Optimizing the Production of Valuable Metabolites using a Hybrid of Constraint-based Model and Machine Learning Algorithms: A Review

Kauthar Mohd Daud[1], Ridho Ananda[2], Suhaila Zainudin[3],
Chan Weng Howe[4], Kohbalan Moorthy[5], Nurul Izrin Binti Md Saleh[6]
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia[1,2,3]
Institut Teknologi Telkom Purwokerto, Indonesia[2]
UTM Big Data Centre, Faculty of Computing, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor Malaysia[4]
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang Malaysia[5]
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia[6]

***Abstract*—The advances in genome sequencing and metabolic engineering have allowed the reengineering of the cellular function of an organism. Furthermore, given the abundance of omics data, data collection has increased considerably, thus shifting the perspective of molecular biology. Therefore, researchers have recently used artificial intelligence and machine learning tools to simulate and improve the reconstruction and analysis by identifying meaningful features from the large multi-omics dataset. This review paper summarizes research on the hybrid of constraint-based models and machine learning algorithms in optimizing valuable metabolites. The research articles published between 2020 and 2023 on machine learning and constraint-based modeling have been collected, synthesized, and analyzed. The articles are obtained from the Web of Science and Scopus databases using the keywords: "Machine learning", "flux balance analysis", and "metabolic engineering". At the end of the search, this review contained 13 records. This review paper aims to provide current trends and approaches in *in silico* metabolic engineering while providing research directions by highlighting the research gaps. In addition, we have discussed the methodology for integrating machine learning and constraint-based modeling approaches.**

***Keywords*—Flux balance analysis; genome-scale metabolic model; machine learning; metabolic engineering**

I. Introduction

Microorganisms have been used in industrial sectors such as food processing, chemical manufacturing, pharmaceuticals, fermentation, and others. Advances in genome sequencing have resulted in several innovations that allow researchers to gain in-depth knowledge and information about an organism. One of these advancements is metabolic engineering, which reengineers the cellular function of an organism. In the 1990s, metabolic engineering was introduced to describe recombinant DNA technology for optimizing microbial activity [1]. Metabolic engineering aims to optimize the synthesis of desired metabolites by directing the metabolic flow and the fluxes toward the desired metabolites. The designs are categorized into two types: [1] targeting metabolic network components, such as gene/reaction knockout/knock-in, and [2] enhancing the metabolic network by altering it using network reconstruction tools or incorporating new non-native pathways into the host.

Over the previous few decades, there has been a noticeable breakthrough, such as incorporating adenosylcobinamide phosphate biosynthesis from *Rhadobacter capsulatus* into the *E.coli* strain, which improves the vitamin B_{12} to 307 $\mu g/g$ [2]. In another case, the yeast was engineered to improve the production of rubusoside and rebaudiosides, leading to 1368.6 mg/L and 132.7 mg/L, respectively [3]. Although metabolic pathway optimization technologies have shown promise, an incomplete understanding of the connection between target cell phenotype and genotype impedes their further development. This results in the prevalent utilization of conventional trial-and-error methodologies and indirectly remains tedious, costly, and time-consuming.

Therefore, constraint-based modeling (CBM) approaches have been used to analyze organisms by providing significant phenotypic knowledge based on genotypic perturbations. CBM approaches, which include Flux Balance Analysis (FBA) and its variants (Minimization of Metabolic Adjustment, MoMA; Regulatory on/off minimization, ROOM; and Flux Variability Analysis, FVA), are used to reveal metabolic phenotypes by analyzing the optimality of an organism [4], [5]. However, a significant challenge in CBM is that the desired flux is not limited to a single solution due to biological network redundancy and complex genome-scale metabolic model (GSMM), thus permitting alternate optimum solutions. Furthermore, due to the intricacy and interdependence of components in the metabolic network, selecting appropriate and optimal reactions/genes for knockout is difficult, laborious, and time-consuming [6]. Hence, previous research has combined metaheuristic optimization algorithms such as genetic algorithm (GA), differential search algorithms (DSA), flower pollination algorithm (FA), and others [7], [8], [9], [10].

With the recent advancement of high-throughput technology and the overwhelming amount of omics data, data collection has increased considerably, thus shifting the perspective

on molecular biology [11]. Although big data in biology enables data-driven science to comprehend complex biological systems and events, interpreting data is still complicated. Therefore, machine learning (ML) has been applied to deal with biological omics data for various applications such as prediction, classification, and discovery. The involvement of ML in the data shows a great potential to reveal hidden and detailed information in the data.

It has proven successful in diabetes disease prediction, optical character recognition, face identification, and others [12], [13], [14], [15]. ML is a set of algorithms to improve prediction accuracy by learning and analyzing the patterns from large experimental datasets. Recently, ML has been applied to increase the accuracy of the genotype-phenotype relationship by analyzing the integrated metabolic networks with regulatory or signaling networks. Furthermore, ML requires fewer parameters than other statistical or computation approaches, thus making them useful for various tasks, including predicting the impact of genetic perturbations, reconstructing phylogenetic trees, and others [16], [17].

This paper aims to review how ML techniques are applied in metabolic engineering, specifically to optimize the production of desired metabolites. The paper is organized as follows: Section II introduces the definition of metabolic engineering. Section III provides a brief on constraint-based modeling. Section IV discusses machine learning in metabolic engineering. Then, applications of machine learning in metabolic engineering have been described in Section V. After that results and discussion are provided in Section VI. In the last, the conclusion is given in Section VII.

II. Metabolic Engineering

Each component in biological systems plays a vital role in biological processes and interacts with each other. Therefore, it is crucial to analyze the systems as a whole. The organism's function can be divided into three major biochemical pathways: gene regulatory, signal transduction, and metabolic networks. Gene regulatory involves a set of genes, proteins, and their regulatory mechanisms that determine the expression of the gene. Signal transduction networks communicate between and within cells by mediating, detecting, amplifying, and integrating various external and internal stimuli to govern and coordinate cellular activities. Meanwhile, the metabolic network is a series of biochemical reactions involving the transformation and modification of substrates into different products in which the enzymes act as catalysis agents. The metabolic network is essential in assessing a cell's biochemical and physiological properties. This research is mainly concerned with metabolic networks.

Advancements in genome sequencing have brought about many developments that allow biological researchers to have more profound knowledge and information about an organism. One of the developments is the establishment of metabolic engineering (ME), which allows the researchers to probe in detail the organizations of an organism, including the reactions, pathways, metabolites, and genes, and exploit the organisms for strain optimization. Metabolic engineering aims to optimize the metabolism of organisms by exploiting and manipulating their metabolic capabilities through modeling and, thus, generates economically and industrially viable organisms through optimization and predictive tools. In order to achieve this objective, it is necessary to adapt current metabolic engineering approaches by incorporating automated simulation techniques instead of relying on previous in vivo or in vitro investigations.

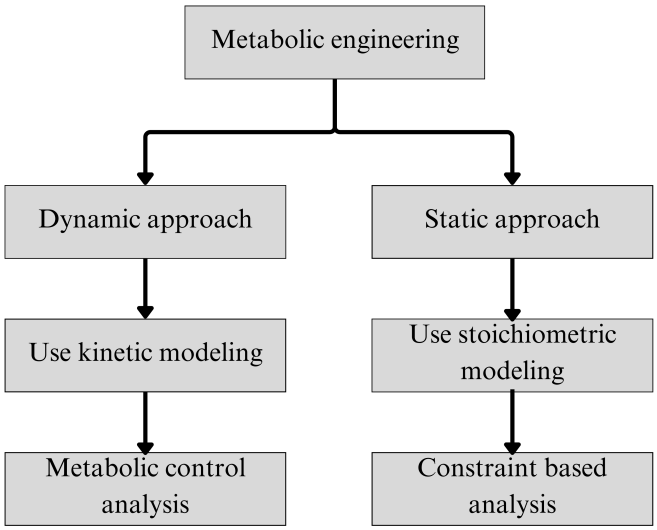

Fig. 1. Approaches in metabolic engineering.

In order to exploit and manipulate the metabolic capabilities of an organism, the metabolic pathways within the cell need to be modeled. A model is a simplified system representation that allows the user to understand, predict, and control the system [18]. An organism can be modeled based on a dynamic or static approach. In ME, the metabolism of the target organism was represented in the mathematical model. Thus, the network's precise respective pathway or reactions that need to be manipulated and optimized can be identified. Various computational modeling approaches and algorithms have been developed and applied to aid the researchers [19]. Different approaches have been developed depending on the representations, as shown in Fig. 1.

The approaches in metabolic engineering can be divided into two, which are the dynamic approach and the static approach. Each approach varies in terms of metabolism representation, whereby the dynamic approach uses kinetic modelling and static approach uses a stoichiometric matrix to represent the metabolic network [20]. Furthermore, the difference between these two approaches is the model used. The dynamic approach uses a kinetic model, and the static approach uses a stoichiometric model or a metabolic network. Both of these models consist of different information and representations. The dynamic approach describes the changes in metabolite concentrations over time, while the static approach does not [21]. Table I defines the difference between the kinetic and the stoichiometric models.

