Synthetic and Systems Biotechnology 9 (2024) 33–42

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Towards a hybrid model-driven platform based on flux balance analysis and a machine learning pipeline for biosystem design

Debiao Wu [1], Feng Xu [1], Yaying Xu, Mingzhi Huang [*], Zhimin Li, Ju Chu

State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai, 200237, People's Republic of China

ARTICLE INFO

Keywords:
Metabolic modeling
Machine learning
Flux balance analysis
Biosystems design
Saccharomyces cerevisiae
Succinate dehydrogenase

ABSTRACT

Metabolic modeling and machine learning (ML) are crucial components of the evolving next-generation tools in systems and synthetic biology, aiming to unravel the intricate relationship between genotype, phenotype, and the environment. Nonetheless, the comprehensive exploration of integrating these two frameworks, and fully harnessing the potential of fluxomic data, remains an unexplored territory. In this study, we present, rigorously evaluate, and compare ML-based techniques for data integration. The hybrid model revealed that the overexpression of six target genes and the knockout of seven target genes contribute to enhanced ethanol production. Specifically, we investigated the influence of succinate dehydrogenase (SDH) on ethanol biosynthesis in *Saccharomyces cerevisiae* through shake flask experiments. The findings indicate a noticeable increase in ethanol yield, ranging from 6 % to 10 %, in SDH subunit gene knockout strains compared to the wild-type strain. Moreover, in pursuit of a high-yielding strain for ethanol production, dual-gene deletion experiments were conducted targeting glycerol-3-phosphate dehydrogenase (GPD) and SDH. The results unequivocally demonstrate significant enhancements in ethanol production for the engineered strains *Δsdh4Δgpd1*, *Δsdh5Δgpd1*, *Δsdh6Δgpd1*, *Δsdh4Δgpd2*, *Δsdh5Δgpd2*, and *Δsdh6Δgpd2*, with improvements of 21.6 %, 27.9 %, and 22.7 %, respectively. Overall, the results highlighted that integrating mechanistic flux features substantially improves the prediction of gene knockout strains not accounted for in metabolic reconstructions. In addition, the finding in this study delivers valuable tools for comprehending and manipulating intricate phenotypes, thereby enhancing prediction accuracy and facilitating deeper insights into mechanistic aspects within the field of synthetic biology.

1. Introduction

Biological systems, encompassing proteins, pathways, and cells in their entirety, have experienced growing utilization across diverse biotechnological applications [1,2]. However, the progress in constructing tailored microbial cell factories has been hindered by the intricate nature of biological systems, which consist of a multitude of components and entail numerous unknown interactions among them. The DBTL cycle, which consists of four fundamental stages: design, build, test, and learn, represents a widely adopted methodology in synthetic biology research [3–6]. The learning stage plays a crucial role in extracting valuable biological insights from test data and leveraging them to inform subsequent designs. Given the substantial amount of data generated by modern biopharmaceutical plants, the automation of the learning stage becomes essential to ensure the efficient implementation of synthetic biology techniques. The advancement of various tools geared towards expediting the DBTL cycle has undoubtedly become crucial in automating the design, construction, and testing phases [3,7]. However, it is worth noting that the transition from the learning stage to the design phase has been relatively slow, with notable advancements observed only in a few specific and narrow applications [8,9]. Moreover, the integration of mechanistic and data-driven strategies remains limited, hindering the iterative and more efficient progression of this cycle. Consequently, these challenges and focal points present significant areas of interest within the realm of synthetic biology research.

Data-driven models and constraint-based mechanistic models represent two effective computational approaches utilized for analyzing biological data and constructing biological system models [10]. Specifically, machine learning (ML), a field that applies statistical and

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author.
E-mail address: huangmz@ecust.edu.cn (M. Huang).
[1] Debiao Wu and Feng Xu contributed equally to this work.

https://doi.org/10.1016/j.synbio.2023.12.004
Received 18 July 2023; Received in revised form 22 December 2023; Accepted 22 December 2023
Available online 29 December 2023

computer science methods to enable automated learning, prediction, and inference from experimental data by computer systems, has emerged as a widely applied technique in large-scale biological datasets recently. ML facilitates the identification of crucial features and the prediction of new data, thereby enhancing the accuracy and efficiency of data analysis [3,7]. However, data-driven methods often overlook prior biological knowledge during pattern analysis, which imposes limitations on the credibility and interpretability of the resulting models [11]. To this end, constraint-based modeling can be employed to simulate steady-state metabolism at the cellular level. In particular, the utilization of in vitro-generated metabolite fluxes has been incorporated to inform specific ML models, providing predictive advantages in specific instances [12–16]. For example, Yang et al. closely correlated biological signals with measured phenotypes using ML methodologies. To establish causal relationships, they integrated genome-scale metabolic models (GSMMs) into ML, thereby providing metabolic mechanistic insights [16]. However, there is still a lack of a comprehensive practical approach that effectively integrates data-driven models with experimental omics technologies. Such integration would facilitate the incorporation of mechanistic biological knowledge into the learning process [11].

As a proof of concept, this study presents a specific and practical learning framework that combines strain-specific metabolic models with ML algorithms to predict phenotypic traits of interest. The framework is applied to the design of engineered *Saccharomyces cerevisiae* (*S. cerevisiae*) aimed at achieving high-yield bioethanol production. In recent years, the global market share of bioethanol in the automotive fuel sector has witnessed a steady increase, with further growth anticipated in the future [17]. Additionally, the development of biofuels serves as a crucial contributor to energy diversification and a reduction in dependence on fossil fuels [18]. Recognizing the substantial commercial value of bioethanol, prior reviews have outlined commonly employed strategies to enhance production [19,20].

In this study, the challenges arising from limited data and the interpretation of learning results were addressed by integrating ML methods with a flux balance analysis (FBA) approach based on GSMM. To this end, we begin by augmenting the phenotypic dataset of ethanol production using a previously developed online detection model that leverages Raman spectroscopy. Subsequently, FBA was employed to simulate strain-specific metabolism. Reaction fluxes were extracted as additional features, forming a fluxomic dataset. Leveraging the obtained fluxomic dataset, we constructed prediction models for ethanol yield employing twelve diverse ML methods. A noteworthy enhancement in computational strain design for augmenting bioethanol production is observed compared to the utilization of the GSMM alone. Upon validating the superior performance of the proposed hybrid model, which combines mechanistic and data-driven approaches, we identify and validate targets that have not been reported previously. Furthermore, we endeavor to construct high-yield engineered strains to further amplify bioethanol production.

2. Materials and methods

2.1. Strains and cultivation

The strain *S. cerevisiae* BY4741, *E. coli* DH5α, and some plasmids for CRISPR/Cas9 technology were obtained from the laboratory of East China University of Science and Technology. The details of experimental strains utilized in this study are listed in Table S1. Additionally, the plasmids, sgRNA primers for CRISPR/Cas9-mediated gene knockout, donor DNA sequences, and primer lists for verification are included in Tables S2–S4. For cultivation, single colonies were selected from plates treated with antibiotics or activated and were inoculated into 250 mL flasks containing 25 mL of YPD medium. The flasks were incubated overnight at 28 °C and 220 rpm until reaching an OD_{600} of 2, which served as the desired optical density for flask fermentation. In this study, fermentation was carried out using a YSC synthetic medium or YPD medium. A 2 % inoculum volume of the culture was added to 500 mL flasks at 30 °C and 220 rpm for 14 h.

