curr. Protoc. Hum. Genet.*, vol. Chapter 15, p. Unit 15.6, 2006, doi: 10.1002/0471142905.hg1506s48.

[4] N. J. W. Kreger-van Rij, *The yeasts: a taxonomic study*. Elsevier, 2013.

[5] I. Matsumoto, T. Arai, Y. Nishimoto, V. Leelavatcharamas, M. Furuta, and M. Kishida, “Thermotolerant Yeast Kluyveromyces marxianus Reveals More Tolerance to Heat Shock than the Brewery Yeast Saccharomyces cerevisiae.,” *Biocontrol Sci.*, vol. 23, no. 3, pp. 133–138, 2018, doi: 10.4265/bio.23.133.

[6] C. Ratledge, “Single cell oils for the 21st century,” in *Single cell oils*, Elsevier, 2010, pp. 3–26.

[7] T. Dulermo and J.-M. Nicaud, “Involvement of the G3P shuttle and β-oxidation pathway in the control of TAG synthesis and lipid accumulation in Yarrowia lipolytica.,” *Metab. Eng.*, vol. 13, no. 5, pp. 482–491, 2011, doi: 10.1016/j.ymben.2011.05.002.

[8] P. Martorell, M. Stratford, H. Steels, M. T. Fernández-Espinar, and A. Querol, “Physiological characterization of spoilage strains of Zygosaccharomyces bailii and Zygosaccharomyces rouxii isolated from high sugar environments.,” *Int. J. Food Microbiol.*, vol. 114, no. 2, pp. 234–242, 2007, doi: 10.1016/j.ijfoodmicro.2006.09.014.

[9] D. Touchette *et al.*, “Novel Antarctic yeast adapts to cold by switching energy metabolism and increasing small RNA synthesis.,” *ISME J.*, 2021, doi: 10.1038/s41396-021-01030-9.

[10] J. Blazeck *et al.*, “Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production.,” *Nat. Commun.*, vol. 5, p. 3131, 2014, doi: 10.1038/ncomms4131.

[11] M. N. Larroque, F. Carrau, L. Fariña, E. Boido, E. Dellacassa, and K. Medina, “Effect of Saccharomyces and non-Saccharomyces native yeasts on beer aroma compounds.,” *Int. J. Food Microbiol.*, vol. 337, p. 108953, 2021, doi: 10.1016/j.ijfoodmicro.2020.108953.

[12] V. Mukherjee, D. Radecka, G. Aerts, K. J. Verstrepen, B. Lievens, and J. M. Thevelein, “Phenotypic landscape of non-conventional yeast species for different stress tolerance traits desirable in bioethanol fermentation.,” *Biotechnol. Biofuels*, vol. 10, p. 216, 2017, doi: 10.1186/s13068-017-0899-5.

[13] J. M. Cregg, J. L. Cereghino, J. Shi, and D. R. Higgins, “Recombinant protein expression in Pichia pastoris.,” *Mol. Biotechnol.*, vol. 16, no. 1, pp. 23–52, 2000, doi: 10.1385/MB:16:1:23.

[14] S. Rebello *et al.*, “Non-conventional yeast cell factories for sustainable bioprocesses.,” *FEMS Microbiol. Lett.*, vol. 365, no. 21, 2018, doi: 10.1093/femsle/fny222.

[15] J. Nielsen and J. D. Keasling, “Engineering Cellular Metabolism.,” *Cell*, vol. 164, no. 6, pp. 1185–1197, 2016, doi: 10.1016/j.cell.2016.02.004.

[16] J. S. Edwards and B. O. Palsson, “Systems properties of the Haemophilus influenzae Rd metabolic genotype.,” *J. Biol. Chem.*, vol. 274, no. 25, pp. 17410–17416, 1999, doi: 10.1074/jbc.274.25.17410.

[17] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, “Current status and applications of genome-scale metabolic models,” *Genome Biol.*, vol. 20, no. 1, p. 121, 2019.

[18] B. Papp, B. Szappanos, and R. A. Notebaart, “Use of genome-scale metabolic models in evolutionary systems biology.,” *Methods Mol. Biol.*, vol. 759, pp. 483–497, 2011, doi: 10.1007/978-1-61779-173-4_27.