In stoichiometric models, the biochemical reactions in the metabolic network are represented as a set of stoichiometric equations, whereby the elements of different metabolites in the metabolic network are denoted as stoichiometric coefficients in the stoichiometric matrix. Consequently, the intracellular metabolic fluxes can be determined at the steady state using the

TABLE I. DIFFERENCES BETWEEN THE KINETIC AND STOICHIOMETRIC MODELS

Characteristics	Kinetic model	Stoichiometric model
Definition	Describes changes in metabolite concentrations over time.	Assumes the system is at steady-state conditions, where the concentrations of the metabolites are constant over time.
What information resides in the model?	1) Metabolites concentrations 2) Kinetic parameters	Stoichiometric information of all specified reactions and genes
How do they represent the model?	Ordinary differential equations (ODE)	Linear equation
Size of the metabolic network for applicability	Small-scale metabolic network	Large-scale metabolic network
How do they work?	It uses kinetic rate laws obtained from biochemical and mechanistic information.	Imposes constraints and objective functions
Drawback	1) Requires many parameters 2) Sometimes leads to uncertainty in the model prediction 3) Not fully utilized in ME	1) Lead to underdetermine system; the number of equations is larger than the number of variables 2) Generate many possible solutions 3) Solutions might not be unique
Time-consuming	High	Low
Computational extensive	High	Low
Accuracy	High	Low

mass balance constraints. However, stoichiometric models are often underdetermined and eventually lead to many possible non-unique solutions. Thus, the models require additional constraints to narrow the range of possible phenotypic solutions. These constraints may include physicochemical, biological, mass conservation, and thermodynamics. Stoichiometric models have been used to enumerate the fluxes in a metabolic network by employing an objective function. The main application of stoichiometric models is on metabolic networks, specifically in metabolic engineering strategies [7], [8], [22], [23].

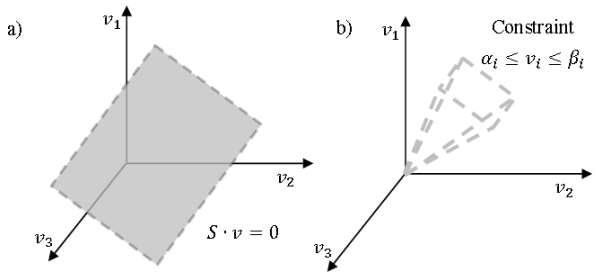

Fig. 2. Unconstrained (a) and Constrained (b) Solution space.

III. CONSTRAINT-BASED MODELING

The constraint-based method (CBM) is an approach to investigating the optimality of an organism by predicting and describing the metabolic phenotypes [24]. In CBM, constraints are applied to the systems, thus creating feasible flux distribution space. Different types of constraints can be categorized into physicochemical, topo-biological, environmental, and regulatory [25], [26]. These constraints can be expressed as equality or inequality constraints, as shown below, and have been reviewed by [26]. The equation that describes the incoming and outgoing fluxes accumulation for each metabolite in the metabolic network is described in 1.

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{S} \times \mathbf{v} \tag{1}$$

where $\mathbf{S}$ is the stoichiometric matrix of size $m \times n$ (m is the number of metabolites and n is the number of reactions), $\mathbf{X}$ is the m concentration vector, and v is the n flux vector. Each metabolite's production rate must equal the consumption rate at the steady state. Therefore, the above equation is simplified to Eq. (2).

$$\mathbf{S} \times \mathbf{v} = \mathbf{0} \tag{2}$$

The imposition of constraints will further reduce the number of allowable flux distributions and constraints taken upon the form in Eq. (3).

$$\alpha_i \leq v_i \leq \beta_i \tag{3}$$

where i is the length of m reactions, α_i and β_i are the lower and upper limits for the i reaction, respectively. The values for α_i and β_i are determined based on reactions' reversibility or irreversibility and measured uptake rates. These constraints may restrict specific phenotypes from existing in the solution space. Fig. 2 illustrates the differences between unconstrained and constrained solution space of feasible steady-state flux distributions.

As shown in Fig. 2, unconstrained steady-state solution space is underdetermined due to the ratio of reactions typically exceeding the number of metabolites. Eq. (1) provides a hyperplane that defines the allowable flux distributions. Considering different constraints, the solution space is limited to specific desired phenotypes. Therefore, CBM aims to describe and predict the desired phenotypes of an organism by describing the metabolic networks of an organism using the stoichiometric framework and a series of constraints. Despite the imposition of constraints and steady-state assumption, the solutions generated are not limited to a single solution. Instead, the solutions generated are limited to the desired phenotypes.

In order to solve the underdetermined system, the problem of measuring internal fluxes is solved using an optimization problem [28]. Thus, an objective function is defined, as illustrated in Fig. 3. Generally, an objective function is a biological assumption that an organism can be achieved. Then, linear optimization is used to find the solution that optimizes the desired objective function. Examples of objective functions include minimizing ATP production and nutrient uptake and maximizing growth rate. The most common objective function is growth rate since organisms maximize their growth after evolutionary pressures [29]. Referring to the above equations, Eq. 1 to 3, the objective function for maximizing the growth rate is mathematically represented by Eq. 4.

$$\max Z = v_{biomass} \tag{4}$$

Generally, there are four CBM approaches - flux balance analysis (FBA), flux variability analysis (FVA), minimization of metabolic adjustment (MoMA), and regulatory on/off minimization (ROOM). Table II portrays the characteristics of the four CBM approaches and the applications that have been carried out.

As shown in Table II, FBA is a classical CBM method and has become one of the most common approaches researchers use [7], [8], [25], [30], [31]. Despite FBA's non-uniqueness due to the exclusion of regulatory and kinetic parameters, FBA excels in handling vast data within metabolic networks compared to other approaches, such as predicting higher steady states for biological objectives such as growth rate and production rate. Moreover, despite the incompleteness of metabolic network models, FBA can still determine the organism's steady-state fluxes.

FVA employs linear programming to identify multiple biologically optimal solutions with the same objective value. These solutions are non-unique due to the metabolic network's ability to achieve the same objective value through different equivalent pathways, often represented by recessive phenotypes. Unlike FBA, which examines the distribution of flux within pathways, FVA focuses on determining the feasible ranges of minimum and maximum fluxes for each reaction. Meanwhile, MoMA employs quadratic programming to minimize the Euclidean distance on flux space between the wild-type and mutant, while ROOM predicts the post-genetic perturbation steady state of metabolic networks. In contrast to MoMA, ROOM identifies flux distributions that yield high-rate solutions while minimizing flux deviations between wild-type and mutant and preserving the linearity of fluxes based on experimental measurements [10], [32]. Additionally, ROOM can discover shorter alternative pathways for rerouting fluxes after genetic perturbations, employing mixed integer linear programming (MILP) to meet the same constraints as FBA.

IV. Machine Learning in Metabolic Engineering

In silico metabolic engineering comprises computer simulations that predict and analyze an organism's metabolic network to improve the organism's cellular activities [8]. The improvement involves manipulating metabolic, signal, or regulatory networks. One approach to investigating the effects of genetic changes on metabolite synthesis is in silico reaction knockout modeling. The organism's behavior can be predicted through constraint-based modeling (CBM) methods by analyzing the effects of phenotypic and genotypic perturbations on the organisms.

High-throughput technologies such as gene sequencing, protein purification/quantification, mass spectrometry, and others have enabled a new era of biological information in which the amount of biological data has significantly expanded over

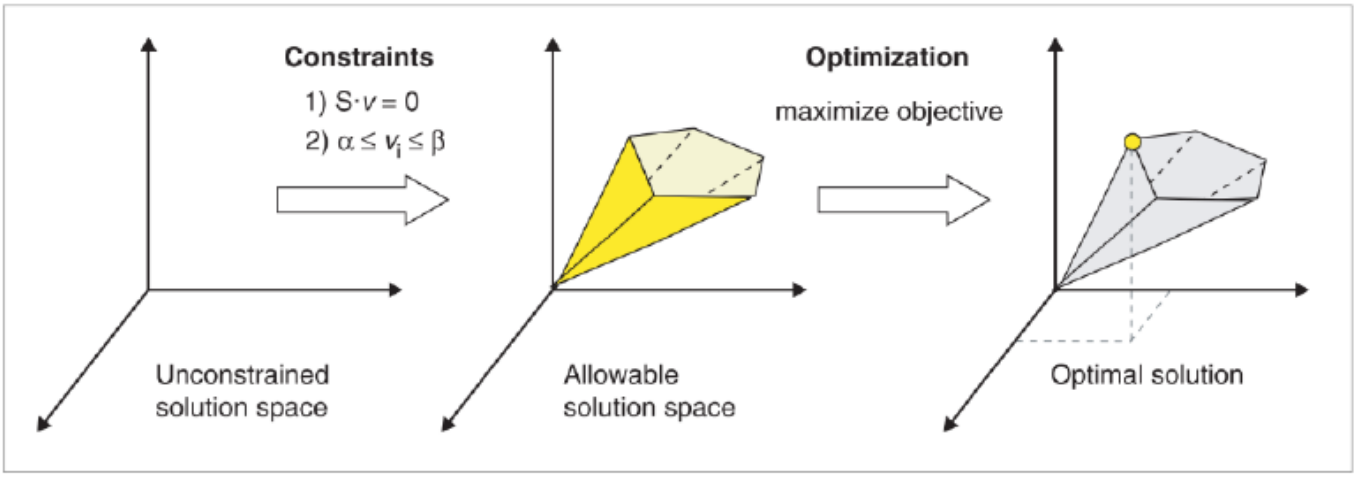

Fig. 3. The conceptual basis of CBM [27].