2.2. Genome-scale metabolic model

The GSMM encompasses all the documented biochemical reactions and transmembrane transporters that occur within an organism. Mathematically, the reaction network is represented as a stoichiometric matrix **S**, capturing the precise ratios of reactants and products involved in each biochemical conversion [20]. Under the assumption of metabolic steady-state, reaction rates (fluxes) are governed by mass and energy balance principles and can be described by a vector **v**, which represents the fluxes across the metabolic network. These fluxes are subjected to constraints defined by lower and upper limits, $\mathbf{v}_{\mathrm{lb}}$ and $\mathbf{v}_{ub}$, respectively. By adjusting these constraints, it becomes feasible to simulate different genetic or environmental factors, thereby creating context-specific metabolic models that align with experimental data [11].

$$subjectto\ W^T V = f \tag{1}$$

$$S\,v = 0 \tag{2}$$

$$v_{lb} \le v \le v_{ub} \tag{3}$$

Where the binary vector W represents the biomass pseudo-response as a distinct target, while *f* denotes the maximum achievable growth rate by the network, considering the imposed constraints. In this study, the GSMM of yeast 8.0.0 was utilized [21]. The simulation and analysis were primarily conducted using the COBRA Toolbox 3.0 [22] implemented in MATLAB, along with the Gurobi solver (Gurobi Optimization, LLC). Gene deletions were simulated using the *singleGeneDeletion* function within the COBRA Toolbox. The constraints incorporated the actual values of ethanol, glucose, glycerol, and biomass observed during fermentation for FBA. To determine the alterations in glucose, ethanol, glycerol, and biomass following a 14 h fermentation period, we relied on a Raman spectroscopy model developed through ML methods [23]. Briefly, the study generated predictions for glucose, ethanol, and biomass at 3-min intervals using a pre-established online monitoring model [23]. The profile data were converted into specific rate data and employed as constraints for the GSMM. In addition, the oxygen uptake rate was set to 1 mmol/gDCW/h based on the previous report [24], and the biomass reaction served as the objective function for the FBA simulations.

2.3. Establishment and application of hybrid models

2.3.1. Data preprocessing

To establish the correlation between metabolic flux data and actual ethanol yield, preprocessing steps were undertaken, encompassing feature selection and data dimensionality reduction. In particular, transport reactions and exchange reactions were initially eliminated. Likewise, pseudo reactions and diffusion reactions were excluded from consideration. This exclusion allowed for a focused examination of the essential metabolic pathways of interest. For feature selection, variance analysis, and univariate selection methods were employed primarily in this study (Fig. S1). Variance analysis was instrumental in identifying metabolic reactions that exhibited minimal fluctuations. These reactions, with a variance of 0, were deemed suitable for removal from the dataset. Throughout the univariate selection process, the correlation between reactions and the actual ethanol biosynthesis rate in the samples was evaluated, leading to the exclusion of reactions exhibiting low correlation. Following the preprocessing step, the number of features in the metabolic flux data was effectively reduced from 3496 in the GSMM to a more streamlined set of 331 selected features. This refined feature

set, accompanied by 883 data samples, served as the training data for the machine learning algorithm.

2.3.2. ML model selection, training and testing

In this study, ML algorithms from the open-source Python library Scikit-learn (https://scikit-learn.org) were employed to establish a regression model that correlated the flux data from the GSMM with the actual ethanol production in *S. cerevisiae*. Samples were randomly divided into training, validation, and testing subsets, with 70 %, 15 %, and 15 % of the main dataset being allocated, respectively. The training data served as the basis for model fitting, capturing inherent patterns within (619 samples). As numerous methods incorporated hyperparameters affecting the learning process, a grid search was executed on a validation subset to ascertain optimal hyperparameter configurations. To ensure robustness, all methods underwent three rounds of 5-fold cross-validation, employing 80 % of the data in each round. Following hyperparameter determination, models were retrained using the complete training dataset, encompassing validation samples (751 samples). Model performance assessment involved utilizing the trained models to predict outcomes in the independent testing datasets (132 samples). Notably, these samples were neither associated with nor included in the training or hyperparameter selection phases of the study.

2.3.3. Performance metrics, and model interpretation

The predictive performances of the hybrid model were assessed by evaluating the determination coefficients (R^2) value and root mean square error (RMSE). The manifestations of the model on the testing datasets were characterized by specific metrics, and their calculation formulae are outlined as follows. The predictive performance of the model is indicated by the R^2 value, where a closer approach to 1 signifies higher performance. Furthermore, a smaller RMSE implies a narrower discrepancy between predicted and actual values, reflecting superior model performance. The comparison of prediction performance across different models and the selection of the optimal model are facilitated through the assessment of R^2 and RMSE metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{5}$$

Selected models, chosen based on their predictive performance, serve as effective tools for data analysis. The interpretability of ML models was further enhanced by the application of the SHAP framework—a game-theoretic approach designed to elucidate the output of ML models [25]. In addition, some ML models, such as linear regression models, offer high interpretability by generating feature weight coefficients, indicating the direction of their positive or negative impact. For models that do not inherently produce such coefficients, SHAP can be employed to compute specific values for each feature, acting as equivalent positive and negative weight coefficients. Notably, SHAP values consider inter-feature influences, aiding in the evaluation of the reliability of feature weight coefficients, particularly in linear regression models. These models assign weight coefficients to features, where higher values signify a more significant impact on predicting the target variable. The polarity of these coefficients concerning the prediction target enables the determination of the association between data characteristics and the prediction target. Specifically, negative weight coefficients associated with metabolic reactions suggest that increased metabolic flux leads to decreased ethanol production. Conversely, positive weight coefficients indicate that heightened metabolic flux is linked to increased ethanol production.

2.4. Genetic manipulation

The detailed methods for plasmid and strain construction can be referred to published reports [26,27]. In brief, the deletion strains were constructed through CRISPR/Cas9-mediated genome editing [28]. For subsequent gene editing, pCAS9-NAT was introduced into *S. cerevisiae* BY4741. The gRNA plasmid and donor DNA fragments were co-transformed into yeast cells to facilitate gene deletion. Yeast transformation was executed employing the Frozen EZ Yeast Transformation II Kit (Zymo Research, USA), and subsequent selection of transformants was carried out on YPD-NAT (nourseothricin resistance) and HygB plates. To eliminate the gRNA plasmids bearing the HygB marker, the edited transformants were streaked onto YPD-NAT plates and cultured at 30 °C for 2 days, with this process repeated in triplicate.

2.5. Analytical methods

The quantification of target analytes in the fermentation broth was carried out according to prior studies [23]. The dry cell weight (DCW) was determined by calculating the optical density at 600 nm (OD_{600}). The ethanol concentration was analyzed using an Agilent 1490 gas chromatograph (Agilent Technologies, USA). In addition, an Agilent 1290 high-performance liquid chromatograph (Agilent Technologies, USA) was employed to detect glucose and glycerol.

2.6. Data processing and statistical analysis

The experiments were performed in triplicate. Statistical significance between two independent sample groups was assessed using Student's t-test, and correlation analysis was conducted using Pearson's correlation.

3. Results

3.1. Construction of the hybrid model

A ML-based prediction model for ethanol yield was established using flux data from the Scikit-learn platform. The performance of various ML models in predicting ethanol yield is summarized in Table 1. It is noteworthy that linear models, including LinearRegression, Ridge, Lasso, ElasticNet, BayesianRidge, and ARDRegression, exhibit R^2 values closer to 1 and lower RMSE values when compared to other nonlinear models. In this study, the linear models demonstrated a better predictive performance when forecasting ethanol production. The magnitude and direction of the characteristic weight coefficient associated with metabolic reactions directly convey the influence of the characteristic on ethanol production. Taking into account the interdependence among

Table 1
Performance of different machine learning models on ethanol production. KNN, k-Nearest Neighbor; SVR, Support Vector Machine; RR, ridge regression; GBR, Gradient Boosting Regressor; RF, Random Forest Regressor; DTR, Decision Tree Regressor; LassoR, Lasso Regressor; MLPR, Multilayer Perceptron Regressor; LR, Linear Regression; EN, Elastic Net; BR, Bayesian Ridge; ARDR, Automatic Relevance Determination Regression.

