Cell factory design with advanced metabolic modelling empowered by artificial intelligence

Citation for the original published paper (version of record):
Lu, H., Xiao, L., Liao, W. et al (2024). Cell factory design with advanced metabolic modelling empowered by artificial intelligence. Metabolic Engineering, 85: 61-72.
http://dx.doi.org/10.1016/j.ymben.2024.07.003

Metabolic Engineering 85 (2024) 61–72

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/meteng

Cell factory design with advanced metabolic modelling empowered by artificial intelligence

Hongzhong Lu [a,**], Luchi Xiao [a], Wenbin Liao [a,b], Xuefeng Yan [b], Jens Nielsen [c,d,*]

[a] State Key Laboratory of Microbial Metabolism, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, PR China
[b] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, 200237, PR China
[c] BioInnovation Institute, Ole Måløes Vej, DK2200, Copenhagen N, Denmark
[d] Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE412 96, Gothenburg, Sweden

ARTICLE INFO

Keywords:
In silico strain design
Mechanistic metabolic models
Artificial intelligence
Hybrid models

ABSTRACT

Advances in synthetic biology and artificial intelligence (AI) have provided new opportunities for modern biotechnology. High-performance cell factories, the backbone of industrial biotechnology, are ultimately responsible for determining whether a bio-based product succeeds or fails in the fierce competition with petroleum-based products. To date, one of the greatest challenges in synthetic biology is the creation of high-performance cell factories in a consistent and efficient manner. As so-called white-box models, numerous metabolic network models have been developed and used in computational strain design. Moreover, great progress has been made in AI-powered strain engineering in recent years. Both approaches have advantages and disadvantages. Therefore, the deep integration of AI with metabolic models is crucial for the construction of superior cell factories with higher titres, yields and production rates. The detailed applications of the latest advanced metabolic models and AI in computational strain design are summarized in this review. Additionally, approaches for the deep integration of AI and metabolic models are discussed. It is anticipated that advanced mechanistic metabolic models powered by AI will pave the way for the efficient construction of powerful industrial chassis strains in the coming years.

1. Introduction

Synthetic biology enables the development of various efficient cell factories capable of producing bio-based products for societal benefit (Volk et al., 2023). Technological breakthroughs in whole-genome synthesis and precise gene editing have made it feasible to reconstruct the metabolic networks of various cell factories and integrate heterogeneous reaction pathways from natural plants, or even construct artificial synthetic pathways that do not exist in nature (Wang and Doudna, 2023). With the aid of automation, gene manipulation of target strains can be carried out in a high-throughput manner (Si et al., 2017). However, cellular metabolism is highly complex, and a general chassis cell typically has 4000–6000 genes (Nielsen, 2017). Indeed, to improve the cell factory performance in terms of titre, production rate and yield (TRY), there are many possible solutions for overexpression, knockout and knockdown of a set of target genes. It is impossible to test all possible solutions during wet-lab experiments, and therefore it is valuable to determine the optimal engineering strategies for rewiring strain metabolism (King et al., 2015).

Various kinds of metabolic models have been developed and employed to predict the optimal design of strains to improve the TRY of target bioproducts. Among these models, genome-scale metabolic models (GEMs), encompassing detailed gene–protein-reaction relationships (GPRs), are widely used in systematic studies of strain metabolism. For industrial chassis strains, metabolic models can not only be used to predict metabolic gene targets for high productivity but also be valuable scaffolds for integrative analysis of omics data, which can help to reveal the mechanism underlying a certain phenotype, e.g., high productivity of a given product (McCloskey et al., 2013). As a result, GEMs have been built for more than 6239 organisms using manually curated or automatic procedures (Gu et al., 2019). Most GEMs only cover information on metabolic pathways, while constraint

* Corresponding author. BioInnovation Institute, Ole Måløes Vej, DK2200, Copenhagen N, Denmark.
** Corresponding author.
E-mail addresses: hongzhonglu@sjtu.edu.cn (H. Lu), JNI@bii.dk (J. Nielsen).

https://doi.org/10.1016/j.ymben.2024.07.003
Received 14 May 2024; Received in revised form 6 July 2024; Accepted 6 July 2024
Available online 20 July 2024

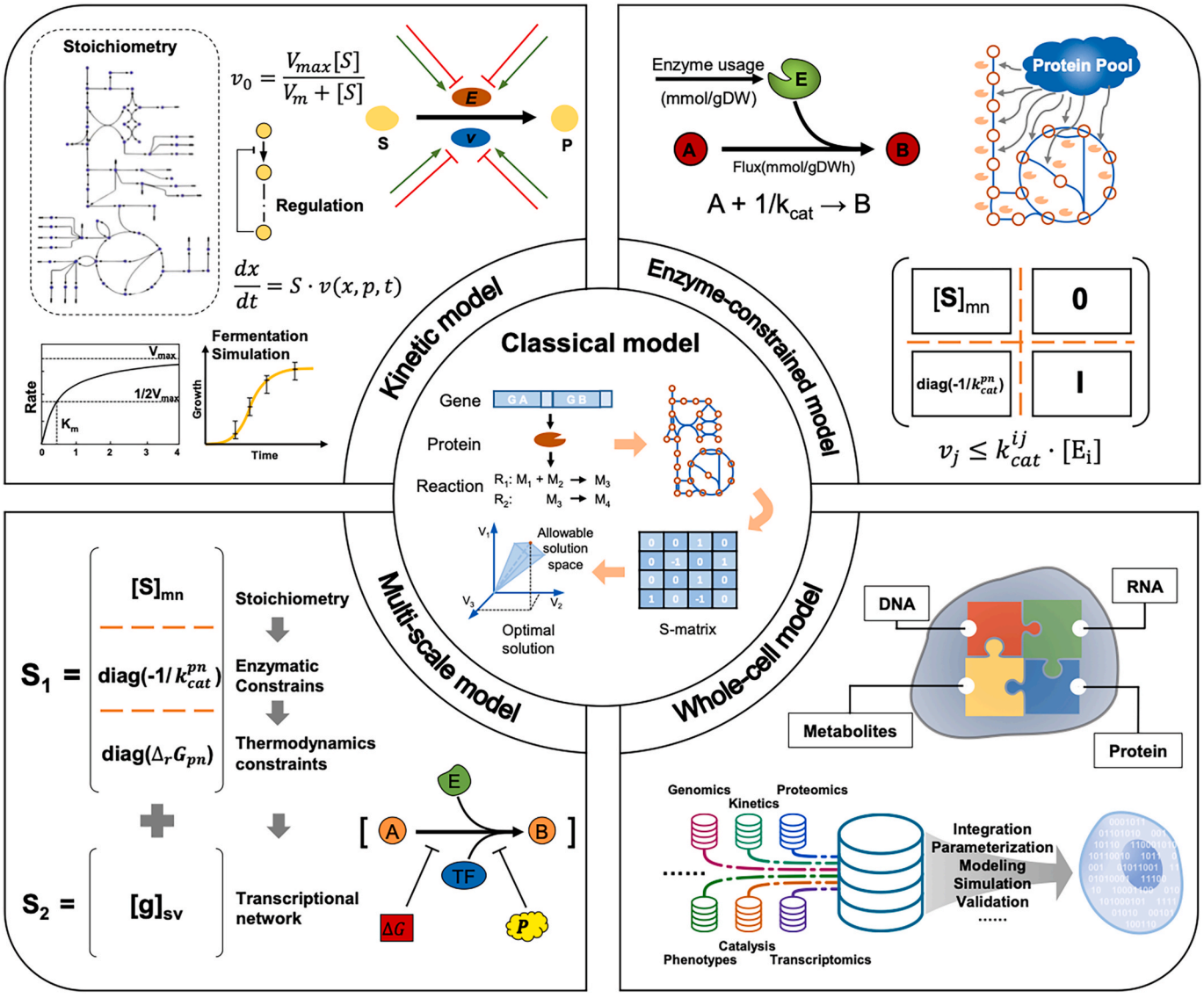

Fig. 1. Summary of several advanced metabolic models that are widely used in computational strain design. In the kinetic model, the Michaelis–Menten equation was used to describe metabolic dynamics over time. In the enzyme-constrained model, the enzyme kinetics and abundance were included to improve the model prediction capabilities. In the multi-scale model, multiple layers of the molecular network, such as the TRN, might be combined with the metabolic network to enhance the model prediction capabilities. Finally, a whole-cell model might encompass a range of metabolic submodules, such as those involved in DNA synthesis and the cell cycle, to represent whole-cell metabolic activities.

information at other levels of metabolism is always omitted. To overcome the bottlenecks of GEMs, more advanced metabolic models, including kinetic models, enzyme-constrained models, multi-scale models and whole-cell models, have been proposed to increase the number of application scenarios for metabolic models (Lu et al., 2021) (Fig. 1). With the larger size and higher prediction accuracy, the latest models display great potential in the rational design of industrial strains. However, in contrast to classical GEMs, the successful reconstruction of these more advanced models requires a certain number of parameters, i.e., the enzyme's k_{cat} and K_m, as well as a delicate tuning of these parameters within the models. In this regard, most industrial strains lack these key enzyme parameters, which hinders the wider application of these advanced models in the field of industrial biotechnology.

Complementary to those mechanistic metabolic models, owing to the tremendous advances in big data generation and AI technology, various machine learning models (**MLs**, also including deep learning models) are now widely used in diverse biological fields, such as enzyme optimization (Yu et al., 2023c), *de novo* pathway design (Zhang and Lapkin, 2023) and strain development (Sabzevari et al., 2022), providing new opportunities for the rational design of next-generation cell factories. Metabolic models are regarded as white-box models, while MLs can be regarded as black-box models because it is difficult to infer the possible mechanisms underlying MLs (Yang et al., 2019). Even having drawbacks in terms of interpretability, MLs still have enormous potential for predicting phenotypes from genotypes by learning hidden features from high-dimensional datasets (Greener et al., 2022), e.g., those obtained by using robots to perform large-scale molecular experiments (Si et al., 2017). Such datasets could be fed into MLs to iteratively improve the predictive performance of MLs. The application of MLs is emerging as an alternative approach to improve metabolic engineering (Patra et al., 2023). The weak interpretability of MLs can be partially addressed by combining the aforementioned mechanistic metabolic models with MLs to understand the features of cellular metabolism learned by MLs (Zampieri et al., 2019).