TABLE II. SUMMARY OF CONSTRAINT-BASED MODELING APPROACHES

Characteristics	FBA	FVA	MoMA	ROOM
Purpose	Measure the optimal flux value	Measure the ranges of each flux	Compare the steady-state fluxes between mutant and wild-type	Minimize the number of significant flux changes between mutant and wild-type
Optimization model	Linear programming	Linear programming	Quadratic programming	Mixed-integer LP
Able to predict the lethality of genes?	Yes	No	No	Yes
Computational time	Short	Long	Long	Long
Size of model	Large	Large	Small	Large
Predicted solutions	Multiple optimal solutions	Assess the robustness of flux distribution	Transient metabolic states	The predicted solutions are nearer to the experimental data

time. The various omics biological datasets, ranging from genomic to metabolomic and fluxomic, can provide direct insight into an organism's phenotype. An alternative approach is therefore needed to analyze and process large amounts of information quickly. Machine Learning (ML) has been increasingly used in metabolic engineering to replace human metabolic engineers [33], [34], [35]. Given its success in pattern recognition, model prediction, and others [36], [37], [38], [39], [40].

Machine learning (ML) is used to generate trial-and-error inferences and improve the predictions from data without a predefined set of rules. ML has been massively used in data analysis and typically allows applications to develop intelligently by understanding patterns in big data [1]. There are two types of ML based on data: labeled and unlabeled (Fig. 4). For the labeled data, algorithms learn from labeled training data to help predict the outcomes of unlabeled data. Meanwhile, unlabeled data use unsupervised learning to seek patterns and clusters in an unlabeled dataset. Examples of supervised learning algorithms include decision trees [41], support vector machines [42], and regression [43], whereas Principal Component Analysis (PCA) [44], [45] and K-means clustering [46] are unsupervised learning algorithms. Another ML type is reinforcement learning, in which the algorithm interacts with experience and learns to maximize the desired goal using experience, data, and trial-and-error interactions. Reinforcement learning does not need labeled input/output but focuses on balancing exploration and exploitation.

ML has recently played a significant role in biological research [16], [39]. These algorithms focus on model performance by training highly heterogeneous data. It is undoubtedly an opportunity to integrate ML algorithms with CBM models in various biological data sets such as gene expression, metabolites, phenotypes, and others [4], [47]. The application of ML in metabolic engineering will provide several benefits. First, ML can be used in various *in silico* metabolic engineering stages, from analyzing the metabolic flux data to designing optimal metabolic pathways. Second, the full integration of omics data, including genomic, transcriptomic, proteomic, and

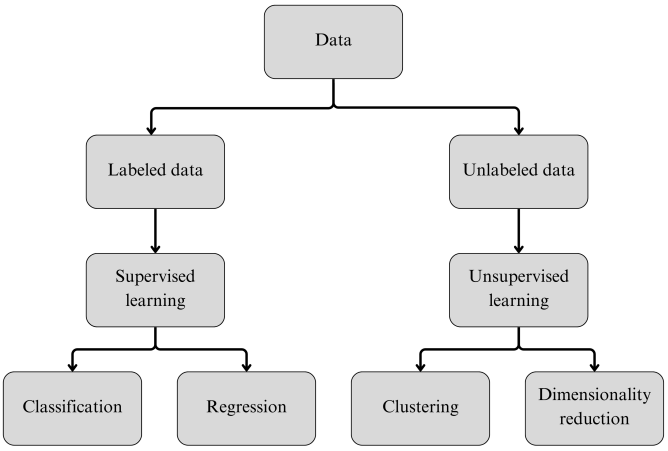

Fig. 4. Machine learning categories.

metabolomic data, is crucial for predicting the metabolic pathway as it provides valuable insight into biological networks [48]. Furthermore, via gene expression analysis using ML, the key regulators of a metabolic pathway can be identified based on the genetic perturbations on cellular metabolism.

Therefore, by merging machine learning with other computational tools in metabolic engineering, researchers may optimize cellular metabolism for enhanced production of biofuels, chemicals, and other essential molecules in a quick, cost-effective, and sustainable way. As shown in Fig. 5, the reactions and metabolites from GSMM are extracted and represented in a stoichiometric matrix. These datasets comprise instances (reactions and metabolites involved in the specific pathway). The coefficient in the stoichiometric matrix represents the knockout (coefficient one) and non-knockout reactions (coefficient zero) involved in that pathway. In this case, different combinations of knockout reactions are obtained. The training data, then, is used to train the chosen ML algorithms and predict the response of the test dataset. The responses

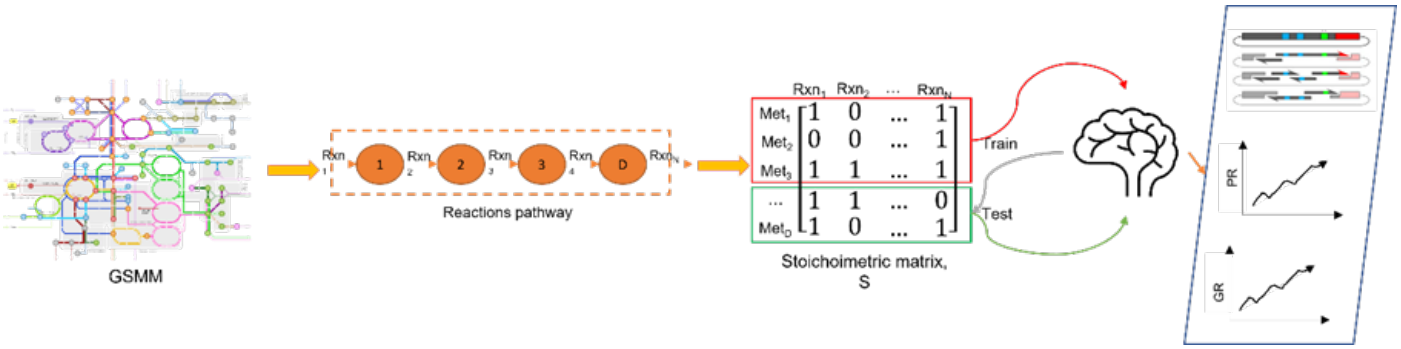

Fig. 5. Overview of the standard workflow of ML in ME.

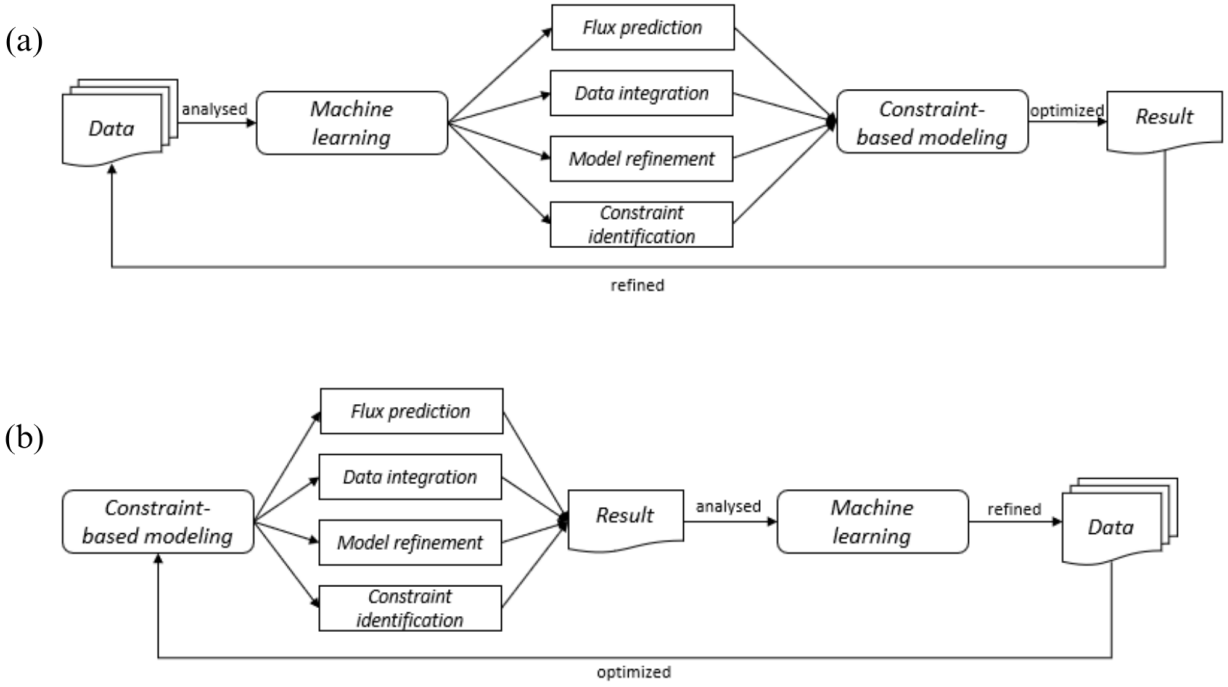

Fig. 6. Integration of machine learning and constraint-based modeling. (a) Refers to the ML as input to CBM, while (b) is CBM as input to ML).

include growth rate, product rate of desired metabolites, and different mutants with different combinations of knockout reactions.

According to [24], [49], the merging of ML and CBM can occur in three approaches. The first approach involves the inclusion of ML after CBM generates fluxomic data by predicting the growth conditions, cellular ML productivity, nutrient consumption, gene essentiality, or biomass concentration. The second approach uses a multi-omics data simplification process before entering the CBM process. The results of fluxomic data from CBM are then combined with the initial multi-omics data for the prediction process using a specific ML algorithm. The last approach uses ML on multi-omics data to get fluxomic data. This paper deduces that merging CBM with ML can occur in two ways, namely, ML as input to the CBM and CBM as input to the ML.