Model	R^2	RMSE
KNN	0.3279	9.8
SVR	0.3329	9.8
RR	0.9999	0.057
GBR	0.9741	1.9
RFR	0.9693	2.1
DTR	0.9365	3.0
LassoR	0.9995	0.27
MLPR	0.8421	4.8
LR	0.9999	0.092
EN	0.9992	0.35
BR	0.9999	0.043
ARDR	0.9999	0.025

features, the SHAP method was employed in this study to quantify the influence of each metabolic reaction feature on the predicted ethanol production. The SHAP values associated with various metabolic reaction features and their correlation with model feature weight coefficients are presented in Fig. 1. The results revealed a positive correlation between SHAP values and metabolic reaction weight coefficients, with a ρ value exceeding 0.8. This indicated that metabolic reactions possessing larger weight coefficients corresponded to larger SHAP values, reinforcing the idea that weight coefficients effectively reflect the impact of metabolic reactions on ethanol production. These findings underscored the reliability of linear models in exploring the association between metabolic reactions and ethanol production. To mitigate potential errors associated with individual model training, this study utilized the average weight coefficients derived from RidgeCV, LassoCV, ElasticNetCV, BayesianRidge, and ARDRegression for subsequent analysis of metabolic reaction weights. In particular, metabolic reactions with positive weights experience enhanced ethanol production rates and yield higher ethanol outputs when metabolic fluxes increase.

In contrast, reducing the metabolic fluxes of metabolic reactions with negative weights proves beneficial for elevating ethanol production rates and achieving greater ethanol yields. During the design phase of the DBTL cycle, the integration of inhibitory agents that target negatively weighted metabolic reactions can be incorporated into yeast fermentation. Genetic manipulations such as gene knockout or downregulation provide control over the enzymes or genes responsible for these reactions at the molecular level. Moreover, to augment the rates of positively weighted metabolic reactions, the introduction of activators or increased precursor concentrations for specific metabolic reactions can be employed during yeast fermentation, or gene overexpression for particular metabolic reactions can be implemented at the molecular level.

3.2. Evaluation of the hybrid model

To scale the weight coefficients of metabolic reactions within the range of [−1, 1], the *MaxAbsScaler* function in machine learning (ML) algorithms is employed. The resulting transformed coefficients are referred to as weight scores. The distribution of metabolic reactions with varying scores across the entire metabolic network is depicted in Fig. 2. Notably, the metabolic reactions highlighted by red arrows in the figure, especially those associated with genes within the demarcated red dashed box, are commonly employed to augment ethanol production via gene overexpression. For instance, target genes such as glutamate synthase (GLT), glutamine synthetase (GLN), glutamate dehydrogenase (GDH), alcohol dehydrogenase (ADH), pyruvate decarboxylase (PDC), and $NADP^+$-dependent glycerol-3-phosphate dehydrogenase (GAPN) were commonly overexpressed when involving high-yield ethanol strains [29, 30]. In addition, metabolic reactions denoted by blue arrows or reactions marked with blue circles were designed for gene knockout or silencing in metabolic engineering for ethanol production (Fig. 2). For instance, ubiquinol-cytochrome c reductase (QCR), cytochrome *c* oxidase (COX) [31], glycerol-3-phosphate dehydrogenase (GPD), aldehyde dehydrogenase 4 (ALD4), alcohol dehydrogenase 2 (ADH2), and glycerol efflux protein (FPS1) [29,32]. These findings aligned well with the distribution of metabolic network scores presented in Fig. 2, thereby validating the reliability of ML-based methods in the exploration of metabolic reaction interpretability.

In addition, a summary of target genes frequently associated with metabolic engineering for ethanol production in *S. cerevisiae*, along with their respective effects on ethanol production as observed in GSMM simulations, is presented in Table 2. However, it is noteworthy that within GSMM simulations, the full prediction of target genes for these established and effective metabolic reactions poses a considerable challenge. For instance, when considering GSMM simulations, the overexpression of genes such as GLT, GLN, GDH, and GAPN does not yield increased ethanol production. Likewise, the knockout of genes such as ALD4 and ADH2 does not lead to enhanced ethanol production in GSMM simulations. In this study, the models, which integrated mechanistic and data-driven approaches, successfully addressed the identified limitations. Such hybrid models offered notable advantages in biosystem analysis and facilitated the targeted selection of genes for metabolic engineering.

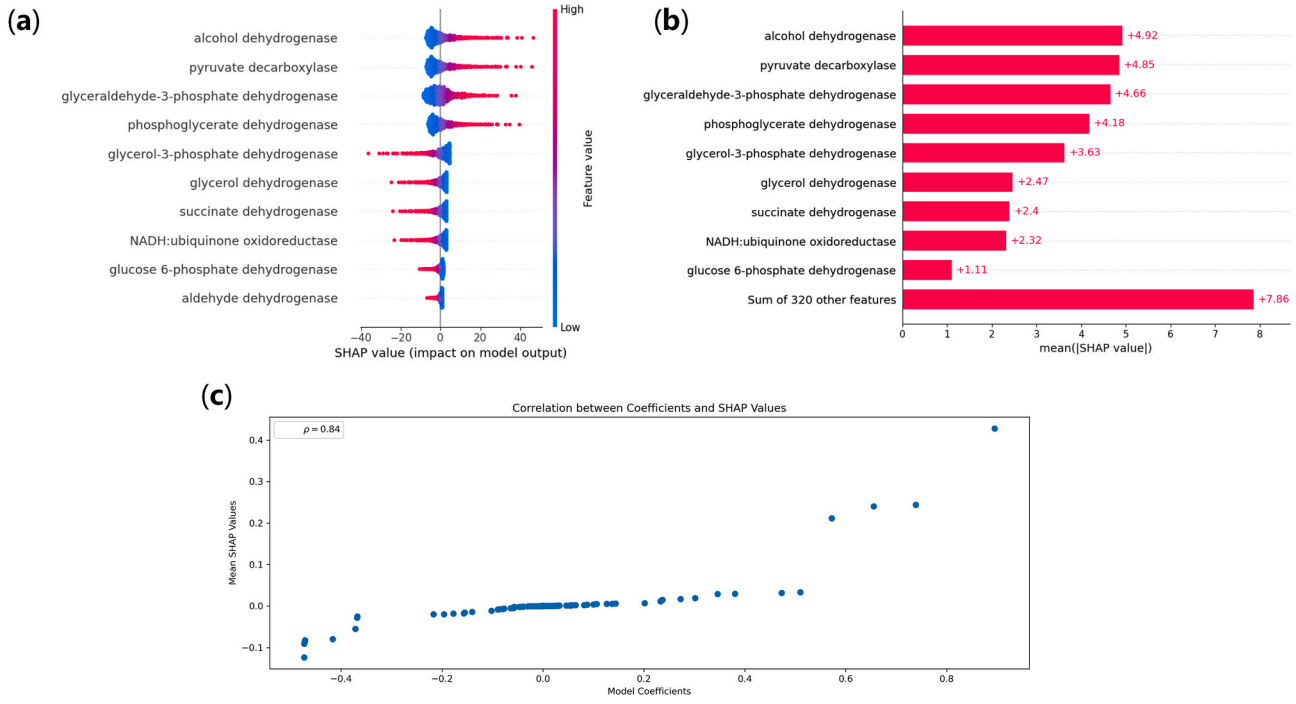

Fig. 1. SHAP value distribution of metabolic reactions and their correlation with model feature weight coefficients. (a) SHAP value distribution of partial metabolic reactions; (b) The average absolute value of the SHAP value of a metabolic reaction. (c) Correlation between metabolic reaction SHAP value and its weight coefficient.