[19] H. Lopes and I. Rocha, “Genome-scale modeling of yeast: chronology, applications and critical perspectives,” *FEMS yeast research*. 2017. doi: 10.1093/femsyr/fox050.

[20] I. Thiele and B. Ø. Palsson, “A protocol for generating a high-quality genome-scale metabolic reconstruction,” *Nat. Protoc.*, vol. 5, no. 1, p. 93, 2010.

[21] H. Wang *et al.*, “RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor,” *PLoS Comput. Biol.*, vol. 14, no. 10, 2018, doi: 10.1371/journal.pcbi.1006541.

[22] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, “The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum.,” *PLoS Comput. Biol.*, vol. 9, no. 3, p. e1002980, 2013, doi:

10.1371/journal.pcbi.1002980.
[23] S. M. D. Seaver *et al.*, “The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes.,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D575–D588, 2021, doi: 10.1093/nar/gkaa746.
[24] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, “Fast automated reconstruction of genome-scale metabolic models for microbial species and communities,” *Nucleic Acids Res.*, vol. 46, no. 15, pp. 7542–7553, 2018, doi: 10.1093/nar/gky537.
[25] M. Aite *et al.*, “Traceability, reproducibility and wiki-exploration for ‘à-la-carte’ reconstructions of genome-scale metabolic models,” *PLoS Comput. Biol.*, vol. 14, no. 5, p. e1006146, 2018.
[26] S. Magnúsdóttir *et al.*, “Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota,” *Nat. Biotechnol.*, vol. 35, no. 1, pp. 81–89, 2017, doi: 10.1038/nbt.3703.
[27] I. Famili, J. Förster, J. Nielsen, and B. O. Palsson, “Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 23, pp. 13134–13139, 2003, doi: 10.1073/pnas.2235812100.
[28] I. Domenzain, F. Li, E. J. Kerkhoven, and V. Siewers, “Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species,” *FEMS Yeast Res.*, vol. 21, no. 1, p. foab002, 2021.
[29] N. Loira, T. Dulermo, J.-M. Nicaud, and D. J. Sherman, “A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica.,” *BMC Syst. Biol.*, vol. 6, p. 35, 2012, doi: 10.1186/1752-0509-6-35.
[30] E. J. Kerkhoven, K. R. Pomraning, S. E. Baker, and J. Nielsen, “Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica,” *NPJ Syst. Biol. Appl.*, vol. 2, p. 16005, 2016.
[31] O. Dias, R. Pereira, A. K. Gombert, E. C. Ferreira, and I. Rocha, “iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis,” *Biotechnol. J.*, 2014, doi: 10.1002/biot.201300242.
[32] S. Marcišauskas, B. Ji, and J. Nielsen, “Reconstruction and analysis of a Kluyveromyces marxianus genome-scale metabolic model,” *BMC Bioinformatics*, vol. 20, no. 1, p. 551, 2019, doi: 10.1186/s12859-019-3134-5.
[33] N. Xu, L. Liu, W. Zou, J. Liu, Q. Hua, and J. Chen, “Reconstruction and analysis of the genome-scale metabolic network of Candida glabrata,” *Mol. Biosyst.*, 2013, doi: 10.1039/c2mb25311a.
[34] L. Caspeta, S. Shoaie, R. Agren, I. Nookaew, and J. Nielsen, “Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials,” *BMC Syst. Biol.*, 2012, doi: 10.1186/1752-0509-6-24.
[35] M. A. Asadollahi, J. Maury, K. R. Patil, M. Schalk, A. Clark, and J. Nielsen, “Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering.,” *Metab. Eng.*, vol. 11, no. 6, pp. 328–334, 2009, doi: 10.1016/j.ymben.2009.07.001.
[36] A. M. Ruffing and R. R. Chen, “Metabolic engineering of Agrobacterium sp. strain ATCC 31749 for production of an alpha-Gal epitope.,” *Microb. Cell Fact.*, vol. 9, p. 1, 2010, doi: 10.1186/1475-2859-9-1.
[37] K. R. Kildegaard *et al.*, “Engineering and systems-level analysis of Saccharomyces cerevisiae for production of 3-hydroxypropionic acid via malonyl-CoA reductase-dependent pathway.,” *Microb. Cell Fact.*, vol. 15, p. 53, 2016, doi: 10.1186/s12934-016-0451-5.
[38] G. Xu, W. Zou, X. Chen, N. Xu, L. Liu, and J. Chen, “Fumaric acid production in Saccharomyces cerevisiae by in silico aided metabolic engineering.,” *PLoS One*, vol. 7, no. 12, p. e52086, 2012, doi: 10.1371/journal.pone.0052086.
[39] X. Chen *et al.*, “Metabolic engineering of Torulopsis glabrata for malate production.,” *Metab. Eng.*, vol. 19, pp. 10–16, 2013, doi: 10.1016/j.ymben.2013.05.002.
[40] S. Li, X. Gao, N. Xu, L. Liu, and J. Chen, “Enhancement of acetoin production in Candida glabrata by in silico-aided metabolic engineering.,” *Microb. Cell Fact.*, vol. 13, no. 1, p. 55, 2014, doi: 10.1186/1475-2859-13-55.
[41] Z. A. Irani, E. J. Kerkhoven, S. A. Shojaosadati, and J. Nielsen, “Genome-scale metabolic model of Pichia pastoris with native and humanized glycosylation of recombinant proteins.,”