In the prior case, machine learning methods can improve metabolic models' accuracy and predictive power by predicting and refining metabolic models. The metabolic fluxes from omics data predicted using ML algorithms are input constraints for the metabolic model. Furthermore, ML can assist in identifying essential features (genes or reactions) for improving specific metabolite production. Considering that the metabolic model is complex, identifying crucial genes or reactions is essential while maintaining the viability of a cell. Meanwhile, in the post case, CBM can provide features, labels, and model selection to machine learning. Constraint-based approaches have been used to model the GSMM for simulating the phenotypic behavior after genotype perturbations. With the inclusion of machine learning methods, the selected features

from CBM can be used to train ML models for predicting the pathway activity, thus optimizing the metabolic model. Fig. 6 illustrates the integration of machine learning and constraint-based modeling approaches.

V. Application of Machine Learning in Metabolic Engineering

An unprecedented amount of information has now been used to seek biological mechanisms at the molecular level. The recent advancement of high-throughput technologies has significantly boosted data collecting and fundamentally altered how people view molecular biology [50]. However, predicting bioproduction titers from microbial hosts has been challenging due to complicated interactions between regulatory networks, signaling, and metabolic networks [50]. There are several ways to carry out experiments concerning metabolic engineering. Machine learning, which has undoubtedly led to significant improvements in recent research and is expected to surge shortly, is a critical tool for analyzing, understanding, and exploiting omic data.

A novel approach for predicting yeast metabolome using machine learning based on quantitative proteomic data of kinase knockouts was presented by [51]. The results showed that the ML algorithm accurately predicts the metabolome with complex genetic modification. However, the study assumes that protein expression levels are proportional to changes in metabolic flux. Nevertheless, when post-transcriptional or post-translational modifications occur, the protein expression levels may differ and not proportionate to the changes in metabolic flux. Additionally, the dataset used is relatively small. Thus, expanding the dataset to include a broader range of genetic perturbations and experimental conditions could improve the generalizability of the ML models.

In another research, the integration of knowledge mining, genome-scale modeling, and ML for predicting the bioproduction of *Yarrowia lipolytica* has been proposed [50]. The proposed framework integrates different data, including genomics, metabolomics, and literature, to construct a knowledge-based and optimal GSMM. Then, ML algorithms are applied to predict bioproduction yields based on gene expression data and environmental conditions. They have successfully outperformed the traditional methods. However, the complexity of GSMM and lack of comprehensive knowledge may hinder accurate predictions. Thus, further development and validation are crucial to enhance its applicability and reliability.

Furthermore, [52] have proposed multi-omics data to analyze and characterize key molecular pathways and features essential for yeast growth based on different environmental conditions. The pipeline incorporates biological knowledge in the machine learning model to improve predictions. The proposed pipeline outperforms traditional ML methods and gives insight into the underlying biological mechanisms regulating cell growth. However, the pipeline has several limitations that need to be addressed. For instance, the pipeline relies on the quality and completeness of data sources, which may vary and be limited across different organisms.

A machine learning framework to assess microbial factories' performance was proposed by [1], which thos microbial are microorganisms that can produce various valuable compounds. Like [50], [52], the researchers proposed the integration of different data, including genomics, transcriptomics, metabolomics, and fermentation data. This integration framework is used to model the relationship between genetic and environmental factors and the production of target compounds. The proposed framework uses feature selection, regression, and classification algorithms to predict yields, identify genetic targets for strain engineering, and optimize the conditions. Although the proposed framework successfully demonstrated promising results, however, the framework relies on the availability of data sources. Furthermore, the complexity of metabolic networks and the lack of kinetic transcriptional or genomics data may affect the accuracy of prediction and strain engineering.

In addition, Tachibana and his colleagues prepared a study on Green Fluorescent Protein (GFP) extracted from engineered Escherichia coli. They conducted using Deep Neural Network (DNN) [53]. Before being assessed by machine learning to assign the GFP intensities into a reasonable range for analysis with the DNN technique, the GFP intensities were scaled down by five orders of magnitude. All machine learning methods utilized data from the yeast extract for double-validation calculations. The remaining data were divided into learning and test datasets for random cross-validation. DNNs were built using tanh activations and four hidden layers (200, 100, 50, and 20 units). The average Mean Squared Error (MSE), determined from the rearranged matrices for each variable, was used to measure representative importance in their study. Their research discovered that DNN showed high coefficients of determination and low MSE values.

Different ML algorithms, including random forest, support vector machine, and neural networks have been evaluated by [54], to assess their accuracy in predicting the phenotypic traits of three organisms: yeast, rice, and wheat. The study also investigates the impact of different feature selection methods and data preprocessing techniques on predictive performance. Based on the research, the authors found that combinations of ML algorithms and feature selection methods can achieve high accuracy in predicting phenotypic traits based on genetic data. In another domain, elastic net logistic regression has been proposed to determine the functional and structural brain alterations in female schizophrenia patients [55]. The study combines functional magnetic resonance imaging and diffusion tensor imaging to identify brain regions associated with the disease. The elastic net logistic regression selects relevant features and builds a predictive model. The study found that the model improves the accuracy of classifying the patients.

The developed framework or pipeline proposed by previous researchers demonstrates that machine learning can achieve high accuracy in predicting phenotypic traits based on genotypic perturbations. Moreover, multi-omics data integration has allowed ML algorithms to improve the accuracy of strain engineering in selecting the optimal genetic perturbations. However, there are some limitations and challenges that need to be addressed. Firstly, transcriptomic and genetic data availability is only limited to specific organisms. Thus, predicting and simulating genetic perturbations for less researched organisms is challenging. In addition, the complexity of metabolic networks, thus the complexity of integrated networks, may hinder the predictive capabilities and strain engineering. Therefore,

further development and validation, including biological validation, is needed to enhance ML's interpretability, robustness, and applicability in predicting phenotype changes.

Shimizu and Toya in 2021 experimented with evaluating the cellular performances of 13C- metabolic flux analysis using artificial gene deletion [56]. It is essential to understand the physiology of the metabolism in practical bioprocesses to evaluate the efficiency of the desired model. They stated that the quantitative imaging of microbial cells for metabolic engineering is enabled by metabolic flux analysis (MFA). The nonlinear least squares approach is used to compute metabolic fluxes. A mathematical model that includes carbon atom transfers and molecular mass balancing is provided. Based on the solution space, it is possible to calculate the best trajectory for a given growth and output rate. For the growth phase, the individual growth rate is kept at its highest level and shifted to the critical value, which produces the highest specific production rate.

A. Combination of Unsupervised and Supervised Techniques

Moreover, many articles have reviewed the recent advances in model-assisted metabolic engineering, which aims to design and optimize the metabolic pathways of organisms to improve the production of desired metabolites [39], [57], [58]. Mainly, the review articles discussed the use of ML to assist in predicting the effects of genetic perturbations for integrated multi-omics data. Previously, metaheuristics optimization algorithms, such as Genetic Algorithm (GA), Differential Search Algorithm (DSA), flower pollination algorithm, Bee Algorithm, Particle Swarm Optimization (PSO), and others, have been used to improve the design of strain. The improved production of desired metabolites has proved the success of metaheuristic algorithms. However, with multi-omics data integration, the strain design becomes more challenging. Thus, using ML approaches is highly needed to enhance the accuracy of model predictions.

B. Unsupervised Techniques

Unsupervised techniques identify patterns based on predetermined mathematical criteria (such as the number of clusters or variance independence). Large-scale biological datasets have been analyzed using both learning techniques, which have also been combined with FBA. For the unsupervised technique, Sahu et al. developed the "Split Lipids into Measurable Entities reactions" (SLIMEr) approach to model the lipids in genome-scale metabolic models in yeast [59]. SLIMEr later divides lipid components into acyl chain distributions and lipid classes using a mathematical framework, imposing limitations on both [59]. Subsequently, Sahu and his colleagues also established a framework to examine growth-related mechanisms of several S. cerevisiae strains by combining FBA with Multimodal Artificial Neural Networks [59]. The study was to use mechanistic knowledge to integrate data-driven ML techniques to overcome their "black-box" restriction in flux distributions. The framework was evaluated using 1,484 strains of *S. cerevisiae* with single gene knockouts. Growth rates were designated as constraints in pFBA. The study shows that Multimodal Artificial Neural Networks and FBA can train and predict the individual gene expression data for analyzing the flux distributions.

Jalili et al. (2021) performed cancer-specific metabolic signatures using Random forest classification with PCA and FBA [60]. For each cancer model, flux distributions were computed using FBA. After that, using PCA and Random Forests techniques, FBA-based characteristics were generated. PCA generates the variation of flux distributions in cancer models representing the response variables. Random Forests then employed these response variables to categorize crucial fluxes (which showed the impacted sub-cellular systems). Based on their study, the authors discovered that the pentose phosphate route, extracellular transport, mitochondrial transporters, fatty acid production, and other metabolic characteristics are the factors that distinguish between normal and abnormal cell metabolisms for the cancer model.

Meanwhile, unsupervised ML mainly creates clusterings or representations of the unlabeled dataset to reduce the dimensional complexity of data. Principal Component Analysis (PCA) and K-means clustering are examples of unsupervised ML. In ME, unsupervised ML techniques can be implemented to identify the appropriate and non-appropriate reactions involved in producing desired metabolites. Moreover, unsupervised clustering techniques have been used to distinguish different cell types, such as healthy and non-healthy, cancer and non-cancer markers, and stressed and non-stressed. Fig. 7 below illustrates the unsupervised methods in ME.