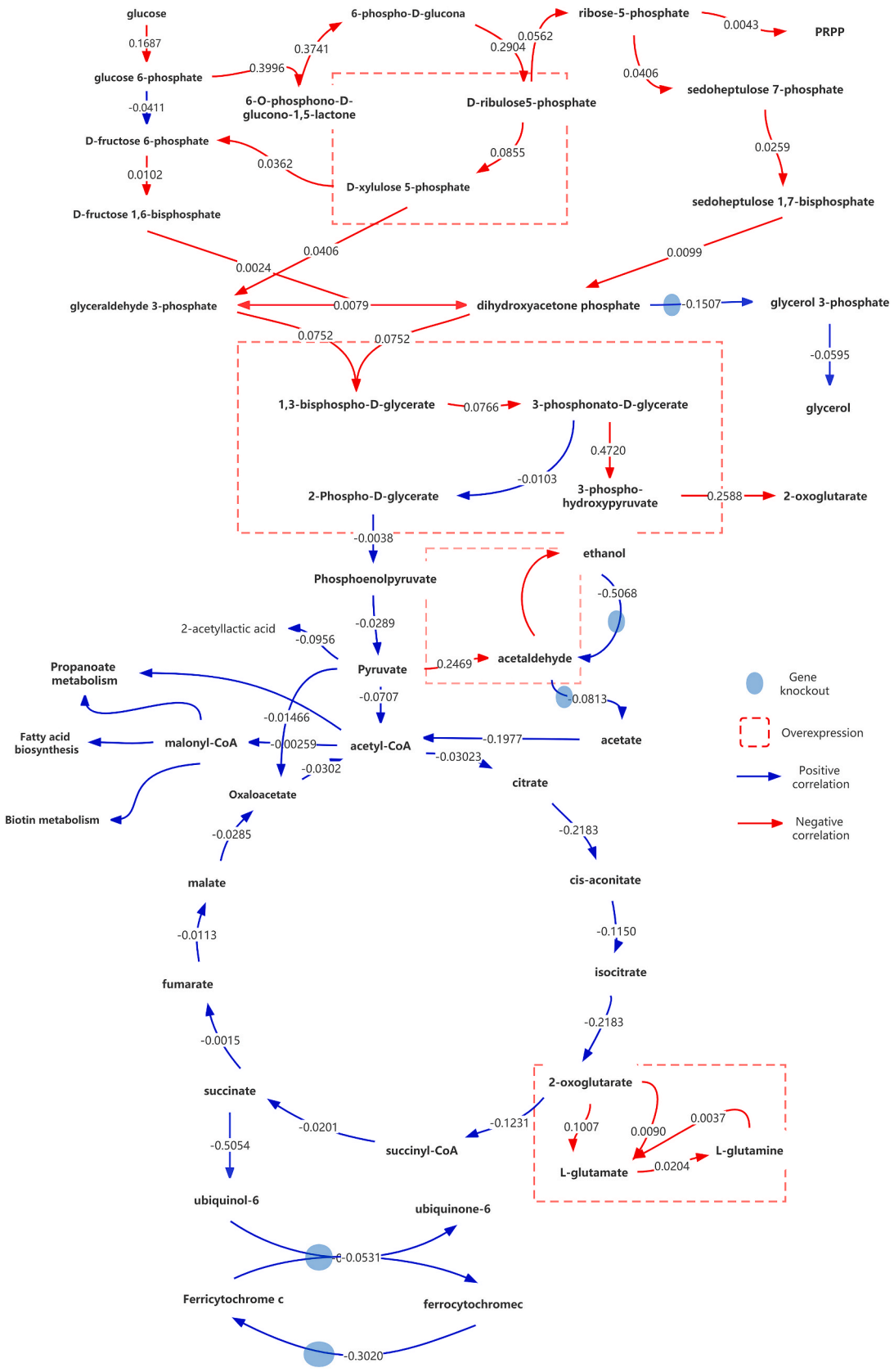

Fig. 2. Distribution of scores for metabolic responses involved in model training. Red represents metabolic reactions or pathways with positive scores, while blue represents metabolic reactions or pathways with negative scores.

Table 2
The predicted performance of common gene targets in ethanol metabolization in GEM simulation and hybrid model, respectively. GLT, Glutamate synthase; GLN, Glutamine synthetase; GDH, Glutamate dehydrogenase; ADH, Alcohol dehydrogenase; PDC, Pyruvate decarboxylase; GAPN, Glyceraldehyde-3-phosphate dehydrogenase; QCR, Ubiquinol-cytochrome c reductase; COX, Cytochrome-c Oxidase; GPD, Glycerol-3-phosphate dehydrogenase; ALD4, Acetaldehyde dehydrogenase 4; ADH2, Alcohol dehydrogenase 2; FPS1, Glycerol export protein.

	Genes	Prediction in GSMM simulation	Prediction by hybrid model	Reference
Targeted genes for overexpression	GLT	–	+	Literature validation
	GLN	–	+	Literature validation
	GDH	–	+	Literature validation
	ADH	+	+	Literature validation
	PDC	+	+	Literature validation
	GAPN	–	+	Literature validation
Targeted genes for knockout	QCR	+	+	Literature validation
	COX	+	+	Literature validation
	GPD	–	+	Literature validation
	ALD4	–	+	Literature validation
	ADH2	–	+	Literature validation
	FPS1	–	+	Literature validation

'+' indicates that the model can predict accurately, '-' indicates that the model cannot accurately predict.

3.3. Application of the hybrid model

While gene knockout of negatively weighted metabolic reactions holds the potential for enhancing ethanol production, it is essential to acknowledge that the weight coefficients of these reactions, as described by ML models, exhibit probabilistic characteristics in their correlation with ethanol production. This relationship lacks determinism and cannot be solely explained from a metabolic mechanism perspective. Therefore, the weighted scores of metabolic reactions were integrated with the metabolic mechanisms outlined in GSMM to identify novel and effective gene knockout targets in this study. To enable a systematic and objective selection process, an exhaustive examination of the impact of diverse intracellular metabolic pathways on ethanol metabolism in *S. cerevisiae* was conducted. As depicted in Fig. 3, weighted scores for metabolic reactions spanning diverse pathways were obtained. The findings highlighted significant pathways that negatively impacted ethanol production, encompassing oxidative phosphorylation, the tricarboxylic acid (TCA) cycle, pyruvate metabolism, fatty acid and glycerolipid metabolism, and several amino acid metabolism pathways. Conversely, notable pathways that positively impact ethanol production encompass carbon metabolism, the pentose phosphate pathway, acetaldehyde, and dicarboxylic acid metabolism, along with specific amino acid metabolism pathways. Pathways involving multiple reactions were primarily associated with amino acid biosynthesis or metabolism. Notably, most amino acid metabolism pathways negatively affected ethanol production due to their close association with cell growth. The biosynthesis of these amino acids necessitated the conversion of coenzyme NAD^+ to NADH, resulting in a reduction in ethanol production. However, specific amino acid metabolism pathways exhibited a favorable impact on ethanol production, as they are essential for cell growth or ethanol tolerance [33,34]. For instance, the overexpression of genes associated with the tryptophan biosynthesis pathway enhanced ethanol tolerance, while the overexpression of genes involved in ATP and $NADH^-$consuming reactions within the glutamate biosynthesis pathway contributes to increased ethanol production. Moreover, the biosynthesis of phenylalanine, tyrosine, and tryptophan, along with the metabolism