Biotechnol. Bioeng., vol. 113, no. 5, pp. 961–969, 2016, doi: 10.1002/bit.25863.
[42] T. Osterlund, I. Nookaew, and J. Nielsen, “Fifteen years of large scale metabolic modeling of yeast: developments and impacts.,” *Biotechnol. Adv.*, vol. 30, no. 5, pp. 979–988, 2012, doi: 10.1016/j.biotechadv.2011.07.021.
[43] B. J. Sánchez and J. Nielsen, “Genome scale models of yeast: towards standardized evaluation and consistent omic integration,” *Integr. Biol. (United Kingdom)*, 2015, doi: 10.1039/c5ib00083a.
[44] Y. Chen, G. Li, and J. Nielsen, “Genome-Scale Metabolic Modeling from Yeast to Human Cell Models of Complex Diseases: Latest Advances and Challenges.,” *Methods Mol. Biol.*, vol. 2049, pp. 329–345, 2019, doi: 10.1007/978-1-4939-9736-7_19.
[45] R. Adadi, B. Volkmer, R. Milo, M. Heinemann, and T. Shlomi, “Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters,” *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002575, 2012, doi: 10.1371/journal.pcbi.1002575.
[46] B. J. Sánchez, C. Zhang, A. Nilsson, P. Lahtvee, E. J. Kerkhoven, and J. Nielsen, “Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints,” *Mol. Syst. Biol.*, vol. 13, no. 8, p. 935, 2017, doi: 10.15252/msb.20167411.
[47] P. S. Bekiaris and S. Klamt, “Automatic construction of metabolic models with enzyme constraints.,” *BMC Bioinformatics*, vol. 21, no. 1, p. 19, 2020, doi: 10.1186/s12859-019-3329-9.
[48] I. Domenzain *et al.*, “Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0,” *bioRxiv*, p. 2021.03.05.433259, 2021, doi: 10.1101/2021.03.05.433259.
[49] C. J. Lloyd *et al.*, “COBRAme : A computational framework for genome-scale models of metabolism and gene expression,” pp. 1–14, 2018.
[50] A. Goelzer *et al.*, “Quantitative prediction of genome-wide resource allocation in bacteria.,” *Metab. Eng.*, vol. 32, pp. 232–243, 2015, doi: 10.1016/j.ymben.2015.10.003.
[51] P. Salvy and V. Hatzimanikatis, “The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models.,” *Nat. Commun.*, vol. 11, no. 1, p. 30, 2020, doi: 10.1038/s41467-019-13818-7.
[52] O. Oftadeh, P. Salvy, M. Masid, M. Curvat, L. Miskovic, and V. Hatzimanikatis, “A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics.,” *Nat. Commun.*, vol. 12, no. 1, p. 4790, 2021, doi: 10.1038/s41467-021-25158-6.
[53] I. E. Elsemman *et al.*, “Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies,” *bioRxiv*, p. 2021.06.11.448029, 2021, doi: 10.1101/2021.06.11.448029.
[54] J. R. Karr, K. Takahashi, and A. Funahashi, “The principles of whole-cell modeling.,” *Curr. Opin. Microbiol.*, vol. 27, pp. 18–24, 2015, doi: 10.1016/j.mib.2015.06.004.
[55] C. Ye *et al.*, “Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C.,” *Biotechnol. Bioeng.*, vol. 117, no. 5, pp. 1562–1574, 2020, doi: 10.1002/bit.27298.
[56] L. Österberg, I. Domenzain, J. Münch, J. Nielsen, S. Hohmann, and M. Cvijovic, “A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism.,” *PLoS Comput. Biol.*, vol. 17, no. 4, p. e1008891, 2021, doi: 10.1371/journal.pcbi.1008891.
[57] J. M. Gutierrez, A. Feizi, S. Li, T. B. Kallehauge, and H. Hefzi, “Prediction of specific productivity and limiting amino acids in animal cells using an expanded computational reconstruction of metabolism and protein secretion”.
[58] Z. Dermoun *et al.*, “TM0486 from the hyperthermophilic anaerobe Thermotoga maritima is a thiamin-binding protein involved in response of the cell to oxidative conditions.,” *J. Mol. Biol.*, vol. 400, no. 3, pp. 463–476, 2010, doi: 10.1016/j.jmb.2010.05.014.
[59] D. Dikicioglu and S. G. Oliver, “Extension of the yeast metabolic model to include iron metabolism and its use to estimate global levels of iron-recruiting enzyme abundance from cofactor requirements.,” *Biotechnol. Bioeng.*, vol. 116, no. 3, pp. 610–621, 2019, doi: 10.1002/bit.26905.
[60] G. Li *et al.*, “Bayesian genome scale modelling identifies thermal determinants of yeast metabolism,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–12, 2021.
[61] D. B. Bernstein, S. Sulheim, E. Almaas, and D. Segrè, “Addressing uncertainty in genome-