Currently, despite the enormous demand for the reconstruction of powerful chassis cells, rational strain design based on advanced model

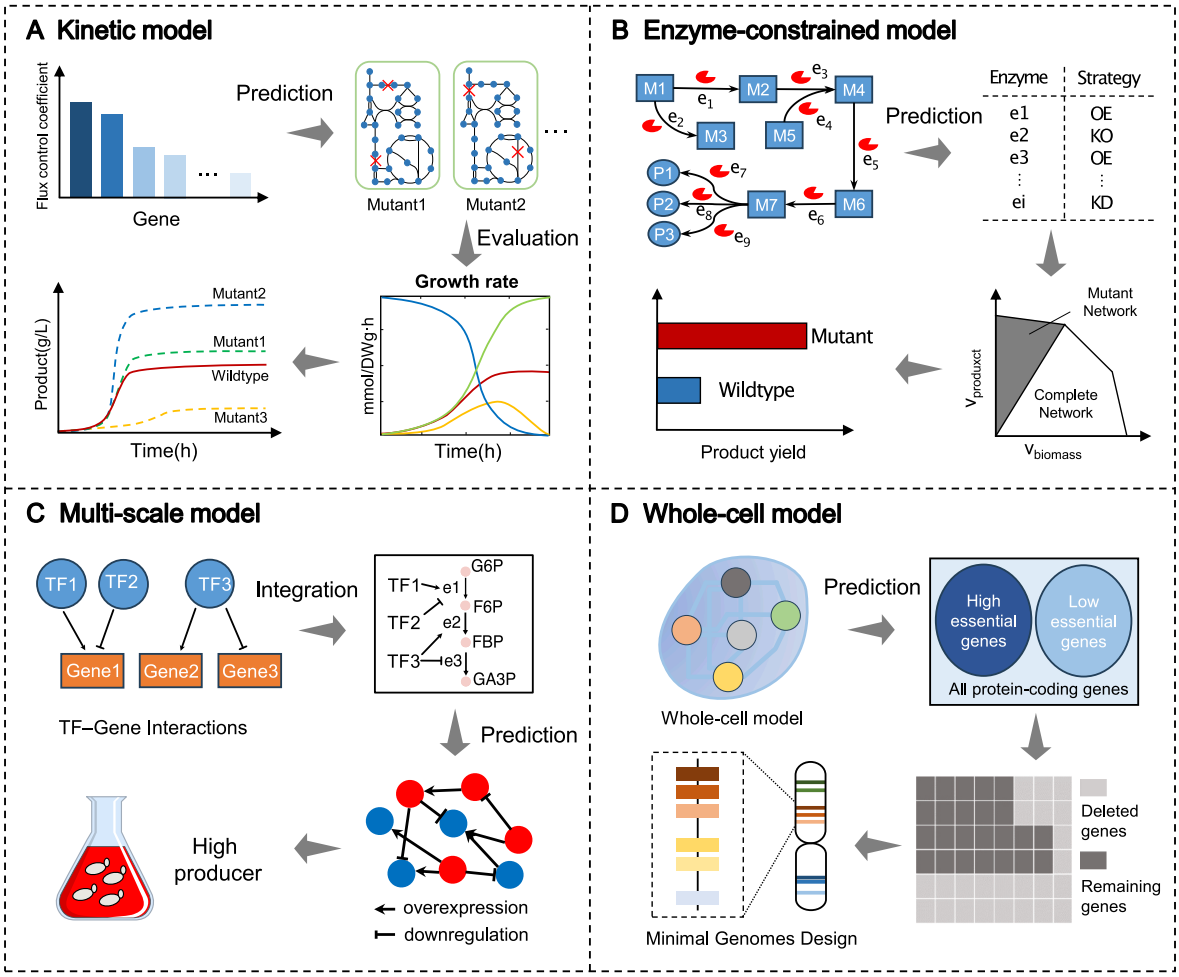

Fig. 2. Diverse applications of mechanistic metabolic models in computational strain design. The application of kinetic models in strain trait and internal flux prediction, with which the cellular growth rate and product concentration over time across mutants could be characterized (A). The application of enzyme-constrained models in enzyme sensitivity analysis and protein demand analysis, which helps to identify key enzymes influencing the productivity of cell factories (B). The application of multi-scale models. For example, GEMs could be combined with transcriptional regulatory networks to identify both TFs and gene targets for designing high-production strains (C). The application of whole-cell models (WCMs) in rational design of minimal genomes. With the aid of WCMs, all genes could be classified as essential or nonessential genes before the experimental reconstruction of a minimal genome (D).

prediction is still in its infancy. The latest progress in AI technology will promote the development of mechanistic models and their large-scale application in industrial biotechnology. However, how AI and complex mechanistic models can be combined remains to be determined. Here, we summarize the recent advances in computational strain design based on advanced modelling of cellular metabolism and its combination with MLs. The progress in the development of metabolic models based on MLs is also summarized. We will further discuss how cutting-edge MLs can be combined with metabolic modelling to accelerate the reconstruction of high-performance strain factories in the near future.

2. Retrospection of *in silico* strain design based on classical metabolic models

To date, most strain design algorithms have been developed based on stoichiometric metabolic models. Among those models, GEMs are widely used since there are multiple powerful computational platforms accessible, such as COBRApy (Ebrahim et al., 2013), COBRA Toolbox (Heirendt et al., 2019) and RAVEN (Wang et al., 2018), which include a variety of algorithms for this type of models. With several automated toolboxes, such as gapseq (Zimmermann et al., 2021), CarveMe (Machado et al., 2018) and model SEED (Henry et al., 2010), it is now possible to construct GEMs for any organism with whole-genome sequencing information. As a result, the computational strain design algorithms developed based on GEMs can be easily extended to a wide range of strains used in metabolic engineering, particularly for non-model organisms. Meanwhile, numerous *in silico* strain design algorithms have been developed. Some of them are frequently used, such as FSEOF (Choi et al., 2010) and optForce (Ranganathan et al., 2010), among which multiobjective optimization has also been incorporated into GEM-based *in silico* strain design. As a typical example, ModCell2 can systematically identify genetic modifications to design modular cells that can be coupled with a variety of production modules and exhibit a minimal trade-off among modularity, performance, and robustness (Garcia and Trinh, 2019). A detailed summary of these algorithms can be found in previous reviews (Lu et al., 2023; Machado and Herrgard, 2015).

Recently, based on GEMs, several packages for systematic design of strains have been developed using open-source platforms such as COBRApy (Ebrahim et al., 2013), which makes it more accessible for end-users from wet labs to employ these computational toolboxes. For example, MEWpy is a comprehensive platform written in Python that can execute strain design workflows with diverse types of metabolic models as input (Pereira et al., 2021). Within MEWpy, multiple evolutionary algorithms, including genetic algorithms and multiobjective optimization algorithms, have been adopted to carry out *in silico* strain

optimization. StrainDesign is another versatile Python package for computational strain design that uses constraint-based metabolic models and integrates various strain optimization algorithms (Schneider et al., 2022). Key algorithms such as OptKnock (Burgard et al., 2003), RobustKnock (Tepper and Shlomi, 2009), and OptCouple (Jensen et al., 2019), as well as the minimal cut sets (MCS) approach (Klamt et al., 2020), are included as separate optimization modules, requiring minimal input for the design objective. This package's standout feature is its ability to combine distinct modules, providing flexible choices for *in silico* strain design.

Although lots of algorithms based on GEMs have been developed, some common drawbacks exist in these toolboxes. First, as GEMs lack the constraints from protein synthesis, enzyme abundance and enzyme kinetics, they cannot be used to accurately predict the quantitative effects of gene overexpression, weakening, knockout and combinations of multiple gene manipulations on cellular growth and productivity (Gudmundsson and Nogales, 2021). The introduction of heterologous synthesis pathways into chassis cells is often accompanied by additional metabolic burden (Wu et al., 2016). Since simulation using GEMs cannot reflect the protein resource allocation among metabolic reactions/pathways, it is impossible to quantitatively predict the effects of metabolic perturbations caused by the introduction of exogenous genes without additional constraints and assumptions (Alsiyabi et al., 2022). These shortcomings of the stoichiometric metabolic model have proven to be bottlenecks in developing more efficient *in silico* strain design algorithms.

3. Rational strain design based on advanced metabolic models

Based on the aforementioned GEMs, additional constraints (i.e., enzyme kinetic parameters and abundance) and metabolic submodules (i.e., protein synthesis and degradation), have been added to develop more advanced models, including but not limited to kinetic models, enzyme-constrained models, multi-scale models and whole-cell models (Fig. 1). As these newly added constraints and metabolic submodules could narrow the solution space and expand the predictive capabilities, the mechanistic models have distinct advantages in comprehensively modelling cellular metabolism, thus providing new opportunities for the holistic design of various industrial microorganisms (Lu et al., 2021). In this section, we summarize the typical applications of these newly developed models for the computational design of widely used cell factories, i.e., *E. coli* and *S. cerevisiae.*

3.1. Kinetic models

In contrast to enzyme-constrained models, kinetic models could encompass detailed enzyme parameters, reaction kinetics and thermodynamics. Thus, kinetic models can be used to simulate the dynamic changes in metabolic fluxes, as well as the concentrations of metabolites and proteins under constraints from physiological and regulatory interactions (Hu et al., 2023; Mishra et al., 2023; St. John et al., 2019) (Fig. 2A). Currently, several algorithms based on kinetic models have been developed to guide the rational strain design. Chowdhury et al. first proposed the strain design algorithm k-OptForce (Chowdhury et al., 2014) based on kinetic models, referring to the published algorithm OptForce (Ranganathan et al., 2010). Similar to OptForce, the set of reactions associated with gene upregulation, downregulation and knockout was initially identified by comparing the fluxes between the wild-type and mutant strains before further filtering and refinement. In the output of k-OptForce, it was found that the predicted interventions avoid larger rearrangement of flux distributions due to constraints imposed by metabolite concentrations. With sensitivity analysis, it was also verified that the number of predicted interventions could be affected by the bounds of metabolic concentrations added as constraints to the model. Compared to the gene targets from stoichiometric models, k-OptForce could be used to find some novel targets for strain engineering, for example, the specific interventions to reduce the inhibition of enzymes by substrates. k-OptForce has been tested on large kinetic models of central metabolism for *E. coli* and *S. cerevisiae*, displaying its good performance in rational strain design. Recently, Khodayari et al. developed a medium-size kinetic model for *E. coli,* k-ecoli457, with 457 reactions and corresponding substrate-enzyme interactions (Khodayari and Maranas, 2016). Compared to traditional procedures using flux balance analysis (FBA), k-ecoli457 could better predict the yields of mutant strains for 24 bioproducts, and the Pearson correlation coefficient between the measured and predicted values reached 0.84, demonstrating the high predictive accuracy of this kind of model.