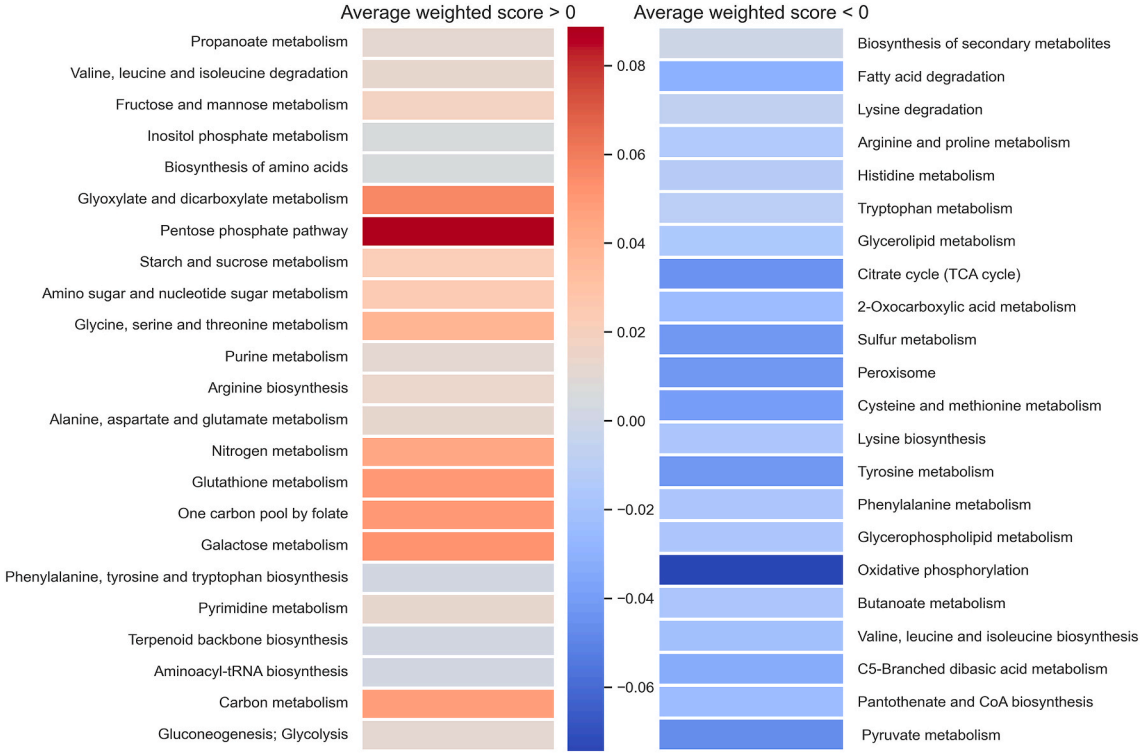

Fig. 3. Distribution of heat map for different metabolic pathways.

of alanine, aspartate, and glutamate, demonstrated a positive correlation with ethanol production. In this study, our specific focus was on pathways exhibiting negative scores to identify potential gene knockout targets associated with ethanol production. Through the evaluation of the weighted scores of metabolic pathways, the oxidative phosphorylation pathway stood out as a promising target for gene knockout identification, characterized by the highest negative weighted score (Fig. 3).

It is worth noting that the identification of the oxidative phosphorylation pathway is based on the correlation between metabolic reactions and their effects on ethanol production, although establishing a definitive causal relationship presents challenges. Therefore, to further elucidate the metabolic mechanism, simulation analyses utilizing GSMM were employed to identify novel gene knockout targets. Fig. S2 presents several reactions with comparably high weight scores within the oxidative phosphorylation pathway. Previous studies have demonstrated that the gene encoding iron-cytochrome c reductase QCR (catalyzing the conversion between ferrocytochrome c and ubiquinone-6) and the cytochrome *c* oxidase COX gene (catalyzing the conversion of ferrocytochrome c to ferricytochrome c) exhibited increased ethanol production in their respective gene deletion mutants [31]. Deletion of COX9 or QCR9 led to a significant 37 % and 27 % increase in ethanol production, respectively, compared to the parental strain, despite minor growth defects. On the other hand, the deletion of QCR6 resulted in a noteworthy 24 % increase in ethanol production without compromising growth. Interestingly, the potential influence of succinate dehydrogenase (SDH), a gene situated between oxidative phosphorylation and the TCA cycle, on ethanol production remains largely unexplored. Succinate dehydrogenase (SDH) holds a crucial position in the TCA cycle and oxidative phosphorylation pathways, as indicated by its involvement in metabolic reactions with notably high weight scores. Notably, GEM-based simulations of single-gene knockouts demonstrated that deleting SDH had minimal impact on growth, rendering it an attractive target for knockout modifications.

3.4. Validation of the hybrid model

3.4.1. Construction of the SDH subunit gene deletion strain

SDH is composed of multiple subunits that function in synergy. However, the effects of gene deletions targeting specific SDH subunits on ethanol production and growth have been sparsely documented. To examine the impact of SDH subunit genes and assembly factor genes on ethanol production, CRISPR/Cas9 technology was employed to individually knock out the sdh1-8 genes. The results demonstrated that, except for the *Δsdh3* strain, all other *sdh* deletion strains exhibited elevated ethanol production in both YPD and YSC media, leading to improvements ranging from 6 % to 10 % (Fig. 4). The engineered strains exhibited distinct growth patterns when cultured in YPD and YSC media. In YSC media, the strains, except for Δsdh3, exhibited unhindered growth and even demonstrated slight enhancements when compared to the BY4741 strain. However, all modified strains demonstrated decreased growth in the YPD medium. Notably, *Δsdh*5 and *Δsdh*6 strains exhibited the most significant increase in ethanol production under both media conditions. These observations suggested that their influence on ethanol production is not dependent on the composition of the growth medium. The experimental findings further supported the validity of the gene knockout targets identified through the hybrid model analysis.

3.4.2. Construction of a high-yielding ethanol-producing strain

Glycerol, a significant byproduct in ethanol production, is tightly regulated by the GPD gene in the glycerol biosynthesis pathway. Numerous studies have consistently indicated that deletion of the GPD gene could result in enhanced ethanol production. Furthermore, the GPD gene plays a pivotal role in maintaining intracellular redox balance. Therefore, CRISPR/Cas9 technology was employed to generate double-gene deletion mutants targeting GPD and SDH in this study. the biomass, ethanol concentration, and relative ethanol production of the engineered strains compared to the parental strain BY4741 were illustrated in Fig. 5. The results revealed a significant enhancement in ethanol production for the engineered strains *Δsdh*4*Δgpd*1, *Δsdh*5*Δgpd*1, *Δsdh*6*Δgpd*1, *Δsdh*4*Δgpd*2, *Δsdh*5*Δgpd*2, and *Δsdh*6*Δgpd*2 when compared to BY4741. However, strains, namely *Δgpd*2*Δsdh*3, *Δgpd*1*Δsdh*1, and *Δgpd*1*Δsdh*2, experienced a noticeable reduction in ethanol production

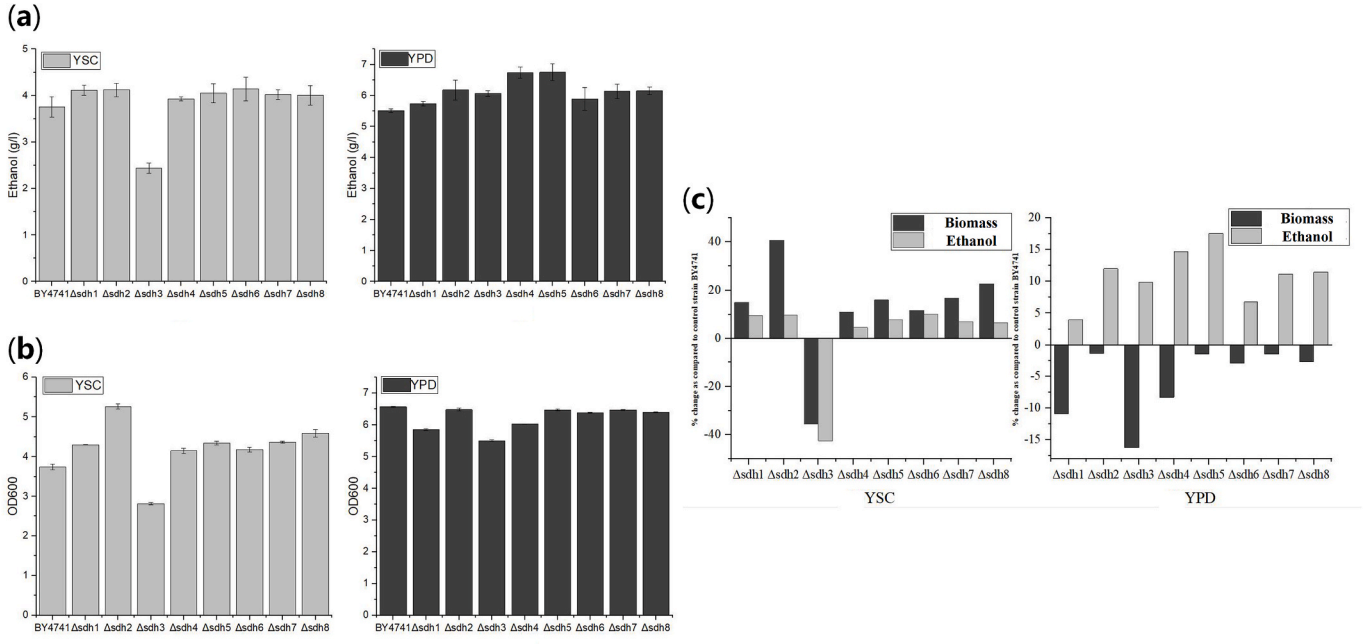

Fig. 4. The ethanol concentration (a) and biomass (b) of different strains on YSC and YPD medium; Biomass and ethanol concentration changing rate of the modified strain on YSC and YPD medium (c).