scale metabolic model reconstruction and analysis," *Genome Biol.*, vol. 22, no. 1, p. 64, 2021, doi: 10.1186/s13059-021-02289-z.
[62] Y. Chen and J. Nielsen, "Mathematical modelling of proteome constraints within metabolism," *Curr. Opin. Syst. Biol.*, 2021.
[63] D. Heckmann *et al.*, "Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models," *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018.
[64] A. Bar-Even *et al.*, "The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters," *Biochemistry*, vol. 50, no. 21, pp. 4402–4410, 2011, doi: 10.1021/bi2002289.
[65] C. K. Barlowe and E. A. Miller, "Secretory protein biogenesis and traffic in the early secretory pathway.," *Genetics*, vol. 193, no. 2, pp. 383–410, 2013, doi: 10.1534/genetics.112.142810.
[66] M. Delic, M. Valli, A. B. Graf, M. Pfeffer, D. Mattanovich, and B. Gasser, "The secretory pathway: exploring yeast diversity.," *FEMS Microbiol. Rev.*, vol. 37, no. 6, pp. 872–914, 2013, doi: 10.1111/1574-6976.12020.
[67] A. C. Horton and M. D. Ehlers, "Secretory trafficking in neuronal dendrites.," *Nat. Cell Biol.*, vol. 6, no. 7, pp. 585–591, 2004, doi: 10.1038/ncb0704-585.
[68] G. K. Gouras, C. G. Almeida, and R. H. Takahashi, "Intraneuronal Abeta accumulation and origin of plaques in Alzheimer's disease.," *Neurobiol. Aging*, vol. 26, no. 9, pp. 1235–1244, 2005, doi: 10.1016/j.neurobiolaging.2005.05.022.
[69] H. W. Aung, S. A. Henry, and L. P. Walker, "Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism," *Ind. Biotechnol.*, vol. 9, no. 4, pp. 215–228, 2013, doi: 10.1089/ind.2013.0013.
[70] R. Chowdhury, A. Chowdhury, and C. D. Maranas, "Using Gene Essentiality and Synthetic Lethality Information to Correct Yeast and CHO Cell Genome-Scale Models," *Metabolites*, vol. 5, no. 4, pp. 536–570, 2015, doi: 10.3390/metabo5040536.
[71] S. T. Hellerstedt *et al.*, "Curated protein information in the Saccharomyces genome database," *Database (Oxford).*, vol. 2017, no. 1, p. bax011, 2017, doi: 10.1093/database/bax011.
[72] P. D. Karp *et al.*, "The BioCyc collection of microbial genomes and metabolic pathways," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1085–1093, 2019, doi: 10.1093/bib/bbx085.
[73] A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, 2018, doi: 10.1093/nar/gkx1132.
[74] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs.," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, 2017, doi: 10.1093/nar/gkw1092.
[75] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017, doi: 10.1093/nar/gkw1099.
[76] S. Moretti, V. D. T. Tran, F. Mehl, M. Ibberson, and M. Pagni, "MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models.," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D570–D574, 2021, doi: 10.1093/nar/gkaa992.
[77] B. D. Heavner and N. D. Price, "Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction," *PLOS Comput. Biol.*, vol. 11, no. 11, p. e1004530, 2015, doi: 10.1371/journal.pcbi.1004530.
[78] M. H. Saier *et al.*, "The Transporter Classification Database (TCDB): 2021 update.," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D461–D467, 2021, doi: 10.1093/nar/gkaa1004.
[79] K. Correia and R. Mahadevan, "Pan-Genome-Scale Network Reconstruction: Harnessing Phylogenomics Increases the Quantity and Quality of Metabolic Models.," *Biotechnol. J.*, vol. 15, no. 10, p. e1900519, 2020, doi: 10.1002/biot.201900519.
[80] J. Huerta-Cepas *et al.*, "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D286–D293, 2015.
[81] L. V Rimareva, M. B. Overchenko, N. I. Ignatova, N. V Shelekhova, E. M. Serba, and A. Y. Krivova, "Study of intracellular ion composition of yeast Saccharomyces cerevisiae biomass," *Russ. Agric. Sci.*, vol. 43, no. 2, pp. 186–189, 2017.
[82] B. Hucker, L. Wakeling, and F. Vriesekoop, "Vitamins in brewing: presence and influence of thiamine and riboflavin on wort fermentation," *J. Inst. Brew.*, vol. 122, no. 1, pp. 126–137, 2016.