More recently, based on kinetic models, Narayanan et al. developed a new framework to carry out strain design, termed nonlinear dynamic model-assisted rational metabolic engineering design (NOMAD) (Narayanan et al., 2024). Within this framework, gene targets can be predicted by tuned kinetic models, and the outputs of these models still satisfy the constraints from real strain physiology. Therefore, the robustness of the engineered strains will be mostly maintained at levels comparable to that of reference strains. To achieve this goal, the kinetic models were carefully assessed and screened by comparing the model output with the actual strain physiology. To verify the unique value of this framework, Narayanan et al. used NOMAD to predict potential gene targets for the overproduction of anthranilate by *E. coli*. Eight gene targets identified in previous studies could be predicted by NOMAD, and novel gene targets were also generated by NOMAD for experimental verification. NOMAD was developed based on the reduced model of *E. coli;* thus, increasing the scope of kinetic models will further promote the application of NOMAD and other similar tools.

In addition to the aforementioned algorithms based on kinetic models, simulations generated by these models play a pivotal role in revealing the intrinsic dynamics of metabolism under experimental conditions. For instance, Lao-Martil et al. developed a physiology-informed kinetic model of yeast glycolysis intricately linked to central carbon metabolism. This model comprehensively accounts for the influence of anabolic reactions, precursors, mitochondria, and the trehalose cycle (Lao-Martil et al., 2023). Through this model, the metabolic dynamics underlying a 110 mM glucose pulse and various steady-state growth rates could be characterized in detail. Moreover, the model could help to elucidate the intracellular metabolic dynamics during the feast-famine regimen and reveal metabolic responses to alternative carbon sources, such as fructose, sucrose, and maltose.

3.2. Enzyme-constrained models

In the past several years, enzyme quantity- and enzyme kinetics-constrained genome-scale metabolic models (ecGEMs) have been introduced as refined metabolic models constructed on the basis of GEMs by adding additional constraints such as enzyme kinetic information and enzyme usage into the metabolic network (Bekiaris and Klamt, 2020; Sanchez et al., 2017). Compared to traditional GEMs, ecGEMs can be immediately used for the integration and analysis of omics data. For example, quantitative proteomics can be employed as additional constraints for the model to refine the maximal flux through each reaction (Fig. 1), thus ecGEMs have significantly improved accuracy in predicting metabolic fluxes and complex phenotypes (Domenzain et al., 2022). Until now, ecGEMs have been successfully used to predict the Crabtree effect of model microorganisms (such as baker's yeast) at higher growth rate (Sanchez et al., 2017) and the maximal growth rate of target strains under different carbon and nitrogen sources (Lu et al., 2019). Interestingly, the predictive capabilities of ecGEMs could be further extended by adding new constraints. For example, the cellular transcriptional regulatory network (TRN) could be combined with ecGEMs to quantitatively predict the effects of changes in transcription factor activity on cellular phenotypes (Österberg et al., 2021). Moreover, cofactor synthesis information could also be merged

into ecGEMs to quantitatively predict the impact of cofactor deficiency on cellular physiology (Chen et al., 2021).

Due to their ease of use and good predictive performance, ecGEMs have become an excellent platform for *in silico* strain design (Fig. 2B). By utilizing ecGEMs, the effects of gene manipulations, such as gene knockout, weakening, and strengthening, on production and strain phenotype can be quantitatively predicted, thereby accelerating the rational engineering of cell factories. In this regard, Li et al. used ec_iML1515 to calculate the protein demand for *E. coli* growth, as well as for shikimate production. Through simulation, 7 high-demand proteins (which needed to be overexpressed) and 5 low-demand proteins (which needed to be downregulated) were identified to improve shikimate biosynthesis. According to experimental validation, 11 of these 12 gene targets could successfully enhance strain performance in shikimate production (Li et al., 2023), thus demonstrating the efficiency of prediction by ecGEMs. With more constraints from enzymes, ecGEMs have also made it possible to predict combinations of multiple gene manipulations for systematic metabolic engineering. For instance, with the aid of the yeast ecGEM (ecYeast8), Ishchuk et al. successfully predicted the combination of 11 gene targets to enhance heme production (Ishchuk et al., 2022).

Owing to the advantageous flexibility and maintainability of ecGEMs, several computational strain design toolboxes have been built with this kind of model. For instance, Yao et al. developed a toolbox named PROSO (Yao and Yang, 2023), which could leverage ecGEMs to conduct *in silico* strain design. Multiple algorithms are included in PROSO, such as PC-OptKnock and the minimization of proteomics adjustments. These function modules in PROSO offer opportunities for systematic strain design in synthetic biology. Moreover, in our previous work, the classical FSEOF algorithm and the excellent predictive capabilities of ecGEMs were combined to develop a new computational platform (ecFactory) used for strain design (Domenzain et al., 2023). In ecFactory, the calculation of protein cost and the variability analysis of protein usage can be easily conducted to filter out unreasonable gene targets. More importantly, the minimal genetic modifications, including gene overexpression, knockout and knockdown, can be identified and used for experimental validation. It is anticipated that the novel strain design based on ecGEMs will further improve the accuracy in the prediction of potential gene targets for rational metabolic engineering.

3.3. Multi-scale models

The phenotype of a cell is determined by regulation at many different levels (Carthew, 2021). Multi-scale models could integrate information from different scales of cellular metabolism, which is useful for predicting gene targets and their regulatory role in order to improve the productivity of engineered strains.

Transcriptional regulation network models (TRNs) reflect the regulation between transcription factors (TFs) and their target genes (Chung et al., 2021), and GEMs integrated with TRNs can therefore be used to predict the pathways that are active under different conditions and thereby used for *in silico* strain design (Fig. 2C). Shen et al. developed an algorithm, OptRAM, based on TRN-regulated models (Shen et al., 2019). With OptRAM, gene manipulations, including overexpression, knockdown and knockout of both TFs and metabolic genes, could be predicted simultaneously. This procedure relies on high-quality TRNs. It is difficult to reconstruct high-quality TRNs, but various procedures have been developed to infer TRNs for different organisms, including machine learning (Erbe et al., 2022), statistical inference and chromatin immunoprecipitation (Pavesi, 2017). Furthermore, Liu et al. built a multi-scale model, etiBsu1209, for *Bacillus subtilis* by integrating enzymatic constraints, thermodynamic constraints and TRNs (Bi et al., 2023). This is currently the most comprehensive model for *B. subtilis*. This model was successfully used for *in silico* strain design, in which two knockout genes were correctly predicted to improve the titre of the nutraceutical menaquinone-7 by twofold. It is anticipated that other gene targets can also be predicted using this model. However, until now, few *in silico* strain design algorithms have been developed based on TRN constrained models. The main obstacle might be that it is currently challenging to quantitatively predict how TFs modulate the expression profiles of target genes.

In silico strain design can also be carried out based on more advanced multi-scale models, including ME models (O'Brien et al., 2013), pcModels (Elsemman et al., 2022) and ETFL models (Salvy and Hatzimanikatis, 2020), within which extra bioprocesses, including but not limited to gene transcription and translation, are coupled with normal biochemical metabolic pathways (Fig. 1). With these comprehensive multi-scale models, it is convenient to predict how enzyme efficiency affects cellular traits. To display the potential of ME models, Dinh et al. developed an *in silico* strain design pipeline that integrates predictions from both traditional GEMs and ME models (Dinh et al., 2018). Using GEMs, the knockout gene targets that make the product growth-coupled were first identified. Next, the robustness of these knockout gene targets was further evaluated using ME models by sampling the k_{cat} values of the enzymes. With such a strategy, 42 high-confidence designs could be identified from a total initial 634 significant growth-coupled production designs. This work demonstrated that enzyme efficiency is a decisive factor determining whether production is coupled with growth. Multi-scale models can be further extended for new applications. Li et al. integrated secretion pathways into a proteome-constrained model of yeast to develop a new kind of model, pcSecYeast (Li et al., 2022a). To characterize the secretion pathways, template reactions were defined to represent the detailed steps related to protein synthesis, protein modification, transport and secretion. With pcSecYeast, accurate correlations between productivity and growth could be predicted for various proteins. This shows that at a lower growth rate, protein production is coupled with growth, while as the growth rate further increases, the production rate decreases, which may be due to resource limitation within the cells. More interestingly, pcSecYeast can be used to predict gene targets from secretion pathways, thus experimentally increasing the production level of α-amylase.

Moreover, based on multi-scale models, Oftadeh et al. developed a novel computation toolbox, rETFL, by considering the consumption of protein and energy resources in the expression of plasmids (Oftadeh and Hatzimanikatis, 2024). It shows that rETFL could reflect how the number of plasmids affected growth and production simultaneously. Interestingly, it successfully predicted a reduction in the growth of *E. coli*, as well as a trade-off between cellular growth and protein production along with an increase in the copy number of plasmids. More importantly, with enzymatic constraints, rETFL could reflect the detailed enzyme resource reallocation under the insertion of plasmids. It reveals that once the resources become more limited when expressing recombinant proteins, the cell tends to synthesize enzymes with higher catalytic capacity. It is anticipated that such a framework will have more applications in characterizing metabolic rewiring for mutant strains with multiple heterologous genes. However, until recently, published *in silico* strain algorithms based on advanced multi-scale models have been scarce, which has hindered the application of these models for systematic metabolic engineering of industrial strains.