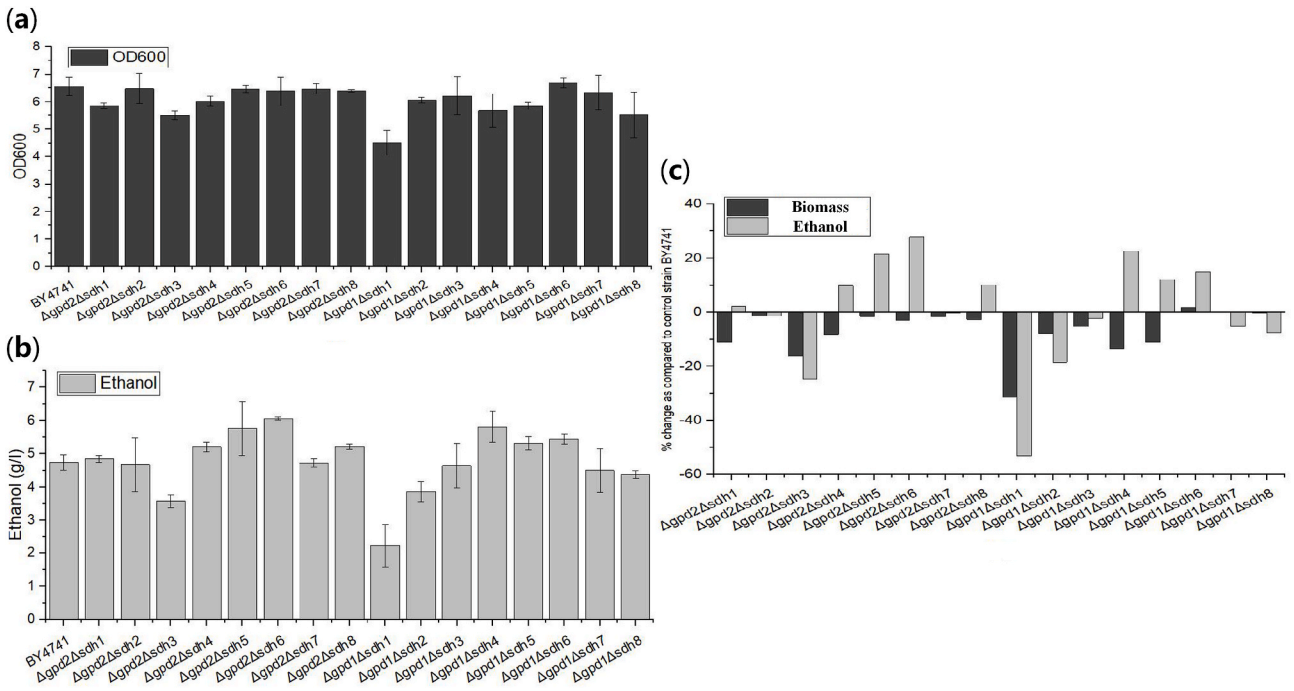

Fig. 5. The growth of the GPD and SDH double-gene knockout strains (a); The ethanol yield of the GPD and SDH double-gene knockout strains (b); and biomass and ethanol concentration changing rate of the GPD and SDH double-gene knockout strains (c).

relative to BY4741. These strains manifested lower OD values compared to BY4741, indicating a dual impact arising from respiratory defects due to the absence of the glycerol biosynthesis pathway and deficiencies in SDH. This situation led to compromised growth and ethanol production capabilities. Among them, *Δgpd*1*Δsdh*1 displayed the most significant effects, with a 30 % decrease in growth and a 50 % reduction in ethanol production compared to BY4741. Importantly, *Δgpd*2*Δsdh*5, *Δgpd*2*Δsdh*6, and *Δgpd*1*Δsdh*4 exhibited the most pronounced increase in ethanol production, with respective increments of 21.6 %, 27.9 %, and 22.7 % compared to BY4741. Notably, *Δgpd*2*Δsdh*6 demonstrated the most substantial improvement in ethanol production, reaching 6.06 g/L under shaken flask conditions. In summary, significant advancements in ethanol production were achieved through data analysis of the hybrid model and the metabolic engineering of strains.

4. Discussion

In this study, we introduce a hybrid model-driven platform that integrates ML algorithms with GSMMs. The primary objective of this platform is to refine the design of biological systems, specifically targeting the augmentation of fluxes within the ethanol biosynthesis pathway through the utilization of data derived from *S. cerevisiae* fermentation. The effectiveness and reliability of the hybrid framework were thoroughly evaluated through the application of various ML approaches. The integration of GSMMs and data-driven techniques establishes a valuable foundation for future benchmarking efforts. In addition, the hybrid model could surpass the insights derived solely from metabolic reconstruction, commonly used for generating flux maps, in terms of both prediction accuracy and biological comprehension.

The integration of ML models with GSMM has garnered increasing interest among researchers in the life sciences. For example, Kim employed ML methods to predict dynamic alterations in metabolic pathways and combined these predictions with GSMMs to unravel the regulatory mechanisms governing these pathways [13]. However, conducting wet experiments for designing biological systems can be costly, time-consuming, and error-prone [11]. To this end, we collected continuous data on the metabolic processes of ethanol, glucose, glycerol, and cell concentration at various time points during the fermentation cycle of *S. cerevisiae* using a soft sensor established with Raman spectroscopy [23]. Subsequently, a GSMM was employed to simulate the fermentation process, leveraging the collected process data as constraints in FBA. By establishing a correlation between the metabolic flux data and actual ethanol production, the hybrid model facilitated the prediction of ethanol yield. In conclusion, the integration of flux data from GSMM with biological data enhanced the analytical capabilities of ML models, leading to a deeper understanding of the intricate relationship between metabolic fluxes and ethanol production in *S. cerevisiae*. This integration enabled a comprehensive analysis that shed light on the underlying mechanisms governing ethanol biosynthesis. Notably, the study primarily aims to elucidate the impact of genetic modifications on ethanol production. However, it is recognized that the attained titer falls short of the theoretical yields potentially achievable by *S. cerevisiae*. Subsequent investigations could include fed-batch experiments designed to evaluate the industrial potential and robustness of the engineered strains, and optimize process conditions for maximizing ethanol production.