In another study, Barbosa and the team researched the effects of production factors such as sugar, nitrogen level, and fermentation temperature on wine quality in non-*Saccharomyces* yeasts [61]. The Exploratory Data Analysis (EDA) activity was enhanced by employing unsupervised machine learning on the entire experimental data set. Latent variable techniques, such as Principal Component Analysis, were used to investigate the responses of multivariate structure. Using agglomerative hierarchical clustering (AHC), 18 responses of natural groups were found. Consequently, the forward stepwise variable selection method is used to determine the input variables for the regression model. The study successfully found direct patterns between different production factors, signifying positive and negative correlations.

They stated that the correlation distance was used to identify clusters or groups of functionally related fermentation metabolites [61]. It was anticipated that the first principal component for the cluster-specific PCA models would explain the majority of the overall variability in the cluster due to highly correlated variables generating clusters. Upon completing PCA, supervised ML was also applied. They used a forward stepwise variable selection method to determine which input variables (experimental factors and their higher-order terms) should be included in the regression model. The stepwise selection technique involved picking and incorporating components one at a time. When there are no variables whose inclusion or exclusion from the model would result in a change in the model's explanatory power that is statistically significant, the method finally ends.

For unsupervised ML, they found that clear patterns of linked variables can be seen in the loading plots, such as those that cluster together or lie in the other direction, signifying positive and negative correlations, respectively, as in Fig. 8. This exploratory PCA analysis supports the necessity to investigate the modular structure of the answers in more detail

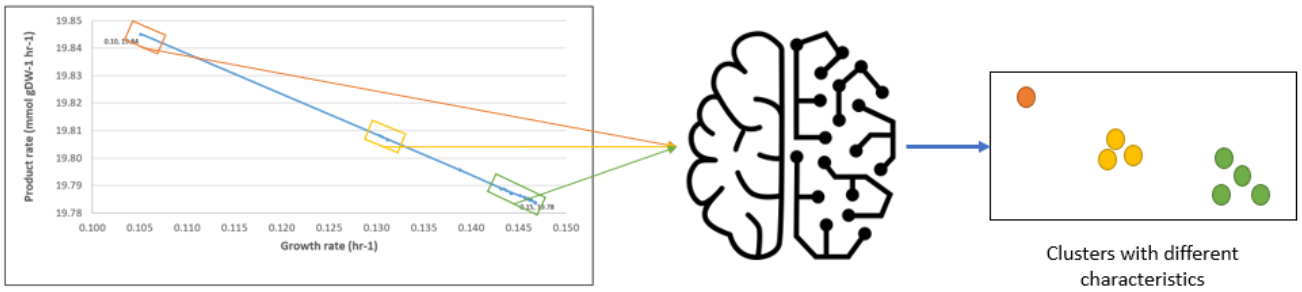

Fig. 7. Unsupervised ML in ME.

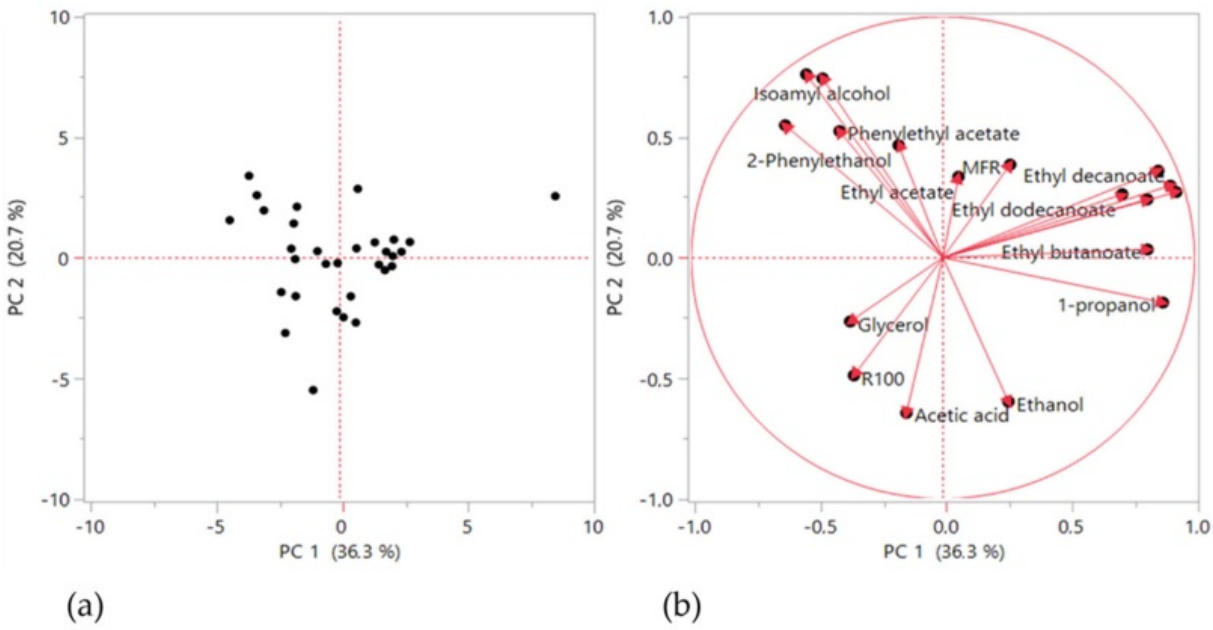

Fig. 8. Result using PCA. a) Scores plot, b) Loading plot [61].

and to identify the linked responses' natural building blocks. In order to uncover the natural blocks or clusters of variables, they implemented an agglomerative hierarchical variable-clustering approach (AHC). As a result, the variables are closer to each other, thus reducing the agglomeration distances [61]. For supervised ML, the considerable changes seen in the exploratory analysis section are confirmed by the modeling results utilizing main effects, second-order interactions, and quadratic terms, indicating the critical influence of the parameters on the fermentation process.

VI. Result and Discussion

The first activity of this review is collecting the references. We first searched the Scopus and Web of Science databases with the keywords "Machine learning", "Flux balance analysis", and "Metabolic engineering" to find relevant literature in recent years. Then, we filtered for references related to the integration between ML and CBM from the results obtained.

After searching the keywords "Machine learning", "Flux balance analysis", and "Metabolic engineering", 223 research studies were extracted through automatic search from Scopus and Web of Science databases. Of the majority of these 223 studies, 32 were duplicate studies and review papers and thus were eliminated from the list. Based on the title, abstracts, and keywords, the remaining 191 research studies were examined, and 90 studies were excluded. Next, the remaining 101 studies were further selected, in which papers published from 2020 to 2023 were selected and left with 13 studies.

Then, the selected relevant references are synthesized. Table III illustrates the synthesis results of 13 relevant studies in the sources. In the table, 17 machine learning approaches are integrated into constraint-based modeling, namely, binary classifier, random forest, PCA, SVM, KNN, Decision tree, gradient tree boosting, DNN, CNN, t-SNE, ensemble learning, kMeans, lasso, multiview neural network, regularized logistic regression, ANN, and reinforcement learning. Out of those 13 studies, only two use the kinetic model, whereas most use the stoichiometric model. Since the stoichiometric model does

TABLE III. SUMMARY OF SELECTED RELEVANT STUDIES

ID	Title	Year	Model	CBM	ML	Purpose	Strategy	Ref.
S1	Prediction of gene essentiality using machine learning and genome-scale metabolic models	2022	S	FBA	Binary classifier	Improves the identification of the essential genes	1	[31]
S2	Identifying metabolic shifts in Crohn's disease using' omics-driven contextualized computational metabolic network models	2023	S	FBA	Random forest	Identify biomarkers for Crohn's disease	1	[62]
S3	Genome-scale modeling of Chinese hamster ovary cells by hybrid semi-parametric flux balance analysis	2022	S	FBA	PCA	Integrate parametric and non-parametric constraints for reducing the search space and improve the prediction of FBA	2	[29]
S4	Computational Framework for Machine-Learning-Enabled C-13 Fluxomics	2022	S	MFA	SVM, KNN, decision tree, random forest, gradient tree boosting, DNN	Predict the flux ratio based on solvability and feature screening	2	[63]
S5	Machine learning-guided evaluation of extraction and simulation methods for cancer patient-specific metabolic models	2022	S	FBA	CNN, t-SNE	Identify the biological features based on cancer patient-specific GEMs	1	[64]
S6	Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction	2021	S	FBA	Ensemble learning	Reconstruct Yarrowia lipolytica GSM to improve organic acids' productions.	2	[50]
S7	Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance	2021	S	FBA	Ensemble learning	Identify biomarkers that are associated with radiation resistance	1	[65]
S8	In silico Design for Systems-Based Metabolic Engineering for the Bioconversion of Valuable Compounds From Industrial By-Products	2021	S	FBA	Random forest	Improve the production of glycerol by integrating transcriptomics data with metabolic network	1	[39]
S9	A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria	2020	S	rFBA	PCA, kMeans, Lasso	Identify the key cross-omics features	1	[66]
S10	A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth	2020	S	pFBA	Multiview neural network	Improve the prediction of phenotypic traits of interest.	1	[52]
S11	A biochemically-interpretable machine learning classifier for microbial GWAS	2020	S	popFVA	PCA, regularized logistic regression	Estimate the functional effects of genetic-associated alleles	1	[67]
S12	A Machine Learning Approach for Efficient Selection of Enzyme Concentrations and Its Application for Flux Optimization	2020	K	FBA	PCA, ANN	Select the optimized enzyme concentration for optimal yield	1	[68]
S13	Strain design optimization using reinforcement learning	2022	K	FBA	Reinforcement learning	Improve production of L-tryptophan	1	[69]

Note : S represents the stoichiometric model; K represents the kinetic model; 1 represents CBM as input to ML; 2 represents ML as input to CBM.

not require intracellular experimental parameters, which are hardly known, stoichiometric models are more favorable for biologists to exploit the detailed capabilities of cell metabolism and [70] outperform kinetic model when the dataset used has large networks [71]. Though kinetic models provide detailed quantitative descriptions of the processes involved in the systems, thus revealing a system's actual dynamic biological behavior, the kinetic model is only limited to the small-scale and newly curated metabolic network [25].