[83] M. L. Pallotta, “Evidence for the presence of a FAD pyrophosphatase and a FMN phosphohydrolase in yeast mitochondria: a possible role in flavin homeostasis.,” *Yeast*, vol. 28, no. 10, pp. 693–705, 2011, doi: 10.1002/yea.1897.
[84] J. D. M. Patring, J. A. Jastrebova, S. B. Hjortmo, T. A. Andlid, and I. M. Jägerstad, “Development of a simplified method for the determination of folates in baker’s yeast by HPLC with ultraviolet and fluorescence detection.,” *J. Agric. Food Chem.*, vol. 53, no. 7, pp. 2406–2411, 2005, doi: 10.1021/jf048083g.
[85] C. Lieven *et al.*, “Memote: A community driven effort towards a standardized genome-scale metabolic model test suite,” *bioRxiv*, p. 350991, 2018, doi: 10.1101/350991.
[86] B. J. Sánchez, F. Li, E. J. Kerkhoven, and J. Nielsen, “SLIMEr: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework,” *BMC Syst. Biol.*, vol. 13, no. 1, p. 4, 2019.
[87] J. Peter *et al.*, “Genome evolution across 1,011 Saccharomyces cerevisiae isolates.,” *Nature*, vol. 556, no. 7701, pp. 339–344, 2018, doi: 10.1038/s41586-018-0030-5.
[88] D. A. Skelly *et al.*, “Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast.,” *Genome Res.*, vol. 23, no. 9, pp. 1496–1504, 2013, doi: 10.1101/gr.155762.113.
[89] P. K. Strope *et al.*, “The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen.,” *Genome Res.*, vol. 25, no. 5, pp. 762–774, 2015, doi: 10.1101/gr.185538.114.
[90] P. Patra, M. Das, P. Kundu, and A. Ghosh, “Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts.,” *Biotechnol. Adv.*, vol. 47, p. 107695, 2021, doi: 10.1016/j.biotechadv.2021.107695.
[91] X. X. Shen *et al.*, “Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum,” *Cell*, 2018, doi: 10.1016/j.cell.2018.10.023.
[92] F. Carly and P. Fickers, “Erythritol production by yeasts: a snapshot of current knowledge.,” *Yeast*, vol. 35, no. 7, pp. 455–463, 2018, doi: 10.1002/yea.3306.
[93] C.-L. Flores and C. Gancedo, “Construction and characterization of a Saccharomyces cerevisiae strain able to grow on glucosamine as sole carbon and nitrogen source.,” *Sci. Rep.*, vol. 8, no. 1, p. 16949, 2018, doi: 10.1038/s41598-018-35045-8.