To investigate the effects of SDH gene deletion on intracellular metabolic flux distribution and mechanisms, we conducted single-gene knockouts of SDH using the yeast 8.0.0 model. The average metabolic flux distributions surrounding the oxidative phosphorylation pathway were calculated for both the wild-type strain (BY4741) and the SDH deletion strain (*ΔSDH*). The calculations yielded the average flux values for specific metabolic pathways, providing insights into the effects of SDH deletion. Fig. 6 illustrates the influence of SDH deletion on some metabolic pathways by GSMM simulations. The findings demonstrated a notable reduction in metabolic fluxes within the TCA cycle, oxidative phosphorylation, and the pyruvate, aspartate, and glutamate metabolism pathways in the *ΔSDH* strain compared to the wild-type strain. Given the previous discussion on the impact of negatively weighted pathways on ethanol productivity, the deletion of SDH might partially hinder the activity of these metabolic pathways, leading to an enhancement in ethanol production. In contrast, the glycolytic pathway exhibits enhanced metabolic activity. As indicated by the positively

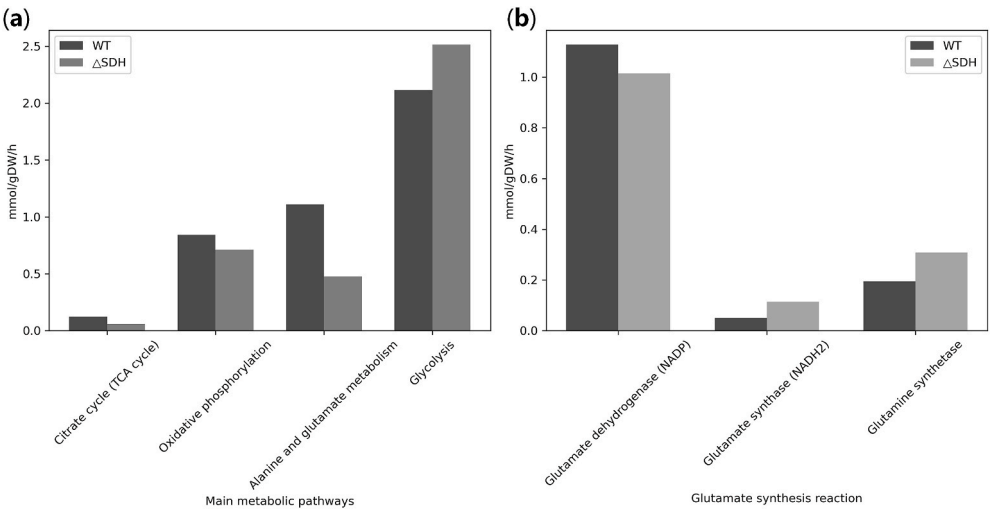

Fig. 6. The effect of SDH gene knockout on metabolic pathways by GSMM simulation.

weighted reactions in the metabolic network distributions, the enhancement of the pentose phosphate pathway and glycolysis might hold a crucial role in promoting favorable ethanol metabolism. In addition, previous studies have proved that the deletion of the SDH2 subunit gene of succinate dehydrogenase resulted in a decrease in organic acids, such as citric acid, while concurrently increasing acetone accumulation in the glycolytic pathway [35,36]. These metabolic changes have been found to positively influence ethanol production.

Furthermore, deletion of the *sdh* gene could affect the oxidative phosphorylation pathway, which was consistent with previous reports [37–39]. Deletion mutants of the *sdh* gene in *S. cerevisiae* exhibit respiratory system defects that limit oxygen consumption, thereby promoting increased ethanol production [40,41]. Interestingly, unexpected outcomes were observed in the metabolic pathways involving pyruvate, aspartate, and glutamate. Despite the overall decrease in metabolic activity within these pathways, the metabolic flux of crucial reactions governed by glutamate synthase and glutamine synthetase in the glutamate biosynthesis pathway increases, whereas the metabolic flux of reactions regulated by glutamate dehydrogenase decreases. Previous studies have demonstrated that manipulating the expression of key genes in these pathways, such as GDH1/GDH3 for glutamate dehydrogenase, GLT1 for glutamate biosynthesis, and GLN1 for glutamine synthetase, can enhance redox balance, reduce glycerol byproduct formation, and promote ethanol production [42,43].

Metabolic fluxes possess a fundamental mechanistic interpretation owing to its robust association with underlying biochemistry. Notably, ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) provides a more accurate understanding of intracellular flux distributions when comparing with GSMM [44]. Likewise, ^{13}C-MFA often assumes that the system is in a quasi-steady state during the labeling experiment. This assumption may not hold in rapidly changing or dynamic metabolic systems. In addition, ^{13}C-labeling experiments can be expensive and time-consuming, especially when considering the production of labeled substrates, sample preparation, and the analysis of isotopic labeling patterns. In contrast, we initially utilized the previously constructed online monitoring model to acquire crucial physiological metabolic parameters during the process. By iteratively differentiating the long-term dynamic process (transitioning from hours or days to minutes), we assumed that cellular metabolism was at a steady state within short time intervals. This approach allows for the continuous inclusion of the dynamic distribution of intracellular metabolic fluxes in the monitoring of biological processes. Therefore, the integration of data-driven metabolic network models not only enhanced predictive capabilities but also yielded valuable mechanistic insights into metabolite interactions under specific conditions, significantly contributing to phenotypic outcomes. This integration confers distinct advantages by facilitating the development and utilization of more biologically interpretable ML models, particularly in scenarios where understanding the effects of cellular or metabolic engineering manipulations is paramount [10]. Hence, our findings strongly support the expansion of this robust hybrid model, driven by data and knowledge, to encompass bioengineering and other related objectives on phenotypic outcomes, such as the secretion of metabolites for drug development purposes.

Data availability

The datasets used in this study can be found in the Supplementary Material or requested to the corresponding author.

CRediT authorship contribution statement

Debiao Wu: Data curation, Visualization, Writing – original draft. **Feng Xu:** Data curation, Visualization, Writing – original draft. **Yaying Xu:** Data curation, Visualization. **Mingzhi Huang:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Zhimin Li:** Supervision, Writing – review & editing. **Ju Chu:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing, All authors have read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant NO. 32071461) and the National Key Research and Development Program of China (Grant NO. 2019YFA0904300).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2023.12.004.

References

[1] Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. Nature 2012;485:185–94. https://doi.org/10.1038/nature11117.

[2] Nielsen J, Keasling JD. Engineering cellular metabolism. Cell 2016;164:1185–97. https://doi.org/10.1016/j.cell.2016.02.004.

[3] Chao R, Mishra S, Si T, Zhao H. Engineering biological systems using automated biofoundries. Metab Eng 2017;42:98–108. https://doi.org/10.1016/j.ymben.2017.06.003.

[4] Du J, Shao Z, Zhao H. Engineering microbial factories for synthesis of value-added products. J Ind Microbiol Biotechnol 2011;38:873–90. https://doi.org/10.1007/s10295-011-0970-3.

[5] Liu Y, Shin HD, Li J, Liu L. Toward metabolic engineering in the context of system biology and synthetic biology: advances and prospects. Appl Microbiol Biotechnol 2015;99:1109–18. https://doi.org/10.1007/s00253-014-6298-y.

[6] Chen Y, Nielsen J. Advances in metabolic pathway and strain engineering paving the way for sustainable production of chemical building blocks. Curr Opin Biotechnol 2013;24:965–72. https://doi.org/10.1016/j.copbio.2013.03.008.

[7] HamediRad M, Chao R, Weisberg S, Lian J, Sinha S, Zhao H. Towards a fully automated algorithm driven platform for biosystems design. Nat Commun 2019;10:5150. https://doi.org/10.1038/s41467-019-13189-z.

[8] Thurow K. Automation for life science laboratories. Adv Biochem Eng Biotechnol 2022;182:3–22. https://doi.org/10.1007/10_2021_170.

[9] King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, Kell DB, Oliver SG. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 2004;427:247–52. https://doi.org/10.1038/nature02236.

[10] Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput Biol 2019;15:e1007084. https://doi.org/10.1371/journal.pcbi.1007084.

[11] Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proc Natl Acad Sci U S A 2020;117:18869–79. https://doi.org/10.1073/pnas.2002959117.

[12] Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E. Metabolic network prediction of drug side effects. Cell Syst 2016;2(3):209–13. https://doi.org/10.1016/j.cels.2016.03.001.

[13] Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat Commun 2016;7:13090. https://doi.org/10.1038/ncomms13090.

[14] Yaneske E, Angione C. The poly-omics of ageing through individual-based metabolic modelling. BMC Bioinf 2018;19:415. https://doi.org/10.1186/s12859-018-2383-z.