3.4. Whole-cell models

Whole-cell models (WCMs) are among the most complex models used to simulate strain performance from genomic and environmental inputs (Goldberg et al., 2018). Currently, whole-cell models have been built for *Mycoplasma genitalium* (Karr et al., 2012), *S. cerevisiae* (Ye et al., 2020), *E. coli* (Macklin et al., 2020) and JCVI-syn3A (a minimal cell with a reduced genome of 493 genes) (Thornburg et al., 2022). In contrast to the aforementioned models, whole-cell models cover more metabolic modules related to gene and protein synthesis (as well as regulation) (Carrera and Covert, 2015) (Fig. 1). Thus, WCMs can be employed to predict how gene targets at a larger genome level influence the

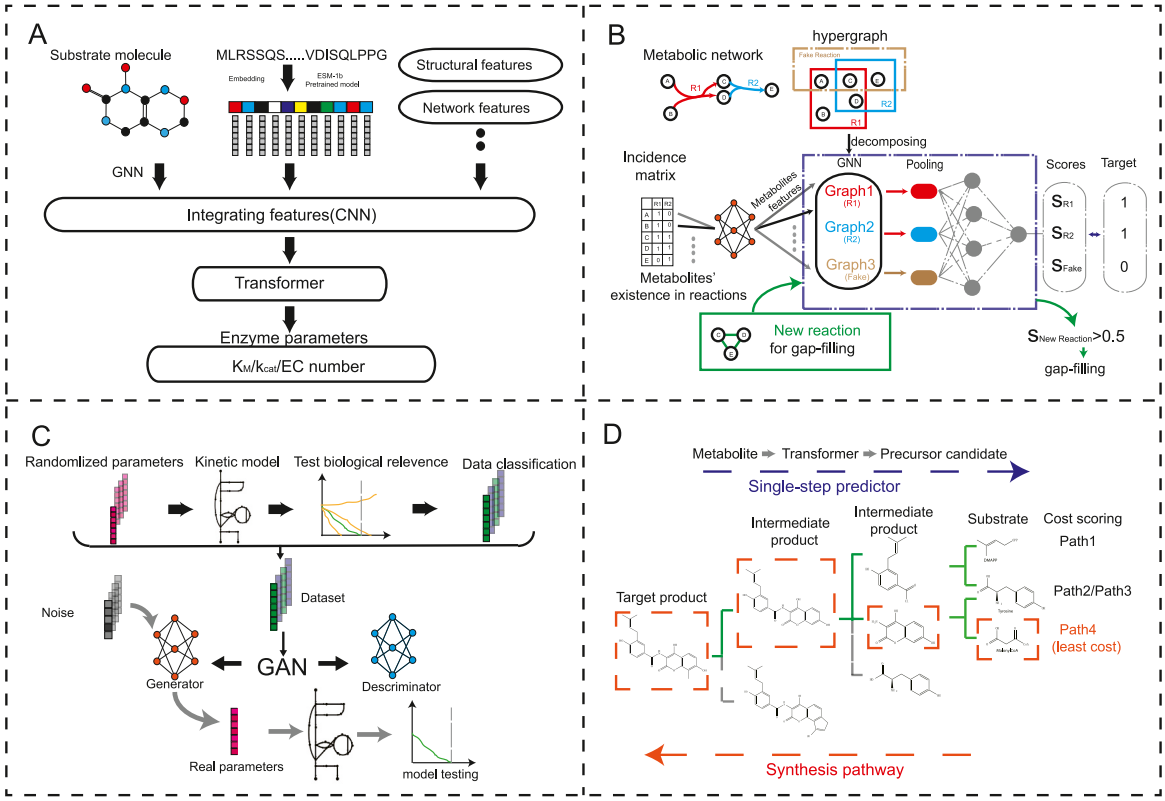

Fig. 3. Roles of MLs in cellular modelling from the prediction of enzyme properties to the reconstruction of multi-scale metabolic models. With the aid of graph neural networks (GNNs) and convolutional neural networks (CNNs), features are extracted from protein sequences and substrate structures to predict key enzymatic parameters such as k_{cat} and K_M (A). Gaps always exist in newly built GEMs, which can be found and filled by various MLs. For instance, a deep learning model, CHESHIRE, based on a hypergraph was established to find candidate reactions suitable for filling gaps in draft GEMs. In the hypergraph, the metabolic network was represented as the metabolite-reaction incidence matrix, which was then fed into GNNs to extract the network topology features (B). To optimize the parameters of kinetic models, random sampling is initially used to generate parameters for kinetic model reconstruction, and the corresponding model performances are evaluated one by one. Subsequently, the dataset is employed to train a generative adversarial network (GAN), where the generator is capable of producing candidate parameters for refinement of kinetic models (C). A transformer-based ML is utilized to design *de novo* biosynthesis pathways, starting from the target product and progressively tracing back to the original substrates. Afterwards, those pathways are further scored and filtered to screen the best pathway for experimental validation (D).

efficiency of cellular growth and production (Marucci et al., 2020).

The acquisition of a minimal genome is extremely important for understanding the fundamentals of cellular metabolism and its evolution (Moger-Reischer et al., 2023). However, it is difficult to test the combinations of all essential genes experimentally; therefore, WCMs can be applied for the computational design of a minimal genome based on their powerful predictive capabilities (Fig. 2D). To redesign the minimal genome of *M. genitalium* (Rees-Garbutt et al., 2020, 2021), Rees-Garbutt et al. developed two algorithms, Minesweeper and Guess/Add/Mate (GAM), to produce a novel design of a minimal genome for *M. genitalium* based on the published WCMs for *M. genitalium* (Karr et al., 2012). Through design-simulate-test cycles with WCMs, Rees-Garbutt et al. identified 10 low essential genes that could reduce the size of the minimal genome compared to the reported one, JCVI-Syn3.0 (Hutchison et al., 2016) (a version of the minimal genome for *M. genitalium*), which provides novel clues for further experimental evaluation. Although the prediction of a minimal genome from WCMs is still not very accurate, subsequent experiments can improve the performance of WCMs by providing more data reflecting the correlations between genotype and phenotype. In addition to the prediction of a minimal genome, deep curation of datasets from different layers or different sources is also very useful for identifying steps impacting cell phenotypes in WCMs. As WCMs can be used to characterize cellular phenotypes in a holistic way, the role of each individual enzyme parameter could be evaluated (Macklin et al., 2020). Thus, WCMs will be useful in curating datasets for robust strain design.

4. MLs can accelerate the reconstruction of computational models

Currently, mechanistic metabolic models only cover some of the main cellular metabolic activities and ignore metabolic activities with unclear mechanisms. For example, the stress response processes under extreme growth conditions, i.e., low pH and high temperature, have not been considered in the model reconstruction for most non-model organisms. Sometimes, due to the highly nonlinear and dynamic characteristics of metabolic activities, it is challenging to accurately delineate the complex interactions among metabolites, enzymes and pathways. Therefore, the computational targets or strategies predicted for strain improvement from metabolic models may be limited, slowing down the construction of more efficient strains. Complementary to mechanistic models, machine learning can integrate numerous variables and predict the strain performance from high-throughput datasets based on so-called black-box models (Angermueller et al., 2016). Thus, MLs could to some extent overcome the bottlenecks of metabolic models and promote the computational design of cell factories based on the learned knowledge from big data.

4.1. MLs for characterizing enzyme kinetic properties

The reconstruction of a mechanistic model always relies on various kinds of parameters, most of which are used for characterizing enzyme properties, i.e., k_{cat} and K_m. Previously, the lack of enzyme kinetic parameters made it difficult to construct enzyme- or kinetics-constrained models (Nilsson et al., 2017). The development of MLs for the prediction of k_{cat} and K_m is now filling this gap (Fig. 3A). There are currently two types of methods for predicting enzyme kinetic parameters using MLs. One is to train traditional MLs (such as random forests and support vector machines) based on structured datasets. As one typical example, Heckmann et al. used different ML algorithms (linear, PLSR, elastic net, random forest and deep neural network) to predict *in vitro* and *in vivo* k_{cat} values (Heckmann et al., 2018). The prediction of k_{cat} could aid to enhance the coverage of k_{cat} for enzymes from the proteome of *E. coli*. Furthermore, it revealed that the predicted k_{cat} at the proteome scale could improve the model accuracy in the prediction of protein abundance. Feature importance analysis showed that fluxes through each reaction and the structural features of proteins, such as the active site depth and active site exposure, contribute significantly to the k_{cat} prediction. However, with limited datasets from the model organism *E. coli*, different MLs exhibited similar performances in terms of *in vitro* and *in vivo* k_{cat} prediction. It should also be noted that the aforementioned MLs were only trained for *E. coli* and cannot be directly applied to other organisms.

To mitigate the above issues, various deep learning models were trained based on large-scale enzymatic datasets stored in the BRENDA and SYBIO databases. In 2021, Kroll et al. developed a deep learning model to directly predict the K_m of an enzyme based on protein sequence and substrate structure information. Although the goodness of fit (R^2) between the predicted values and the experimental ones was only 0.53, this model could predict K_m values for a variety of enzyme–substrate combinations in non-model organisms (Kroll et al., 2021). At the same time, Li et al. utilized convolutional neural networks to extract protein sequence features and constructed a deep learning model, DLKcat (Li et al., 2022b), which can predict the k_{cat} values of enzymes at a large scale. Notably, in the validation dataset, the R^2 of DLkcat was approximately 0.5, meaning that it would generate unreliable predictions. Recently, Qiu et al. further developed a deep learning model, DLTKcat, which can predict how enzyme k_{cat} values are influenced by protein sequence and temperature (Qiu et al., 2024), thus promoting the reconstruction of metabolic models constrained by temperature. In order to further improve model accuracy in k_{cat} prediction, some latest models, including UniKP (Yu et al., 2023a) and DeepEnzyme (Wang et al., 2023), have been released for the scientific community. However, the current deep learning models cannot fully characterize how a single mutation affects the kinetic parameters of an enzyme; thus, there is room for continuous generation of high-quality datasets and systematic optimization of the deep learning framework to improve the model performance in terms of accuracy and generalization.