[94] Q. Chang, T. A. Griest, T. M. Harter, and J. M. Petrash, “Functional studies of aldo-keto reductases in Saccharomyces cerevisiae.,” *Biochim. Biophys. Acta*, vol. 1773, no. 3, pp. 321–329, 2007, doi: 10.1016/j.bbamcr.2006.10.009.
[95] C. P. Kurtzman, J. W. Fell, and T. Boekhout, *The yeasts: a taxonomic study*. Elsevier, 2011.
[96] S. B. Sohn, T. Y. Kim, J. H. Lee, and S. Y. Lee, “Genome-scale metabolic model of the fission yeast Schizosaccharomyces pombe and the reconciliation of in silico/in vivo mutant growth,” *BMC Syst. Biol.*, vol. 6, p. 49, 2012, doi: 10.1186/1752-0509-6-49.
[97] T. Liu, W. Zou, L. Liu, and J. Chen, “A constraint-based model of Scheffersomyces stipitis for improved ethanol production.,” *Biotechnol. Biofuels*, vol. 5, no. 1, p. 72, 2012, doi: 10.1186/1754-6834-5-72.
[98] P. Mishra *et al.*, “Genome-scale metabolic modeling and in silico analysis of lipid accumulating yeast Candida tropicalis for dicarboxylic acid production.,” *Biotechnol. Bioeng.*, vol. 113, no. 9, pp. 1993–2004, 2016, doi: 10.1002/bit.25955.
[99] M. Tomàs-Gamisans, P. Ferrer, and J. Albiol, “Fine-tuning the P. pastoris iMT1026 genome-scale metabolic model for improved prediction of growth on methanol or glycerol as sole carbon sources,” *Microb. Biotechnol.*, vol. 11, no. 1, pp. 224–237, 2018, doi: 10.1111/1751-7915.12871.
[100] S. Christen and U. Sauer, “Intracellular characterization of aerobic glucose metabolism in seven yeast species by 13C flux analysis and metabolomics.,” *FEMS Yeast Res.*, vol. 11, no. 3, pp. 263–272, 2011, doi: 10.1111/j.1567-1364.2010.00713.x.
[101] M. Workman, P. Holt, and J. Thykaer, “Comparing cellular performance of Yarrowia lipolytica during growth on glucose and glycerol in submerged cultivations.,” *AMB Express*, vol. 3, no. 1, p. 58, 2013, doi: 10.1186/2191-0855-3-58.
[102] Ö. Ata *et al.*, “A single Gal4-like transcription factor activates the Crabtree effect in Komagataella phaffii.,” *Nat. Commun.*, vol. 9, no. 1, p. 4911, 2018, doi: 10.1038/s41467-018-07430-4.
[103] G. G. Fonseca, A. K. Gombert, E. Heinzle, and C. Wittmann, “Physiology of the yeast Kluyveromyces marxianus during batch and chemostat cultures with glucose as the sole carbon source.,” *FEMS Yeast Res.*, vol. 7, no. 3, pp. 422–435, 2007, doi: 10.1111/j.1567-