[15] Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO, Walker GC, Collins JJ. A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell 2019;177:1649–1661.e9. https://doi.org/10.1016/j.cell.2019.04.016.

[16] Ben Guebila M, Thiele I. Predicting gastrointestinal drug effects using contextualized metabolic models. PLoS Comput Biol 2019;15(6):e1007100. https://doi.org/10.1371/journal.pcbi.1007100.

[17] Lelieveld J, Klingmüller K, Pozzer A, Burnett RT, Haines A, Ramanathan V. Effects of fossil fuel and total anthropogenic emission removal on public health and climate. Proc Natl Acad Sci U S A 2019;116:7192–7. https://doi.org/10.1073/pnas.1819989116.

[18] Weber A, Kalema-Zikusoka G, Stevens NJ. Lack of rule-adherence during mountain Gorilla tourism encounters in bwindi impenetrable national park, Uganda, places Gorillas at risk from human disease. Front Public Health 2020;8:1. https://doi.org/10.3389/fpubh.2020.00001.

[19] Adebami GE, Kuila A, Ajunwa OM, Fasiku SA, Asemoloye MD. Genetics and metabolic engineering of yeast strains for efficient ethanol production. J Food Process Eng 2022;45:e13798. https://doi.org/10.1111/jfpe.13798.

[20] Xu F, Lu J, Ke X, Shao M, Huang M, Chu J. Reconstruction of the genome-scale metabolic model of saccharopolyspora erythraea and its application in the overproduction of erythromycin. Metabolites 2022;12:509.

[21] Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, Beber ME, Sonnenschein N, Kerkhoven EJ, Nielsen J. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 2019;10:3586. https://doi.org/10.1038/s41467-019-11581-3.

[22] Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc 2019 Mar;14(3):639–702. https://doi.org/10.1038/s41596-018-0098-2.

[23] Wu DB, Xu YY, Xu F, Shao MH, Huang MZ. Machine learning algorithms for in-line monitoring during yeast fermentations based on Raman spectroscopy. PREPRINT (Version 1). 2023. https://doi.org/10.21203/rs.3.rs-2615036/v1. Research Square.

[24] Quarterman J, Kim SR, Kim PJ, Jin YS. Enhanced hexose fermentation by Saccharomyces cerevisiae through integration of stoichiometric modeling and genetic screening. J Biotechnol 2015;194:48–57. https://doi.org/10.1016/j.jbiotec.2014.11.017.

[25] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765–74. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[26] Xu Y, Li Z. Utilization of ethanol for itaconic acid biosynthesis by engineered Saccharomyces cerevisiae. FEMS Yeast Res 2021;21:foab043. https://doi.org/10.1093/femsyr/foab043.

[27] Xu Y, Li Z. Alleviating glucose repression and enhancing respiratory capacity to increase itaconic acid production. Synth Syst Biotechnol 2022;8:129–40. https://doi.org/10.1016/j.synbio.2022.12.007.

[28] Zhang GC, Kong II, Kim H, Liu JJ, Cate JH, Jin YS. Construction of a quadruple auxotrophic mutant of an industrial polyploid saccharomyces cerevisiae strain by using RNA-guided Cas9 nuclease. Appl Environ Microbiol 2014;80:7694–701. https://doi.org/10.1128/AEM.02310-14.

[29] Naghshbandi MP, Tabatabaei M, Aghbashlo M, Gupta VK, Sulaiman A, Karimi K, et al. Progress toward improving ethanol production through decreased glycerol generation in Saccharomyces cerevisiae by metabolic and genetic engineering approaches. Renew Sustain Energy Rev 2019;115:109353. ARTN109353.

[30] Chu BC, Lee H. Genetic improvement of Saccharomyces cerevisiae for xylose fermentation. Biotechnol Adv 2007;25:425–41. https://doi.org/10.1016/j.biotechadv.2007.04.001.

[31] Quarterman J, Kim SR, Kim PJ, Jin YS. Enhanced hexose fermentation by Saccharomyces cerevisiae through integration of stoichiometric modeling and genetic screening. J Biotechnol 2015;194:48–57. https://doi.org/10.1016/j.jbiotec.2014.11.017.

[32] Wang P-M, Zheng D-Q, Ding R, Chi X-Q, Tao X-L, Min H, et al. Improvement of ethanol production in Saccharomyces cerevisiae by hetero-expression of GAPN and FPS1 deletion. J Chem Technol Biotechnol 2011;86:1205–10. https://doi.org/10.1002/jctb.2634.

[33] Ma M, Liu ZL. Mechanisms of ethanol tolerance in Saccharomyces cerevisiae. Appl Microbiol Biotechnol 2010;87(3):829–45. https://doi.org/10.1007/s00253-010-2594-3.

[34] Fujita K, Matsuyama A, Kobayashi Y, Iwahashi H. The genome-wide screening of yeast deletion mutants to identify the genes required for tolerance to ethanol and other alcohols. FEMS Yeast Res 2006;6(5):744–50. https://doi.org/10.1111/j.1567-1364.2006.00040.x.

[35] Zahoor A, Küttner FTF, Blank LM, Ebert BE. Evaluation of pyruvate decarboxylase-negative Saccharomyces cerevisiae strains for the production of succinic acid. Eng Life Sci 2019;19(10):711–20. https://doi.org/10.1002/elsc.201900080.

[36] Raab AM, Gebhardt G, Bolotina N, Weuster-Botz D, Lang C. Metabolic engineering of Saccharomyces cerevisiae for the biotechnological production of succinic acid. Metab Eng 2010;12:518–25. https://doi.org/10.1016/j.ymben.2010.08.005.

[37] Arikawa Y, Kuroyanagi T, Shimosaka M, Muratsubaki H, Enomoto K, Kodaira R, Okazaki M. Effect of gene disruptions of the TCA cycle on production of succinic acid in Saccharomyces cerevisiae. J Biosci Bioeng 1999;87:28–36. https://doi.org/10.1016/s1389-1723(99)80004-8.

[38] Lemire BD, Oyedotun KS. The Saccharomyces cerevisiae mitochondrial succinate: ubiquinone oxidoreductase. Biochim Biophys Acta 2002;1553:102–16. https://doi.org/10.1016/s0005-2728(01)00229-8.

[39] Tzagoloff A, Dieckmann CL. PET genes of Saccharomyces cerevisiae. Microbiol Rev 1990;54:211–25. https://doi.org/10.1128/mr.54.3.211-225.1990.

[40] Hutter A, Oliver SG. Ethanol production using nuclear petite yeast mutants. Appl Microbiol Biotechnol 1998;49:511–6. https://doi.org/10.1007/s002530051206.

[41] Kim SR, Lee KS, Choi JH, Ha SJ, Kweon DH, Seo JH, Jin YS. Repeated-batch fermentations of xylose and glucose-xylose mixtures using a respiration-deficient Saccharomyces cerevisiae engineered for xylose metabolism. J Biotechnol 2010;150:404–7. https://doi.org/10.1016/j.jbiotec.2010.09.962.

[42] Kim JW, Chin YW, Park YC, Seo JH. Effects of deletion of glycerol-3-phosphate dehydrogenase and glutamate dehydrogenase genes on glycerol and ethanol metabolism in recombinant Saccharomyces cerevisiae. Bioproc Biosyst Eng 2012;35:49–54. https://doi.org/10.1007/s00449-011-0590-3.

[43] Kong QX, Gu JG, Cao LM, Zhang AL, Chen X, Zhao XM. Improved production of ethanol by deleting FPS1 and over-expressing GLT1 in Saccharomyces cerevisiae. Biotechnol Lett 2006;28:2033–8. https://doi.org/10.1007/s10529-006-9185-5.

[44] Xu F, Lu J, Ke X, Shao M, Huang M, Chu J. Reconstruction of the genome-scale metabolic model of saccharopolyspora erythraea and its application in the overproduction of erythromycin. Metabolites 2022;12(6):509. https://doi.org/10.3390/metabo12060509.