4.2. MLs in the reconstruction of advanced metabolic models

When developing mechanistic models, it is essential to infer the detailed functions of enzymes, as it is not feasible to experimentally test the metabolic functions of all enzymes in a molecular network. Fortunately, MLs have good performance in predicting the functions of enzymes, including their corresponding EC number, compartment and catalysed reactions within the cell. As one of the typical examples, Kroll et al. proposed a general MLs that could accurately predict enzyme-substrate pairs (Kroll et al., 2023), thereby helping to uncover the catalytic functions of the target enzymes. Moreover, a recent deep learning model, CLEAN, proposed by Yu et al. could successfully predict the EC number based on protein sequences using a contrast learning strategy (Yu et al., 2023c). Additionally, Kim et al. released a novel deep learning model, DeepECtransformer, which could utilize the transformer to successfully predict the EC number for uncharacterized genes (Kim et al., 2023). It also shows that DeepECtransformer can extract hidden features from enzyme sequences (functional motifs or residue sites) when inferring metabolic functions. This, to some extent, improves the interpretability of MLs. Specially, the latest progress in MLs has made it possible to construct high-quality GEMs for any newly sequenced species using a bottom-up procedure. For example, Sandra et al. used a multi-label ensemble model to predict protein localization in eukaryotic organisms. Additionally, a new tool, CarveFungi, was designed based on the previous GEM automation tool CarveMe to incorporate protein localization information into the pipeline for the automated reconstruction of genome-scale metabolic models (Castillo et al., 2023). Such a strategy could improve the quality of GEMs, as verified by better classification of different species using these newly built GEMs. As a second example, Chen et al. updated gap filling approaches based on hypergraph learning to generate functional GEMs (Chen et al., 2023) (Fig. 3B). With this method, the ability of 49 draft GEMs to predict the secretion of fermentation products and amino acids could be considerably enhanced.

Kinetics- or enzyme-constrained models require much work for tuning the kinetic parameters. Various ML procedures have been used to optimize kinetic or enzyme-constrained models (Fig. 3C). For example, Choudhury et al. developed the deep learning-based framework REKINDLE to quickly reconstruct feasible kinetic models reflecting the dynamic properties of cell metabolism (Choudhury et al., 2022). In this framework, the input kinetic parameters from Monte Carlo sampling were evaluated and classified as biologically relevant or not relevant, after which the generative adversarial network model was trained based on labelled datasets to efficiently generate the optimal parameters with the objective of building kinetic models with biological meaning. To further reduce the computational time in kinetic model reconstruction, Choudhury et al. built a novel strategy named RENAISSANCE, which can be used to efficiently parameterize kinetic models to represent the real dynamic properties of a target organism (Choudhury et al., 2023). As a kind of generative MLs, RENAISSANCE combines the advantages of both artificial networks and natural evolutionary strategies to realize stratified sampling for the generation of desired kinetic models. Such a procedure could reduce the time for the reconstruction of large-scale kinetic models.

In addition to kinetic models, ML algorithms are useful to maximize the performance of different enzyme-constrained models. By fine-tuning enzyme kinetic parameters based on Bayesian statistical learning, Li et al. developed enzyme- and temperature-constrained GEMs (etcGEMs), which could accurately characterize the impact of temperature on enzyme activity and growth rate (Li et al., 2021).

Currently, the biosynthetic pathways for some of the natural products are still not known. Thus, it is highly valuable to design efficient synthetic pathways for these valuable bioproducts (Sveshnikova et al., 2022). Due to the accumulation of biochemical and chemical reactions, it becomes easier to extract reaction rules using advanced MLs (Fig. 3D). With these reaction rules in hand, the synthetic pathways can be predicted ab initio using computational models. For example, utilizing transformer neural networks, Zheng et al. developed a toolkit named BioNavi-NP (Zheng et al., 2022), which can successfully predict single or multiple steps of reactions required for biosynthesis of biochemicals. Although there are still certain difficulties in predicting reactions for the synthesis of extremely complex natural compounds, BioNavi-NP offers alternative options to create synthetic routes that are superior to those created by conventional methods. Additionally, deep learning models based on convolutional neural networks (CNNs) assist in the directed selection of enzymes by screening enzyme candidates for uncharacterized reactions in retrosynthetic pathways, thereby contributing to the refinement and completion of the entire retrosynthetic process (Upadhyay et al., 2023). A more detailed review of MLs for predicting the retrosynthetic pathways of bioproducts can be found in (Yu et al., 2023b).

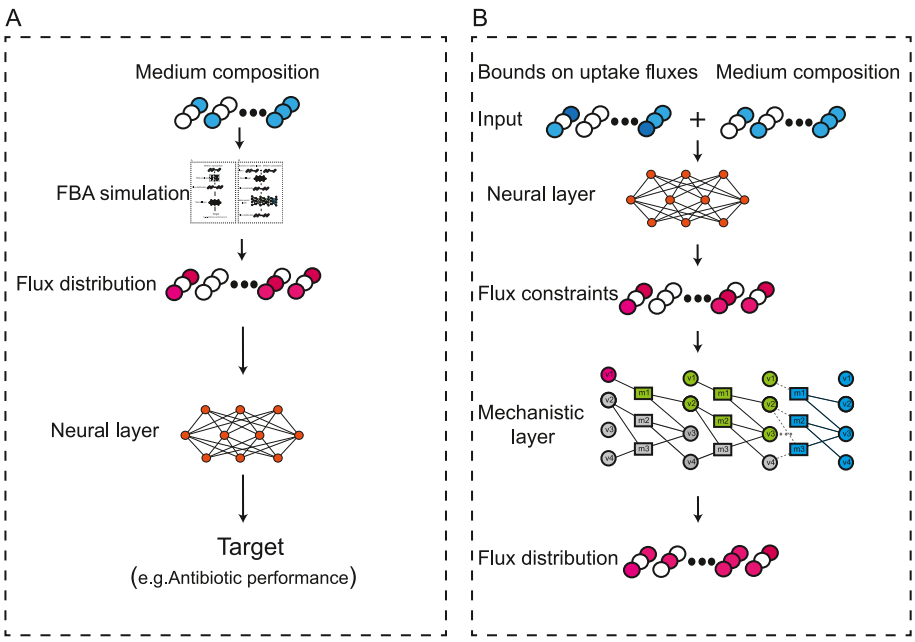

Fig. 4. Two main procedures used for combining MLs and metabolic models. Fluxomic data from FBA simulations based on GEMs can be utilized directly to train neural network models, increasing the interpretability of MLs (A). Mechanistic models of metabolic networks, such as GEMs, are employed in architectures such as RNNs to construct feedforward neural networks, thereby maximizing the learning of the intrinsic mechanisms of metabolic networks (B).

4.3. MLs in prioritizing gene targets for strain engineering

Advances in gene editing have made it possible to modify the genome of strains on a large scale. During the design-build-test-learn (DBTL) cycle, large datasets can be collected to evaluate strain performance. In this respect, advanced data analytics technologies are critical to extract knowledge from the big data for the next round of strain engineering. Multiple types of MLs are well suited for data preprocessing and integration, as well as for feature extraction from high-dimensional, nonlinear datasets. To speed rational metabolic engineering of industrial strains, various computational frameworks based on sophisticated MLs have been proposed to predict gene targets for the optimization of microbial cell factories.

As an exciting example, Radivojević et al. proposed a novel package, the Automated Recommendation Tool (ART) (Radivojevic et al., 2020), which leverages machine learning algorithms from a scikit-learning library (https://scikit-learn.org/) and can guide experiments effectively by automatically predicting gene targets from various inputs, such as proteomics, transcriptomics, and promoter combinations. Additionally, ART offers suggestions for subsequent DBTL cycles, and fresh datasets from more trials can be used to fine-tune the algorithm even further to lower prediction uncertainty. Similar to other MLs, the ART predictive performance relies on the availability of datasets. Notably, once automation is introduced, large-scale datasets become accessible, and ART can predict engineering strategies for synthetic biology more accurately.

The reinforcement learning approach has also been applied to computational strain design. As a proof of concept, Sabzevari et al. devised an approach known as multiagent reinforcement learning (RL) (Sabzevari et al., 2022), which can be used to tune enzyme concentrations from experiments to enhance production levels. As a model-free strategy, the method described by Sabzevari et al. does not require detailed prior knowledge of the metabolic system. It could realize the iterative update of strain performance based on predictions from prior rounds. With the collection of high-throughput datasets, RL will be able to make better use of big data and find smarter strategies for strain design.

It is well known that the reconstruction of kinetic models relies on understanding the detailed mechanism of the reactions and the associated enzyme kinetic parameters (Saa and Nielsen, 2017); however, the mathematical expression used to represent the reaction kinetics might be complex. In contrast, MLs can use convenient and direct approaches to predict the dynamics of metabolic pathways without knowing the detailed mechanisms underlying the molecular kinetics. A study by Costello et al. (Costello and Martin, 2018) showed that, using multilayer omics datasets as input, MLs can be trained to predict the dynamics of metabolic pathways, outperforming classical kinetic models to some extent. In their MLs, the time series of proteomics and metabolomics datasets were used as input to predict the derivative of metabolite concentration. Even with limited training datasets, Costello et al. reported that their MLs exhibited good predictive accuracy (Costello and Martin, 2018). More interestingly, they verified that statistical analysis of MLs, such as PLS, could help to identify critical proteins that have a significant impact on the production of target products, thereby providing clues for rational strain engineering.

5. New paradigm of strain design based on the hybrid of MLs and metabolic models

5.1. How can MLs and metabolic models be linked?

Both MLs and metabolic models have advantages and drawbacks. For example, ML approaches, such as neural networks, require large-scale, high-quality datasets for training due to the large number of network parameters, yet they possess strong learning capabilities. Conversely, mechanistic models represented by GEMs can be used without a large amount of data, but the refinement of these models requires a profound understanding of the functional mechanisms of the biological system. The integration of these two kinds of models will undoubtedly help to address some unresolved issues in precise design of cell factories. However, the differences between MLs and metabolic models make it difficult to directly combine these two kinds of models together since mechanistic models lack the forward propagation ability to automatically adjust parameters, as neural networks do.

To date, two main procedures have been used to integrate metabolic models with MLs. First, the outputs from metabolic models, such as fluxes, can be regarded as additional parameters to train MLs (Fig. 4A). If

the significant features revealed by MLs belong to the fluxes from metabolic models, the interpretability of MLs will be significantly improved. Using this procedure, the potential molecular mechanism underlying antibiotic efficacy was revealed (Yang et al., 2019). Second, the metabolic model itself could be transformed as important features for training MLs (Czajka et al., 2021; Lee et al., 2022). For instance, the condition-specific GEMs produced by omics constraint algorithms may be utilized to illustrate the presence of particular subpathways. These GEMs might subsequently be fed into ML algorithms to train them to predict disease states or classify cell types.