Meanwhile, flux balance analysis (FBA) is the most widely used model assessment method because FBA uses linear programming that is easier to apply than MoMA and ROOM, which use quadratic programming and mixed-integer linear programming. Moreover, although the solutions provided by FBA are non-unique as it does not consider regulatory and signal data, the existing metabolic networks are still incomplete [23]. Regardless of these imperfections, FBA can determine the steady-state fluxes of organisms and predict the optimal long-term evolved state of the cells. In contrast, MoMA and ROOM predict the immediate initial outcome of genetic manipulations. However, cells will evolve from a minimized flux distribution state to an FBA solution [4]. In other words, genetic manipulations will first lead to flux distribution predicted by MoMA and ROOM, eventually converging to a solution predicted by FBA.

TABLE IV. SUMMARY OF MODEL, ADVANTAGE, AND DISADVANTAGES OF MACHINE LEARNING BASED ON THE RELEVANT STUDIES

ID	Dataset	Result	Disadvantage
S1	GEMs of Escherichia coli	The proposed approach showed that will-type FBA solutions contain enough information to predict essentiality, without perturbation such as reaction or gene knockout.	There is no a standar strategy on machine learning utilized for essentiality prediction generally.
S2	RISK cohort data, gene expression data for all mucosal terminal ileal biopsies.	A framework that is a potential to identify pathways of clinical relevance in Crohn's disease, discover of novel diagnostic biomarkers, and therapeutic targets.	There is the discrepancy in the generated metabolic models of Crohn's disease in both RISK-derived tissue and enteroids.
S3	GEMs of Chinese hamster ovary (CHO) cells	The proposed hybrid FBA by involving the mechanistic and non-parametric constraints can efficiently reduce the solution space and improve the prediction result of FBA.	Need the experimental fluxes datasets with the guaranteed high accuracy.
S4	13 C fluxomics	The proposed approach is reliable for fluxomics method readily and applicable to high-throughput metabolic phenotyping.	Computationally expensive especially in the large-scale metabolic network.
S5	Cancer patient-specific GEMs	The results show that tINIT and GIMME has the high performance, but FBA and pFBA has poor performance in cancer metabolism.	Computationally expensive especially in the large GEMs.
S6	GEMs of Yarrowia Lipolytica	This study succeed in integrating knowledge mining, feature extraction, GEMs, and ML for predicting chemical titers in Yarrowia lipolytica.	This model can not capture biosynthesis bottlenecks, consequently, the predictability for low-performance strains is not optimal.
S7	Transcriptomic and genomic datasets	GEMs from patient tumors generated from transcriptomic and genomic datasets. The proposed approach, namely integrating ML and the generated GEMs, can identify prognostic metabolite biomarkers and predict radiosensitivity for individual patients.	Need to collect a larger datasets with the guaranteed high quality.
S8	GEMs of Escherichia coli and transcriptomics data	The proposed method, namely the combination of transcriptome, GEMs, and machine learning can improve the production rate of glycerol.	It does not involve other parameters that influence metabolic processes, such as enzyme, transcriptional regulation, and signaling.
S9	GEMs of Synechococcus sp. PCC 7002, transcriptomics	The proposed approach, namely model-generated flux data, are potential for predicting the growth rates.	Depends on Important information such as the specific metabolite uptake constraints and the nutrient uptake rates that are difficult to obtain directly.

Continued on the next page

TABLE IV. SUMMARY OF MODEL, ADVANTAGE, AND DISADVANTAGES OF MACHINE LEARNING BASED ON THE RELEVANT STUDIES–CONTINUED

ID	Dataset	Result	Disadvantage
S10	Model of Saccharomyces cerevisiae	The proposed framework, namely, a multimodal learning framework, is capable for undestanding and manipulating complex phenotypes and increasing the prediction accuracy.	It does not involve other parameters that influence metabolic processes, such as enzyme, transcriptional regulation, and signaling.
S11	Dataset of drug-tested Mycrobacterium tuberculosis strains	The proposed approach, namely metabolic allele classifier (MACs), can predict antimicrobial resistance (AMR) phenotypes with accuracy on par with mechanism-agnostic ML.	Not suitable for microbial genome-wide association studies.
S12	The input data of 121 balances of four enzymes in the upper part of glycolysis	The ANN algorithms usedto select the enzyme concentration for the upper part of glycolysis	The ANN algorithms that was used to select the enzyme concentration for the upper part of glycolysis could select the optimum enzyme concentrations, improve flux up to 63%, and decrease a cost up to 25%.
S13	GEMs of Escherichia coli, k-ecoli457 and Saccharomyces cerevisiae	The proposed method, namely MARL, could optimize the L-tryptophan production in S. cerevisiae and specific metabolite in the k-ecoli457. MARL could also be used to optimize metabolic gene expression levels.	Its application is still restricted to the particular target enzymes.

As for integrating machine learning with constraint-based models, most those relevant studies employed the first strategy in which biological insights from CBM are used as input to ML. Given the intricacy of biological data and certain biological phenomena or systems that cannot be comprehensively described and examined mechanistically. In the table, there are 10 studies utilized the first strategy to integrate ML into CBM. The task of ML in those studies are to identify, improve, estimate, and select. In identifying, ML have been applied to identify the essential genes [31], biomarker [65], [62], the biological features [64], and the key cross-omics features [66]. Then, the application of ML in the improving process are to improve the production of glycerol [39], the prediction of phenotypic [69], and the production of L-tryptophan [69]. At the rest, ML was applied in estimating the functional effect of genetic-associated aleles and selecting the optimized enzyme concentration for optimal yield.

Nevertheless, some research studies employ a second strategy in which ML analyzes multi-omics data for CBM model reconstruction. In this strategy, ML have useful in the reducing, predicting, and reconstructing processes. In reducing, PCA has been implemented by integrating parametric and non-parametric constraints for reducing the search space in order to improve the prediction of FBA [29]. Then, several machine learning approaches have been utilized in the predicting process to get the optimal flux ratio based on solvability and feature screening [63]. Meanwhile, for reconstructing, Ensemble learning has been applied to reconstruct GSMMs of Yarrowia lipolytica in order to improve organic acids' productions [50], where the reconstruction of GSMM involves multiple steps, including annotation, gap filling, and refinement.

Table IV provides results, dataset used, and disadvantages of the relevant studies from Table III. Based on the synthesis and analysis results obtained from the relevant studies, there are several potentials of ML to contribute in *in silico* metabolic engineering. Integrating ML in traditional algorithms, such as flux balance analysis, can improve the production of the desired metabolites and even promise to guide strain optimization based on hybrid models, namely, the mechanistic and data-driven models. Moreover, ML has given positive influences on the prediction results by involving several experimental data such as fluxomic, transcriptomic, metabolomic, and proteomic in the process of constraint-based modeling. Also, it has been shown that ML can construct GSM, predict the essential genes, reduce the dimensionality of cross-omics features, and study the pattern of omic data. Based on those potentials, ML needs to be considered in metabolic engineering processes using CBM.

VII. CONCLUSION

The advancements in biology, bioinformatics, and computational tools have led to the development of efficient software for modifying organisms for industrial use. Furthermore, the successful reconstruction models of complex biological systems by integrating data from various molecular levels have yielded valuable insights into organisms, thus offering accurate insights into cell activities during organism perturbations. However, this integration can complicate the identification of near-optimal reaction knockouts due to complex biological networks. Therefore, machine learning (ML) and constraint-based modeling (CBM) are employed to facilitate and enhance prediction accuracy.

This review introduced different structure models for representing organisms' systems. Due to the traditional approaches that are costly and irreversible, constraint-based methods have been introduced to overfit the production of valuable metabolites. Though it provides near-optimal solutions, integrating

diverse omics data holds substantial promise in predicting the future state of computational biology systems. Over the coming decade, there will be a growing need for machine learning methods that can be effectively utilized and tailored for these large datasets. Therefore, machine learning methods were integrated into CBM methods to improve the reconstruction of GSMM and the prediction accuracy of genetic perturbations.

We also reviewed several algorithms and applications developed and their different strategies and approaches used in metabolic engineering. As mentioned before, the integration of ML and CBM can happen in two ways. The first way is to apply ML to the integrated biological networks in which ML will identify the essential and meaningful features using the classification technique (supervised ML). This step minimizes the solution space and reconstructs a reduced integrated network for modeling in CBM. The second way is to analyze simulation modeling results from CBM (unsupervised ML).

In conclusion, ML is a superior technique for identifying meaningful features and patterns, which can help reconstruct integrated biological networks that represent the true nature of a cell, thus improving the predictive capabilities of identifying near-optimal reactions knockout for optimizing the production rate of valuable metabolites and growth rates of mutants for industrial purposes.

ACKNOWLEDGMENT

This research was funded by the Fundamental Research Grant Scheme - from the Ministry of Education Malaysia (FRGS/1/2021/ICT02/UKM/02/2).