1364.2006.00192.x.
[104] H. Juergens *et al.*, " Contribution of Complex I NADH dehydrogenase to respiratory energy coupling in glucose-grown cultures of Ogataea parapolymorpha ," *Appl. Environ. Microbiol.*, no. May, 2020, doi: 10.1128/aem.00678-20.
[105] C. A. Brown, A. W. Murray, and K. J. Verstrepen, "Rapid expansion and functional divergence of subtelomeric gene families in yeasts.," *Curr. Biol.*, vol. 20, no. 10, pp. 895–903, 2010, doi: 10.1016/j.cub.2010.04.027.
[106] M. Shakoury-Elizeh *et al.*, "Metabolic response to iron deficiency in Saccharomyces cerevisiae.," *J. Biol. Chem.*, vol. 285, no. 19, pp. 14823–14833, 2010, doi: 10.1074/jbc.M109.091710.
[107] G. P. Holmes-Hampton, N. D. Jhurry, S. P. McCormick, and P. A. Lindahl, "Iron content of Saccharomyces cerevisiae cells grown under iron-deficient and iron-overload conditions.," *Biochemistry*, vol. 52, no. 1, pp. 105–114, 2013, doi: 10.1021/bi3015339.
[108] S. Puig, E. Askeland, and D. J. Thiele, "Coordinated remodeling of cellular metabolism during iron deficiency through targeted mRNA degradation.," *Cell*, vol. 120, no. 1, pp. 99–110, 2005, doi: 10.1016/j.cell.2004.11.032.
[109] W. J. Jo *et al.*, "Novel insights into iron metabolism by integrating deletome and transcriptome analysis in an iron deficiency model of the yeast Saccharomyces cerevisiae.," *BMC Genomics*, vol. 10, p. 130, 2009, doi: 10.1186/1471-2164-10-130.
[110] A. M. Romero, T. Jordá, N. Rozès, M. T. Martínez-Pastor, and S. Puig, "Regulation of yeast fatty acid desaturase in response to iron deficiency.," *Biochim. Biophys. acta. Mol. cell Biol. lipids*, vol. 1863, no. 6, pp. 657–668, 2018, doi: 10.1016/j.bbalip.2018.03.008.
[111] G. Wang, M. Huang, and J. Nielsen, "Exploring the potential of Saccharomyces cerevisiae for biopharmaceutical protein production.," *Curr. Opin. Biotechnol.*, vol. 48, pp. 77–84, 2017, doi: 10.1016/j.copbio.2017.03.017.
[112] A. Feizi, T. Österlund, D. Petranovic, S. Bordel, and J. Nielsen, "Genome-Scale Modeling of the Protein Secretory Machinery in Yeast," *PLoS One*, vol. 8, no. 5, p. e63284, 2013, doi: 10.1371/journal.pone.0063284.
[113] J. M. Gutierrez *et al.*, "Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion.," *Nat. Commun.*, vol. 11, no. 1, p. 68, 2020, doi: 10.1038/s41467-019-13867-y.
[114] J. Sheng, H. Flick, and X. Feng, "Systematic Optimization of Protein Secretory Pathways in Saccharomyces cerevisiae to Increase Expression of Hepatitis B Small Antigen.," *Front. Microbiol.*, vol. 8, p. 875, 2017, doi: 10.3389/fmicb.2017.00875.
[115] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. von Mering, "Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.," *Proteomics*, vol. 15, no. 18, pp. 3163–3168, 2015, doi: 10.1002/pmic.201400441.
[116] J. A. Diderich *et al.*, "Glucose uptake kinetics and transcription of HXT genes in chemostat cultures of Saccharomyces cerevisiae.," *J. Biol. Chem.*, vol. 274, no. 22, pp. 15350–15359, 1999, doi: 10.1074/jbc.274.22.15350.
[117] A. Stolz and D. H. Wolf, "Use of CPY and its derivatives to study protein quality control in various cell compartments.," *Methods Mol. Biol.*, vol. 832, pp. 489–504, 2012, doi: 10.1007/978-1-61779-474-2_35.
[118] L. Paulová, P. Hyka, B. Branská, K. Melzoch, and K. Kovar, "Use of a mixture of glucose and methanol as substrates for the production of recombinant trypsinogen in continuous cultures with Pichia pastoris Mut+.," *J. Biotechnol.*, vol. 157, no. 1, pp. 180–188, 2012, doi: 10.1016/j.jbiotec.2011.10.010.
[119] M. L. Giuseppin, J. W. Almkerk, J. C. Heistek, and C. T. Verrips, "Comparative study on the production of guar alpha-galactosidase by Saccharomyces cerevisiae SU50B and Hansenula polymorpha 8/2 in continuous cultures.," *Appl. Environ. Microbiol.*, vol. 59, no. 1, pp. 52–59, 1993, doi: 10.1128/aem.59.1.52-59.1993.
[120] Y. E. Thomassen, A. J. Verkleij, J. Boonstra, and C. T. Verrips, "Specific production rate of VHH antibody fragments by Saccharomyces cerevisiae is correlated with growth rate, independent of nutrient limitation.," *J. Biotechnol.*, vol. 118, no. 3, pp. 270–277, 2005, doi: 10.1016/j.jbiotec.2005.05.010.
[121] M. Huang, J. Bao, B. M. Hallström, D. Petranovic, and J. Nielsen, "Efficient protein production by yeast requires global tuning of metabolism.," *Nat. Commun.*, vol. 8, no. 1, p. 1131, 2017, doi: 10.1038/s41467-017-00999-2.
[122] A. E. Wentz and E. V Shusta, "A novel high-throughput screen reveals yeast genes that