Recently, several state-of-the-art algorithms were developed to incorporate metabolic models or their network topologies as integral components within deep learning architectures (Fig. 4B), which could be directly used in model training and iterative optimization. In a recent study, Faure et al. proposed three innovative optimizers (Wt, LP, and QP) based on mathematical optimization principles to replace the classical FBA (Faure et al., 2023). Contrary to FBA, which prioritizes maximization of biomass as the objective function, these novel algorithms mainly define a loss function that could gradually minimize discrepancies between model outputs and flux values; the latter were set as the training label through backpropagation of gradients. To reduce the number of iterations, the input flux boundary vector was first used as input for a neural network module prior to its introduction into the optimizer, thereby constructing an artificial metabolic network (AMN) hybrid model that combines mechanistic models and neural networks. This proved that AMN has excellent capabilities in the quantitative prediction of strain growth rates that are not achieved by traditional FBA, while traditional FBA still performs well in classification tasks, i.e., predicting growth and nongrowth. Furthermore, Hasibi et al. developed a new algorithm, FlowGAT, by integrating FBA simulation and MLs to achieve accurate prediction of gene essentiality (Hasibi et al., 2024). In this procedure, the authors first used a graph structure to represent the metabolic fluxes calculated by FBA for the wild-type strain. In such a graph structure, the nodes represent the enzymatic reactions, while the edges represent the mass fluxes among nodes. Based on the fitness data from strains with gene knockout, the graph structure could be integrated into a graph neural network for model training. This finding demonstrated that even under different growth conditions (i.e., changing the substrate), the ability of FlowGAT to predict gene essentiality could be comparable to that of FBA. This procedure fully showcases the great potential of hybrid models by combining the mechanistic insights from metabolic models with the powerful predictive capabilities of deep learning models.

5.2. Typical applications in computational strain design by combining mechanistic models and MLs

Systematic metabolic engineering is pivotal in the development of efficient cell factories wherein the selection of gene target and their subsequent optimization can be achieved through joint prediction using both mechanistic models and MLs. For example, in Zhang's work (Zhang et al., 2020), biological gene targets for improving tryptophan production were first predicted based on the mechanistic model of *S. cerevisiae*, Yeast8. Using this procedure, the four top gene targets, CDC19, TKL1, TAL1, and PCK1, were chosen to guide the wet-lab experiments, along with another gene, PFK1, which was discovered from human experience. Two machine learning algorithms, the Automated Recommendation Tool (ART) and EVOLVE, were employed to predict strain performance when 5 target genes were regulated by distinct promoter combinations (Zhang et al., 2020). Although a limited dataset (covering 250 genotypes after quality filtering) was used, the delicate promoter combinations were identified to further increase tryptophan titre and productivity by 74% and 43%, respectively, compared to the best strain from the training datasets.

In contrast to the above procedure, the fluxes from metabolic models could be used to train MLs to improve their accuracy in prediction of strain phenotyps. For this reason, Czajka et al. first collected phenotypic data for *Yarrowia lipolytica*, including the cultivation conditions, gene backgrounds, and production levels (Czajka et al., 2021). Using these datasets, the fluxes via each reaction in a *Y. lipolytica* GEM could be calculated. Afterwards, these fluxes and other parameters were used as inputs to train MLs, which subsequently could predict the strain production levels based on all input features. This showed that, by integrating the flux features, the ensemble model achieved accurate predictions when production titres surpassed 1 g/L, but further improvements are needed for predictions at lower production levels (<1 g/L). Moreover, the FBA flux was ranked as one of the most important features impacting MLs prediction, illustrating that the reaction fluxes calculated by GEMs could, to some extent, elevate the predictive capability of MLs.

6. Challenges and perspectives

The construction of an efficient cell factory relies on systematic work, including protein engineering, rational design of metabolic pathways, adaptive evolution, and process optimization. Mechanistic metabolic models, such as GEMs, ecGEMs, and ETFLs, encompass detailed functional mapping of genes, proteins and reactions (metabolites), which could provide reliable guidance for reprogramming cellular metabolism to increase productivity. To date, great progress has been made in the rational design of various cell factories using simulations and predictions from both metabolic models and MLs. It has been witnessed that more comprehensive metabolic models are now being created by combining constraints from kinetics, regulation, and cell cycle models with the help of cutting-edge MLs, which considerably improve the efficiency of computational strain design.

However, there are still some limitations in the application of advanced metabolic models for intelligent strain design. First, although large amounts of omics data are generated daily, current models have difficulty in omics integration, especially for those from multi-scale levels. For instance, it seems difficult for current metabolic models to integrate metabolomics, RNA-seq and proteomics simultaneously. There are also inherent inconsistencies between the different levels of omics datasets. Additionally, even in the widely used model organisms *E. coli* and yeast, the regulation underlying flux control, cell division, and stress adaptation has not been fully understood. Thirdly, great efforts are still needed to enhance the quality of metabolic models for most non-model organisms used in industrial biotechnology. Although the most complex models, i.e., WCMs and ETFLs, have been built for some model organisms, it is not convenient to directly transfer them to non-model organisms, which has become one of the major bottlenecks in utilizing these advanced metabolic models on a broader scale. Finally, although many *in silico* strain algorithms have been developed based on stoichiometric metabolic models, i.e., classical GEMs, only a few of them are suitable for more advanced metabolic models, such as ecGEMs and ME models. The lack of user-friendly software limits the wider use of these advanced models in current systematic strain development.

Complementary to mechanistic metabolic models, MLs have obvious advantages in terms of powerful data integration, easy access, a mature ecosystem and high scalability, thus helping to mitigate the aforementioned challenges existing in advanced metabolic models. When sufficient datasets are available, MLs can make accurate predictions through reasonable training, and the learned information from MLs can be transferred from one organism to another, thus increasing the applications of MLs in solving challenges in industrial biotechnology. This could be one reason for the rapid development and deployment of various MLs in synthetic biology platform companies such as Ginkgo and Amyris. Note that the efficient use of MLs requires large amounts of data. In this regard, the high-throughput automatic robotic systems, that can generate high-quality standard datasets, have become indispensable for the development of stronger ML algorithms. Recently, with the breakthroughs in large pre-trained language models, i.e., ChatGPT v4.0 and

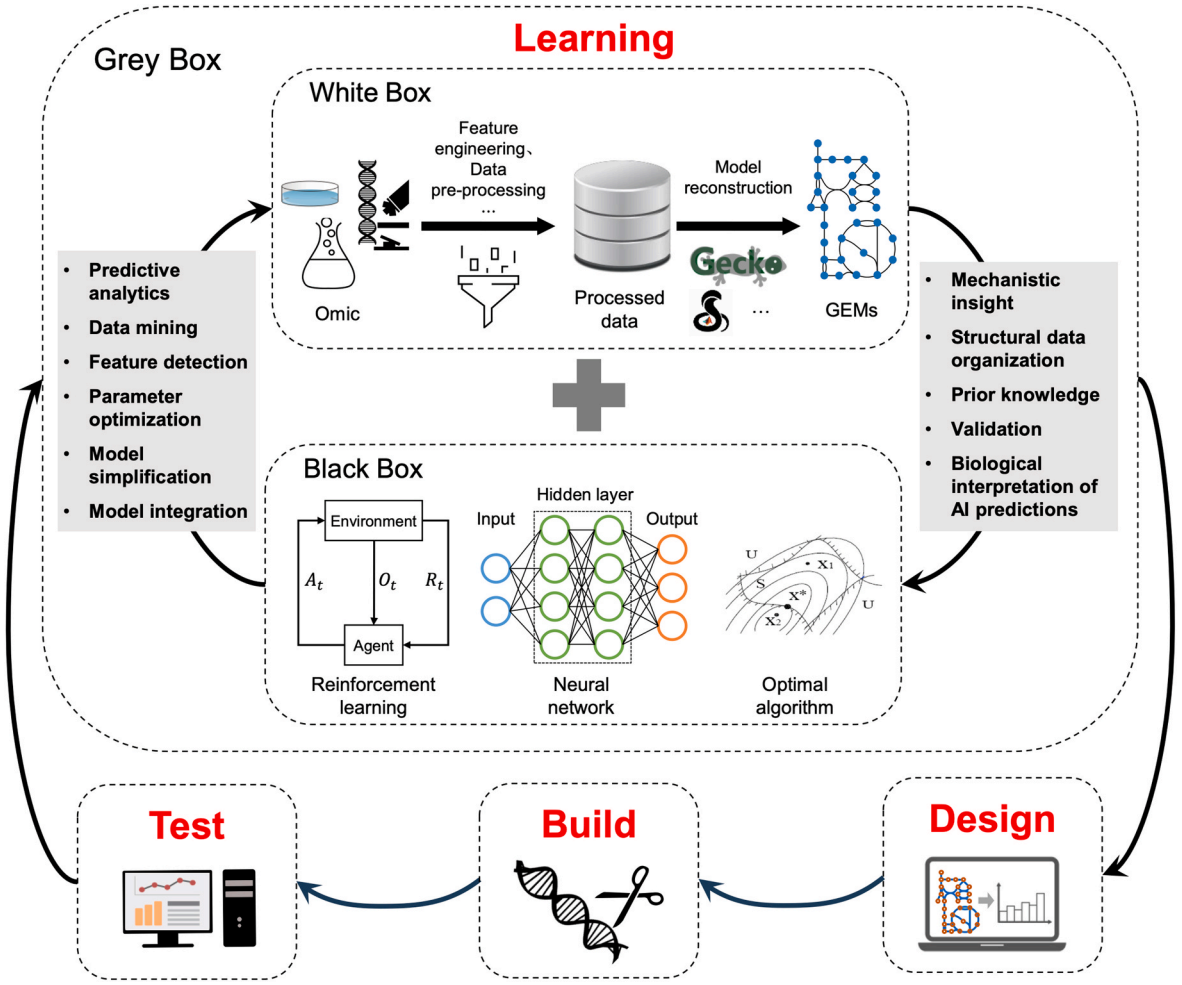

Fig. 5. Consistent integration of mechanistic models and AI in the design-build-test-learn (DBTL) cycle of strain development. The synergy between mechanistic models (white box) and artificial intelligence (black box) within the DBTL cycle helps to overcome the limitations of each type of model. During the learning phase, AI technologies can enhance mechanistic models through enzyme parameter prediction, automatic model construction, holistic model refinement and so on. Mechanistic models could offer biological insights, molecular network topology, and prior knowledge to create hybrid AI models with increased interpretability and prediction accuracy. This deep integration of mechanistic models and sophisticated AI models will provide robust and reliable solutions for the precise design of next-generation cell factories, hence accelerating the DBTL cycle used in systematic metabolic engineering.