REFERENCES

[1] T. Oyetunde, D. Liu, H. G. Martin, and Y. J. Tang, "Machine learning framework for assessment of microbial factory performance," *PLOS ONE*, vol. 14, no. 1, pp. 1–15, 01 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0210558

[2] H. Fang, D. Li, J. Kang, P. Jiang, J. Sun, and D. Zhang, "Metabolic engineering of escherichia coli for de novo biosynthesis of vitamin b_{12}," *Nature Communications*, vol. 9, no. 4917, 11 2018.

[3] Y. Xu, X. Wang, C. Zhang, X. Zhou, X. Xu, L. Han, X. Lv, Y. Liu, S. Liu, J. Li, G. Du, J. Chen, R. Ledesma-Amaro, and L. Liu, "De novo biosynthesis of rubusoside and rebaudiosides in engineered yeasts," *Nature Communications*, vol. 13, no. 3040, 2022.

[4] P. Rana, C. Berry, P. Ghosh, and S. S. Fong, "Recent advances on constraint-based models by integrating machine learning," *Current Opinion in Biotechnology*, vol. 64, pp. 85–91, 2020, analytical Biotechnology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095816691930117X

[5] C. Shene, P. Paredes, L. Flores, A. Leyton, J. A. Asenjo, and Y. Chisti, "Dynamic flux balance analysis of biomass and lipid production by antarctic thraustochytrid oblongichytrium sp. rt2316-13," *Biotechnology and Bioengineering*, vol. 117, no. 10, pp. 3006–3017, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.27463

[6] K. V. Presnell and H. S. Alper, "Systems metabolic engineering meets machine learning: A new era for data-driven metabolic engineering," *Biotechnology Journal*, vol. 14, no. 9, p. 1800416, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.201800416

[7] S. Mutturi, "Focus: a metaheuristic algorithm for computing knockouts from genome-scale models for strain optimization," *Mol. BioSyst.*, vol. 13, pp. 1355–1363, 2017. [Online]. Available: http://dx.doi.org/10.1039/C7MB00204A

[8] K. M. Daud, M. S. Mohamad, Z. Zakaria, R. Hassan, Z. A. Shah, S. Deris, Z. Ibrahim, S. Napis, and R. O. Sinnot, "A non-dominated sorting differential search algorithm flux balance analysis (ndsdsafba) for in silico multiobjective optimization in identifying reactions knockout," *Computers in Biology and Medicine*, vol. 113, pp. 1–13, 2019. [Online]. Available: 10.1016/j.compbiomed.2019.103390

[9] E. Iranmanesh, M. A. Asadollahi, and D. Biria, "Improving l-phenylacetylcarbinol production in saccharomyces cerevisiae by in silico aided metabolic engineering," *Journal of Biotechnology*, vol. 308, pp. 27–34, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168165619309186

[10] K. Shabestary and E. P. Hudson, "Computational metabolic engineering strategies for growth-coupled biofuel production by synechocystis," *Metabolic Engineering Communications*, vol. 3, pp. 216–226, 2016. [Online]. Available: doi.org/10.1016/j.meteno.2016.07.003

[11] H. A. BRAHIM, M. BENLLARCH, N. BENHIMA, S. EL-HADAJ, A. METRANE, and G. BELBARAKA, "New real dataset creation to develop an intelligent system for predicting chemotherapy protocols," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140886

[12] B. A. Al-Hameli, A. A. Alsewari, S. S. Basurra, J. Bhogal, and M. A. H. Ali, "Diabetes disease prediction system using hnb classifier based on discretization method," *Journal of Integrative Bioinformatics*, vol. 20, no. 1, 3 2023.

[13] N. Casano, S. J. Santini, P. Vittorini, G. Sinatti, P. Carducci, C. M. Mastroianni, M. R. Ciardi, P. Pasculli, E. Petrucci, F. Marinangeli, and C. Balsano, "Application of machine learning approach in emergency department to support clinical decision making for sars-cov-2 infected patients," *Journal of Integrative Bioinformatics*, vol. 20, no. 2, p. 20220047, 2023. [Online]. Available: https://doi.org/10.1515/jib-2022-0047

[14] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr)," *IEEE Access*, vol. 8, pp. 142 642–142 668, 2020.

[15] H. Hairani and D. Priyanto, "A new approach of hybrid sampling smote and enn to the accuracy of machine learning methods on unbalanced diabetes disease data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140864

[16] A. V. Colarusso, I. Goodchild-Michelman, M. Rayle, and A. R. Zomorrodi, "Computational modeling of metabolism in microbial communities on a genome-scale," *Current Opinion in Systems Biology*, vol. 26, pp. 46–57, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452310021000123

[17] J. Y. Lee, B. Nguyen, C. Orosco, and M. P. Styczynski, "Scour: a stepwise machine learning framework for predicting metabolite-dependent regulatory interactions," *BMC Bioinformatics*, vol. 22, no. 365, 7 2021.

[18] M. R. Antoniewicz, "A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications," *Metabolic Engineering*, vol. 63, pp. 2–12, 2021, tools and Strategies of Metabolic Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1096717620301683

[19] P. Kumar, P. A. Adamczyk, X. Zhang, R. B. Andrade, P. A. Romero, P. Ramanathan, and J. L. Reed, "Active and machine learning-based approaches to rapidly enhance microbial chemical production," *Metabolic Engineering*, vol. 67, pp. 216–226, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1096717621001087

[20] M. Rocha, P. Maia, R. Mendes, J. P. Pinto, E. C. Ferreira, J. Nielsen, K. R. Patil, and I. Rocha, "Natural computation meta-heuristics for the in silico optimization of microbial strains," *BMC Bioinformatics*, vol. 9, no. 499, 11 2008.

[21] J. K. Khanijou, H. Kulyk, C. Bergès, L. W. Khoo, P. Ng, H. C. Yeo, M. Helmy, F. Bellvert, W. Chew, and K. Selvarajoo, "Metabolomics and modelling approaches for systems metabolic engineering," *Metabolic Engineering Communications*, vol. 15, p. e00209, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214030122000189

[22] T. B. Alter, L. M. Blank, and B. E. Ebert, "Genetic optimization algorithm for metabolic engineering revisited," *Metabolites*, vol. 8, no. 2, 2018. [Online]. Available: https://www.mdpi.com/2218-1989/8/2/33

[23] A. Passi, J. D. Tibocha-Bonilla, M. Kumar, D. Tec-Campos, K. Zengler, and C. Zuniga, "Genome-scale metabolic modeling enables in-depth understanding of big data," *Metabolites*, vol. 12, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2218-1989/12/1/14

[24] M. K. Khaleghi, I. S. P. Savizi, N. E. Lewis, and S. A. Shojaosadati, "Synergisms of machine learning and constraint-based modeling of metabolism for analysis and optimization of fermentation parameters," *Biotechnology Journal*, vol. 16, no. 11, p. 2100212, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.202100212

[25] K. M. Shreya Anand and P. Padmanabhan, "An insight to flux-balance analysis for biochemical networks," *Biotechnology and Genetic Engineering Reviews*, vol. 36, no. 1, pp. 32–55, 2020. [Online]. Available: https://doi.org/10.1080/02648725.2020.1847440

[26] N. D. Price, J. L. Reed, and B. O. Palsson, "Genome-scale models of microbial cells: evaluating the consequences of constraints," *Nature Reviews Microbiology*, vol. 2, pp. 886–897, 11 2004.

[27] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson, "An expanded genome-scale model of Escherichia coli K-12 (iJR904," *Genome Biology*, 2003.

[28] B. García-Jiménez, J. Torres-Bacete, and J. Nogales, "Metabolic modelling approaches for describing and engineering microbial communities," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 226–246, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037020305286

[29] J. R. C. Ramos, G. P. Oliveira, P. Dumas, and R. Oliveira, "Genome-scale modeling of chinese hamster ovary cells by hybrid semi-parametric flux balance analysis," *Bioprocess and Biosystems Engineering*, no. 45, pp. 1889–1904, 10 2022.

[30] S. Tsouka, M. Ataman, T. Hameri, L. Miskovic, and V. Hatzimanikatis, "Constraint-based metabolic control analysis for rational strain engineering," *Metabolic Engineering*, vol. 66, pp. 191–203, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1096717621000392

[31] L. J. Freischem, M. Barahona, and D. A. Oyarzún, "Prediction of gene essentiality using machine learning and genome-scale metabolic models," *IFAC-PapersOnLine*, vol. 55, no. 23, pp. 13–18, 2022, 9th IFAC Conference on Foundations of Systems Biology in Engineering FOSBE 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240589632300006X

[32] D. Kenefake, E. Armingol, N. E. Lewis, and E. N. Pistikopoulos, "An improved algorithm for flux variability analysis," *BMC Bioinformatics*, vol. 23, no. 550, 12 2022.

[33] M. S. Alzboon and M. S. Al-Batah, "Prostate cancer detection and analysis using advanced machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140843

[34] N. A. A. Hassan, R. A. A. A. A. Seoud, and D. A. Salem, "Open information extraction methodology for a new curated biomedical literature dataset," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140783

[35] Q. Yin, X. Ye, B. Huang, L. Qin, X. Ye, and J. Wang, "Stroke risk prediction: Comparing different sampling algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01406115

[36] B. Wingfield, S. Coleman, T. McGinnity, and A. Bjourson, "Robust microbial markers for non-invasive inflammatory bowel disease identification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 6, pp. 2078–2088, 2019.