increase secretion of heterologous proteins.," *Appl. Environ. Microbiol.*, vol. 73, no. 4, pp. 1189–1198, 2007, doi: 10.1128/AEM.02427-06.
[123] T. Lodi, B. Neglia, and C. Donnini, "Secretion of human serum albumin by Kluyveromyces lactis overexpressing KlPDI1 and KlERO1.," *Appl. Environ. Microbiol.*, vol. 71, no. 8, pp. 4359–4363, 2005, doi: 10.1128/AEM.71.8.4359-4363.2005.
[124] M. Wu *et al.*, "Engineering of a Pichia pastoris expression system for high-level secretion of HSA/GH fusion protein.," *Appl. Biochem. Biotechnol.*, vol. 172, no. 5, pp. 2400–2411, 2014, doi: 10.1007/s12010-013-0688-y.
[125] A. Nilsson, J. Nielsen, and B. O. Palsson, "Metabolic models of protein allocation call for the kinetome," *Cell Syst.*, vol. 5, no. 6, pp. 538–541, 2017.
[126] A. D. Shrivastava and D. B. Kell, "FragNet, a Contrastive Learning-Based Transformer Model for Clustering, Interpreting, Visualizing, and Navigating Chemical Space.," *Molecules*, vol. 26, no. 7, 2021, doi: 10.3390/molecules26072065.
[127] J. Zrimec *et al.*, "Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure.," *Nat. Commun.*, vol. 11, no. 1, p. 6141, 2020, doi: 10.1038/s41467-020-19921-4.
[128] I. Schomburg, L. Jeske, M. Ulbrich, S. Placzek, A. Chang, and D. Schomburg, "The BRENDA enzyme information system–From a database to an expert system," *J. Biotechnol.*, vol. 261, pp. 194–206, 2017.
[129] P.-J. Lahtvee *et al.*, "Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast," *Cell Syst.*, vol. 4, no. 5, pp. 495–504, 2017.
[130] J. Björkeroth, K. Campbell, C. Malina, R. Yu, F. Di Bartolomeo, and J. Nielsen, "Proteome reallocation from amino acid biosynthesis to ribosomes enables yeast to grow faster in rich media," *Proc. Natl. Acad. Sci.*, vol. 117, no. 35, pp. 21804–21812, 2020, doi: 10.1073/pnas.1921890117.
[131] J. A. Paulo, J. D. O'Connell, R. A. Everley, J. O'Brien, M. A. Gygi, and S. P. Gygi, "Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 S. cerevisiae proteins across 10 carbon sources," *J. Proteomics*, vol. 148, pp. 85–93, 2016.
[132] T. W. Doughty *et al.*, "Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts," *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, 2020.
[133] K. Kito, H. Ito, T. Nohara, M. Ohnishi, Y. Ishibashi, and D. Takeda, "Yeast interspecies comparative proteomics reveals divergence in expression profiles and provides insights into proteome resource allocation and evolutionary roles of gene duplication," *Mol. Cell. Proteomics*, vol. 15, no. 1, pp. 218–235, 2016, doi: 10.1074/mcp.M115.051854.