Gemini v1.0, artificial general intelligence (AGI) has become a game changer for scientific research, which undoubtedly provides new opportunities for industrial biotechnology. It is believed that the performance of AI could certainly outperform that of mechanistic metabolic models in some respects. However, the interpretability of MLs alone is relatively weak. To overcome this problem, hybrid models integrating MLs (including AGI) and metabolic models will become highly valuable to accelerate the DBTL cycle for strain engineering (Fig. 5). Through such deep integration, rational and precise design for the reconstruction of next-generation cell factories will become feasible, which will greatly promote consistent development towards a bio-based economy.

Funding

This work is supported by grant 2022YFA0913000 from the National Key R&D Program of China, the Shanghai Pujiang Program, and grants 22208211 and 22378263 from the National Natural Science Foundation of China (NSFC).

Conflict of interest statement

Nothing declared.

CRediT authorship contribution statement

Hongzhong Lu: Writing – review & editing, Writing – original draft, Conceptualization. **Luchi Xiao:** Writing – original draft, Investigation. **Wenbin Liao:** Writing – original draft, Investigation. **Xuefeng Yan:** Writing – review & editing. **Jens Nielsen:** Writing – review & editing, Conceptualization.

Data availability

No data was used for the research described in the article.

References

Alsiyabi, A., Chowdhury, N.B., Long, D., Saha, R., 2022. Enhancing in silico strain design predictions through next generation metabolic modeling approaches. Biotechnol. Adv. 54, 107806.
Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., 2016. Deep learning for computational biology. Mol. Syst. Biol. 12, 878.
Bekiaris, P.S., Klamt, S., 2020. Automatic construction of metabolic models with enzyme constraints. BMC Bioinf. 21, 19.
Bi, X., Cheng, Y., Xu, X., Lv, X., Liu, Y., Li, J., Du, G., Chen, J., Ledesma-Amaro, R., Liu, L., 2023. etiBsu1209: a comprehensive multiscale metabolic model for Bacillus subtilis. Biotechnol. Bioeng. 120, 1623–1639.

Burgard, A.P., Pharkya, P., Maranas, C.D., 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84, 647–657.
Carrera, J., Covert, M.W., 2015. Why build whole-cell models? Trends Cell Biol. 25, 719–722.
Carthew, R.W., 2021. Gene regulation and cellular metabolism: an essential partnership. Trends Genet. 37, 389–400.
Castillo, S., Peddinti, G., Blomberg, P., Jouhten, P., 2023. Reconstruction of Compartmentalized Genome-Scale Metabolic Models Using Deep Learning for over 800 Fungi. bioRxiv., 2023.08.23.554328.
Chen, C., Liao, C., Liu, Y.Y., 2023. Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning. Nat. Commun. 14, 2375.
Chen, Y., Li, F., Mao, J., Chen, Y., Nielsen, J., 2021. Yeast optimizes metal utilization based on metabolic network and enzyme kinetics. Proc. Natl. Acad. Sci. U.S.A. 118, e2020154118.
Choi, H.S., Lee, S.Y., Kim, T.Y., Woo, H.M., 2010. In silico identification of gene amplification targets for improvement of lycopene production. Appl. Environ. Microbiol. 76, 3097–3105.
Choudhury, S., Moret, M., Salvy, P., Weilandt, D., Hatzimanikatis, V., Miskovic, L., 2022. Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks. Nat. Mach. Intell. 4, 710–719.
Choudhury, S., Narayanan, B., Moret, M., Hatzimanikatis, V., Miskovic, L., 2023. Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. bioRxiv, 529387, 2023.02.21.
Chowdhury, A., Zomorrodi, A.R., Maranas, C.D., 2014. K-OptForce: integrating kinetics with flux balance analysis for strain design. PLoS Comput. Biol. 10, e1003487.
Chung, C.H., Lin, D.W., Eames, A., Chandrasekaran, S., 2021. Next-generation genome-scale metabolic modeling through integration of regulatory mechanisms. Metabolites 11, 606.
Costello, Z., Martin, H.G., 2018. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. npj Systems Biology and Applications 4, 19.
Czajka, J.J., Oyetunde, T., Tang, Y.J., 2021. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction. Metab. Eng. 67, 227–236.
Dinh, H.V., King, Z.A., Palsson, B.O., Feist, A.M., 2018. Identification of growth-coupled production strains considering protein costs and kinetic variability. Metab. Eng. Commun. 7, e00080.
Domenzain, I., Lu, Y., Shi, J., Lu, H., Nielsen, J., 2023. Computational biology predicts metabolic engineering targets for increased production of 102 valuable chemicals in yeast. bioRxiv, 2023.01.31.526512.
Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E.J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrissey, J.P., Sonnenschein, N., Nielsen, J., 2022. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nat. Commun. 13, 3766.
Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R., 2013. COBRApy: COnstraints-based reconstruction and analysis for Python. BMC Syst. Biol. 7, 74.
Elsemman, I.E., Rodriguez Prado, A., Grigaitis, P., Garcia Albornoz, M., Harman, V., Holman, S.W., van Heerden, J., Bruggeman, F.J., Bisschops, M.M.M., Sonnenschein, N., Hubbard, S., Beynon, R., Daran-Lapujade, P., Nielsen, J., Teusink, B., 2022. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies. Nat. Commun. 13, 801.
Erbe, R., Gore, J., Gemmill, K., Gaykalova, D.A., Fertig, E.J., 2022. The use of machine learning to discover regulatory networks controlling biological systems. Mol. Cell 82, 260–273.
Faure, L., Mollet, B., Liebermeister, W., Faulon, J.L., 2023. A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. Nat. Commun. 14, 4669.
Garcia, S., Trinh, C.T., 2019. Multiobjective strain design: a framework for modular cell engineering. Metab. Eng. 51, 110–120.
Goldberg, A.P., Szigeti, B., Chew, Y.H., Sekar, J.A., Roth, Y.D., Karr, J.R., 2018. Emerging whole-cell modeling principles and methods. Curr. Opin. Biotechnol. 51, 97–102.
Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., 2022. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23, 40–55.
Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., Lee, S.Y., 2019. Current status and applications of genome-scale metabolic models. Genome Biol. 20, 121.
Gudmundsson, S., Nogales, J., 2021. Recent advances in model-assisted metabolic engineering. Curr. Opin. Struct. Biol. 28, 100392.
Hasibi, R., Michoel, T., Oyarzún, D.A., 2024. Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. npj Systems Biology and Applications 10, 24.
Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., Palsson, B.O., 2018. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat. Commun. 9, 5252.
Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., Magnusdóttir, S., Ng, C.Y., Preciat, G., Žagare, A., Chan, S.H.J., Aurich, M.K., Clancy, C.M., Modamio, J., Sauls, J.T., Noronha, A., Bordbar, A., Cousins, B., El Assal, D.C., Valcarcel, L.V., Apaolaza, I., Ghaderi, S., Ahookhosh, M., Ben Guebila, M., Kostromins, A., Sompairac, N., Le, H.M., Ma, D., Sun, Y., Wang, L., Yurkovich, J.T., Oliveira, M.A.P., Vuong, P.T., El Assal, L.P., Kuperstein, I., Zinovyev, A., Hinton, H. S., Bryant, W.A., Aragón Artacho, F.J., Planes, F.J., Stalidzans, E., Maass, A., Vempala, S., Hucka, M., Saunders, M.A., Maranas, C.D., Lewis, N.E., Sauter, T., Palsson, B.Ø., Thiele, I., Fleming, R.M.T., 2019. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. 14, 639–702.
Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., Stevens, R.L., 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat. Biotechnol. 28, 977.
Hu, M., Dinh, H.V., Shen, Y., Suthers, P.F., Foster, C.J., Call, C.M., Ye, X., Pratas, J., Fatma, Z., Zhao, H., Rabinowitz, J.D., Maranas, C.D., 2023. Comparative study of two Saccharomyces cerevisiae strains with kinetic models at genome-scale. Metab. Eng. 76, 1–17.
Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351, aad6253.
Ishchuk, O.P., Domenzain, I., Sanchez, B.J., Muniz-Paredes, F., Martinez, J.L., Nielsen, J., Petranovic, D., 2022. Genome-scale modeling drives 70-fold improvement of intracellular heme production in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S.A. 119, e2108245119.
Jensen, K., Broeken, V., Hansen, A.S.L., Sonnenschein, N., Herrgård, M.J., 2019. OptCouple: joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs. Metab. Eng. 8, e00087.
Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival Jr., B., Assad-Garcia, N., Glass, J.I., Covert, M.W., 2012. A whole-cell computational model predicts phenotype from genotype. Cell 150, 389–401.
Khodayari, A., Maranas, C.D., 2016. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat. Commun. 7, 13806.
Kim, G.B., Kim, J.Y., Lee, J.A., Norsigian, C.J., Palsson, B.O., Lee, S.Y., 2023. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nat. Commun. 14, 7370.
King, Z.A., Lloyd, C.J., Feist, A.M., Palsson, B.O., 2015. Next-generation genome-scale models for metabolic engineering. Curr. Opin. Biotechnol. 35, 23–29.
Klamt, S., Mahadevan, R., von Kamp, A., 2020. Speeding up the core algorithm for the dual calculation of minimal cut sets in large metabolic networks. BMC Bioinf. 21, 510.
Kroll, A., Engqvist, M.K.M., Heckmann, D., Lercher, M.J., 2021. Deep learning allows genome-scale prediction of Michaelis constants from structural features. PLoS Biol. 19, e3001402.
Kroll, A., Ranjan, S., Engqvist, M.K.M., Lercher, M.J., 2023. A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nat. Commun. 14, 2787.
Lao-Martil, D., Schmitz, J.P.J., Teusink, B., van Riel, N.A.W., 2023. Elucidating yeast glycolytic dynamics at steady state growth and glucose pulses through kinetic metabolic modeling. Metab. Eng. 77, 128–142.
Lee, S.M., Lee, G., Kim, H.U., 2022. Machine learning-guided evaluation of extraction and simulation methods for cancer patient-specific metabolic models. Comput. Struct. Biotechnol. J. 20, 3041–3052.
Li, F., Chen, Y., Qi, Q., Wang, Y., Yuan, L., Huang, M., Elsemman, I.E., Feizi, A., Kerkhoven, E.J., Nielsen, J., 2022a. Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints. Nat. Commun. 13, 2969.
Li, F., Yuan, L., Lu, H., Li, G., Chen, Y., Engqvist, M.K.M., Kerkhoven, E.J., Nielsen, J., 2022b. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nat. Catal. 5, 662–672.
Li, G., Hu, Y., Jan, Z., Luo, H., Wang, H., Zelezniak, A., Ji, B., Nielsen, J., 2021. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. Nat. Commun. 12, 190.
Li, Z., Gao, C., Ye, C., Guo, L., Liu, J., Chen, X., Song, W., Wu, J., Liu, L., 2023. Systems engineering of Escherichia coli for high-level shikimate production. Metab. Eng. 75, 1–11.
Lu, H., Kerkhoven, E.J., Nielsen, J., 2021. Multiscale models quantifying yeast physiology: towards a whole-cell model. Trends Biotechnol. 1–15.
Lu, H., Li, F., Sanchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcisauskas, S., Anton, P.M., Lappa, D., Lieven, C., Beber, M.E., Sonnenschein, N., Kerkhoven, E.J., Nielsen, J., 2019. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat. Commun. 10, 3586.
Lu, J., Bi, X., Liu, Y., Lv, X., Li, J., Du, G., Liu, L., 2023. In silico cell factory design driven by comprehensive genome-scale metabolic models: development and challenges. Systems Microbiology and Biomanufacturing 3, 207–222.
Machado, D., Andrejev, S., Tramontano, M., Patil, K.R., 2018. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 46, 7542–7553.
Machado, D., Herrgard, M.J., 2015. Co-evolution of strain design methods based on flux balance and elementary mode analysis. Metab. Eng. 2, 85–92.
Macklin, D.N., Ahn-Horst, T.A., Choi, H., Ruggero, N.A., Carrera, J., Mason, J.C., Sun, G., Agmon, E., DeFelice, M.M., Maayan, I., Lane, K., Spangler, R.K., Gillies, T.E., Paull, M.L., Akhter, S., Bray, S.R., Weaver, D.S., Keseler, I.M., Karp, P.D., Morrison, J.H., Covert, M.W., 2020. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. Science 369, eaav3751.
Marucci, L., Barberis, M., Karr, J., Ray, O., Race, P.R., de Souza Andrade, M., Grierson, C., Hoffmann, S.A., Landon, S., Rech, E., Rees-Garbutt, J., Seabrook, R., Shaw, W., Woods, C., 2020. Computer-aided whole-cell design: taking a holistic approach by integrating synthetic with systems biology. Front. Bioeng. Biotechnol. 8, 942.
McCloskey, D., Palsson, B.O., Feist, A.M., 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol. 9, 661.
Mishra, S., Wang, Z., Volk, M.J., Zhao, H., 2023. Design and application of a kinetic model of lipid metabolism in Saccharomyces cerevisiae. Metab. Eng. 75, 12–18.