[37] I. K. Nti, A. F. Adekoya, B. A. Weyori, and O. Nyarko-Boateng, "Applications of artificial intelligence in engineering and manufacturing: a systematic review," *Journal of Intelligent Manufacturing*, vol. 33, pp. 1581–1601, 4 2021.

[38] X. Lv, A. Hueso-Gil, X. Bi, Y. Wu, Y. Liu, L. Liu, and R. Ledesma-Amaro, "New synthetic biology tools for metabolic control," *Current Opinion in Biotechnology*, vol. 76, p. 102724, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0958166922000581

[39] A. E. Tafur Rangel, W. Ríos, D. Mejía, C. Ojeda, R. Carlson, J. M. Gómez Ramírez, and A. F. González Barrios, "In silico design for systems-based metabolic engineering for the bioconversion of valuable compounds from industrial by-products," *Frontiers in Genetics*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.633073

[40] J. Zhang, S. D. Petersen, T. Radivojevic, A. Ramirez, A. Perez-Manriquez, E. Abeliuk, B. J. Sanchez, Z. Costello, Y. Chen, M. J. Fero, H. G. Martin, J. Nielsen, J. D. Keasling, and M. K. Jensen, "Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism," *Nature Communications*, vol. 11, no. 4880, 9 2020.

[41] T. Islam, A. B. Akhi, F. Akter, M. N. Hasan, and M. A. Lata, "Prediction of breast cancer using traditional and ensemble technique: A machine learning approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140692

[42] R. Siddalingappa and S. Kanagaraj, "A novel ml approach for computing missing sift, provean, and mutassessor scores in tp53 mutation pathogenicity prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01406111

[43] A. Altaf, H. Mahdin, A. Mahmood, M. I. H. Ninggal, A. Altaf, and I. Javid, "Systematic review for phonocardiography classification based on machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140889

[44] S. Swetha, G. N. Srinivasan, and P. Dayananda, "A hybrid multiple indefinite kernel learning framework for disease classification from gene expression data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140690

[45] F. Islam, M. H. Rahman, Nurjahan, M. S. Hossain, and S. Ahmed, "A novel method for diagnosing alzheimer's disease from mri scans using the resnet50 feature extractor and the svm classifier," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01406131

[46] Y. Boutazart, H. Satori, A. R. A. M, M. Hamidi, and K. Satori, "Covid-19 dataset clustering based on k-means and em algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01403105

[47] Z. Wang, X. Peng, A. Xia, A. A. Shah, Y. Huang, X. Zhu, X. Zhu, and Q. Liao, "The role of machine learning to boost the bioenergy and biofuels conversion," *Bioresource Technology*, vol. 343, p. 126099, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960852421014413

[48] A. Antonakoudis, R. Barbosa, P. Kotidis, and C. Kontoravdi, "The era of big data: Genome-scale modelling meets machine learning," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 3287–3300, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037020304335

[49] M. K. Khaleghi, I. S. P. Savizi, N. E. Lewis, and S. A. Shojaosadati, "Synergisms of machine learning and constraint-based modeling of metabolism for analysis and optimization of fermentation parameters," *Biotechnology Journal*, vol. 16, no. 11, p. 2100212, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.202100212

[50] J. J. Czajka, T. Oyetunde, and Y. J. Tang, "Integrated knowledge mining, genome-scale modeling, and machine learning for predicting yarrowia lipolytica bioproduction," *Metabolic Engineering*, vol. 67, pp. 227–236, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1096717621001130

[51] A. Zelezniak, J. Vowinckel, F. Capuano, C. B. Messner, V. Demichev, N. Polowsky, M. Mülleder, S. Kamrad, B. Klaus, M. A. Keller, and M. Ralser, "Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts," *Cell Systems*, vol. 7, no. 3, pp. 269–283.e6, Sep. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405471218303168

[52] C. Culley, S. Vijayakumar, G. Zampieri, and C. Angione, "A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth," *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18 869–18 879, Aug. 2020. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.2002959117

[53] S. Tachibana, T.-Y. Chiou, and M. Konishi, "Machine learning modeling of the effects of media formulated with various yeast extracts on heterologous protein production in escherichia coli," *MicrobiologyOpen*, vol. 10, no. 3, p. e1214, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mbo3.1214

[54] N. F. Grinberg, O. I. Orhobor, and R. D. King, "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat," *Machine Learning*, vol. 109, no. 2, pp. 251–277, Feb. 2020. [Online]. Available: http://link.springer.com/10.1007/s10994-019-05848-5

[55] Y. Wu, P. Ren, R. Chen, H. Xu, J. Xu, L. Zeng, D. Wu, W. Jiang, N. Tang, and X. Liu, "Detection of functional and structural brain alterations in female schizophrenia using elastic net logistic regression," *Brain Imaging and Behavior*, vol. 16, no. 1, pp. 281–290, Feb. 2022. [Online]. Available: https://link.springer.com/10.1007/s11682-021-00501-z

[56] H. Shimizu and Y. Toya, "Recent advances in metabolic engineering–integration of in silico design and experimental analysis of metabolic pathways," *Journal of Bioscience and Bioengineering*, vol. 132, no. 5, pp. 429–436, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138917232100205X

[57] P. F. Suthers, C. J. Foster, D. Sarkar, L. Wang, and C. D. Maranas, "Recent advances in constraint and machine learning-based metabolic modeling by leveraging stoichiometric balances, thermodynamic feasibility and kinetic law formalisms," *Metabolic Engineering*, vol. 63, pp. 13–33, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S109671762030183X

[58] G. Zampieri, S. Vijayakumar, E. Yaneske, and C. Angione, "Machine and deep learning meet genome-scale metabolic modeling," *PLOS Computational Biology*, vol. 15, no. 7, pp. 1–24, 07 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1007084

[59] A. Sahu, M.-A. Blatke, J. J. Szymanski, and N. Topfer, "Advances in flux balance analysis by integrating machine learning and mechanism-based models," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4626–4640, 2021.

[60] M. Jalili, M. Scharm, O. Wolkenhauer, M. Damaghi, and A. Salehzadeh-Yazdi, "Exploring the metabolic heterogeneity of cancers: A benchmark study of context-specific models," *Journal of Personalized Medicine*, vol. 11, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2075-4426/11/6/496

[61] C. Barbosa, E. Ramalhosa, I. Vasconcelos, M. Reis, and A. Mendes-Ferreira, "Machine learning techniques disclose the combined effect of fermentation conditions on yeast mixed-culture dynamics and wine quality," *Microorganisms*, vol. 10, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2076-2607/10/1/107

[62] P. Fernandes, Y. Sharma, F. Zulqarnain, B. McGrew, A. Shrivastava, L. Ehsan, D. Payne, L. Dillard, D. Powers, I. Aldridge, J. Matthews, S. Kugathasan, F. M. Fernández, D. Gaul, J. A. Papin, and S. Syed, "Identifying metabolic shifts in Crohn's disease using 'omics-driven contextualized computational metabolic network models," *Scientific Reports*, vol. 13, no. 1, p. 203, Jan. 2023. [Online]. Available: https://www.nature.com/articles/s41598-022-26816-5

[63] C. Wu, J. Yu, M. Guarnieri, and W. Xiong, "Computational framework for machine-learning-enabled 13c fluxomics," *ACS Synthetic Biology*, vol. 11, no. 1, pp. 103–115, 2022, pMID: 34705423. [Online]. Available: https://doi.org/10.1021/acssynbio.1c00189

[64] S. M. Lee, G. Lee, and H. U. Kim, "Machine learning-guided evaluation of extraction and simulation methods for cancer patient-specific metabolic models," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 3041–3052, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037022002434

[65] J. E. Lewis and M. L. Kemp, "Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance," *Nature Communications*, vol. 12, no. 1, p. 2700, May 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-22989-1

[66] S. Vijayakumar, P. K. S. M. Rahman, and C. Angione, "A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria," *iScience*, vol. 23, no. 12, p. 101818, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004220310154

[67] E. S. Kavvas, L. Yang, J. M. Monk, D. Heckmann, and B. O. Palsson, "A biochemically-interpretable machine learning classifier for microbial GWAS," *Nature Communications*, vol. 11, no. 1, p. 2580, May 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-16310-9

[68] A. Ajjolli Nagaraja, P. Charton, X. F. Cadet, N. Fontaine, M. Delsaut, B. Wiltschi, A. Voit, B. Offmann, C. Damour, B. Grondin-Perez, and F. Cadet, "A machine learning approach for efficient selection of enzyme concentrations and its application for flux optimization," *Catalysts*, vol. 10, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2073-4344/10/3/291

[69] M. Sabzevari, S. Szedmak, M. Penttilä, P. Jouhten, and J. Rousu, "Strain design optimization using reinforcement learning," *PLOS Computational Biology*, vol. 18, no. 6, pp. 1–18, 06 2022. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1010177

[70] T. Hameri, G. Fengos, and V. Hatzimanikatis, "The effects of model complexity and size on metabolic flux distribution and control: case study in Escherichia coli," *BMC Bioinformatics*, vol. 22, no. 1, p. 134, Dec. 2021. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04066-y

[71] K. Tummler and E. Klipp, "The discrepancy between data for and expectations on metabolic models: How to match experiments and computational efforts to arrive at quantitative predictions?" *Current Opinion in Systems Biology*, vol. 8, pp. 1–6, 2018, • Regulatory and metabolic networks • Special Section: Single cell and noise. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452310017301920