Moger-Reischer, R.Z., Glass, J.I., Wise, K.S., Sun, L., Bittencourt, D.M.C., Lehmkuhl, B.K., Schoolmaster, D.R., Lynch, M., Lennon, J.T., 2023. Evolution of a minimal cell. Nature 620, 122–127.
Narayanan, B., Weilandt, D., Masid, M., Miskovic, L., Hatzimanikatis, V., 2024. Rational strain design with minimal phenotype perturbation. Nat. Commun. 15, 723.
Nielsen, J., 2017. Systems biology of metabolism. Annu. Rev. Biochem. 86, 245–275.
Nilsson, A., Nielsen, J., Palsson, B.O., 2017. Metabolic models of protein allocation call for the kinetome. Cell Syst 5, 538–541.
O'Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., Palsson, B.O., 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. 9, 693.
Oftadeh, O., Hatzimanikatis, V., 2024. Genome-scale models of metabolism and expression predict the metabolic burden of recombinant protein expression. Metab. Eng. 84, 109–116.
Österberg, L., Domenzain, I., Münch, J., Nielsen, J., Hohmann, S., Cvijovic, M., 2021. A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism. PLoS Comput. Biol. 17, e1008891.
Patra, P., B R, D., Kundu, P., Das, M., Ghosh, A., 2023. Recent advances in machine learning applications in metabolic engineering. Biotechnol. Adv. 62, 108069.
Pavesi, G., 2017. ChIP-seq data analysis to define transcriptional regulatory networks. Adv. Biochem. Eng. Biotechnol. 160, 1–14.
Pereira, V., Cruz, F., Rocha, M., 2021. MEWpy: a computational strain optimization workbench in Python. Bioinformatics 37, 2494–2496.
Qiu, S., Zhao, S., Yang, A., 2024. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. Briefings Bioinf. 25, bbad506.
Radivojevic, T., Costello, Z., Workman, K., Garcia Martin, H., 2020. A machine learning Automated Recommendation Tool for synthetic biology. Nat. Commun. 11, 4879.
Ranganathan, S., Suthers, P.F., Maranas, C.D., 2010. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput. Biol. 6, e1000744.
Rees-Garbutt, J., Chalkley, O., Grierson, C., Marucci, L., 2021. Minimal genome design algorithms using whole-cell models. In: Marchisio, M.A. (Ed.), Computational Methods in Synthetic Biology. Springer US, New York, NY, pp. 183–198.
Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., Grierson, C., 2020. Designing minimal genomes using whole-cell models. Nat. Commun. 11, 836.
Saa, P.A., Nielsen, L.K., 2017. Formulation, construction and analysis of kinetic models of metabolism: a review of modelling frameworks. Biotechnol. Adv. 35, 981–1003.
Sabzevari, M., Szedmak, S., Penttila, M., Jouhten, P., Rousu, J., 2022. Strain design optimization using reinforcement learning. PLoS Comput. Biol. 18, e1010177.
Salvy, P., Hatzimanikatis, V., 2020. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. Nat. Commun. 11, 30.
Sanchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.J., Kerkhoven, E.J., Nielsen, J., 2017. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol. Syst. Biol. 13, 935.
Schneider, P., Bekiaris, P.S., von Kamp, A., Klamt, S., 2022. StrainDesign: a comprehensive Python package for computational design of metabolic networks. Bioinformatics 38, 4981–4983.
Shen, F., Sun, R., Yao, J., Li, J., Liu, Q., Price, N.D., Liu, C., Wang, Z., 2019. OptRAM: in-silico strain design via integrative regulatory-metabolic network modeling. PLoS Comput. Biol. 15, e1006835.
Si, T., Chao, R., Min, Y., Wu, Y., Ren, W., Zhao, H., 2017. Automated multiplex genome-scale engineering in yeast. Nat. Commun. 8, 15187.
St John, P.C., Strutz, J., Broadbelt, L.J., Tyo, K.E.J., Bomble, Y.J., 2019. Bayesian inference of metabolic kinetics from genome-scale multiomics data. PLoS Comput. Biol. 15, e1007424.
Sveshnikova, A., MohammadiPeyhani, H., Hatzimanikatis, V., 2022. Computational tools and resources for designing new pathways to small molecules. Curr. Opin. Biotechnol. 76, 102722.
Tepper, N., Shlomi, T., 2009. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 26, 536–543.
Thornburg, Z.R., Bianchi, D.M., Brier, T.A., Gilbert, B.R., Earnest, T.M., Melo, M.C.R., Safronova, N., Saenz, J.P., Cook, A.T., Wise, K.S., Hutchison, C.A., Smith, H.O., Glass, J.I., Luthey-Schulten, Z., 2022. Fundamental behaviors emerge from simulations of a living minimal cell. Cell 185, 345–360 e28.
Upadhyay, V., Boorla, V.S., Maranas, C.D., 2023. Rank-ordering of known enzymes as starting points for re-engineering novel substrate activity using a convolutional neural network. Metab. Eng. 78, 171–182.
Volk, M.J., Tran, V.G., Tan, S.I., Mishra, S., Fatma, Z., Boob, A., Li, H., Xue, P., Martin, T. A., Zhao, H., 2023. Metabolic engineering: methodologies and applications. Chem. Rev. 123, 5521–5570.
Wang, H., Marcisauskas, S., Sanchez, B.J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., Kerkhoven, E.J., 2018. Raven 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS Comput. Biol. 14, e1006541.
Wang, J.Y., Doudna, J.A., 2023. CRISPR technology: a decade of genome editing is only the beginning. Science 379, eadd8643.
Wang, T., Xiang, G., He, S., Su, L., Yan, X., Lu, H., 2023. DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D structures. bioRxiv, 2023.12.09.570923.
Wu, G., Yan, Q., Jones, J.A., Tang, Y.J., Fong, S.S., Koffas, M.A.G., 2016. Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications. Trends Biotechnol. 34, 652–664.
Yang, J.H., Wright, S.N., Hamblin, M., McCloskey, D., Alcantar, M.A., Schrubbers, L., Lopatkin, A.J., Satish, S., Nili, A., Palsson, B.O., Walker, G.C., Collins, J.J., 2019. A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell 177, 1649–1661.
Yao, H., Yang, L., 2023. PROSO Toolbox: a unified protein-constrained genome-scale modelling framework for strain designing and optimization. arXiv preprint arXiv: 2308.14869.
Ye, C., Xu, N., Gao, C., Liu, G., Xu, J., Zhang, W., Chen, X., Nielsen, J., Liu, L., 2020. Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C. Biotechnol. Bioeng. 117, 1562–1574.
Yu, H., Deng, H., He, J., Keasling, J.D., Luo, X., 2023a. UniKP: a unified framework for the prediction of enzyme kinetic parameters. Nat. Commun. 14, 8211.
Yu, T., Boob, A.G., Volk, M.J., Liu, X., Cui, H., Zhao, H., 2023b. Machine learning-enabled retrobiosynthesis of molecules. Nat. Catal. 6, 137–151.
Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., Zhao, H., 2023c. Enzyme function prediction using contrastive learning. Science 379, 1358–1363.
Zampieri, G., Vijayakumar, S., Yaneske, E., Angione, C., 2019. Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput. Biol. 15, e1007084.
Zhang, C., Lapkin, A.A., 2023. Reinforcement learning optimization of reaction routes on the basis of large, hybrid organic chemistry–synthetic biological, reaction network data. React. Chem. Eng.
Zhang, J., Petersen, S.D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., Sánchez, B.J., Costello, Z., Chen, Y., Fero, M.J., Martin, H.G., Nielsen, J., Keasling, J. D., Jensen, M.K., 2020. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. Nat. Commun. 11, 4880.
Zheng, S., Zeng, T., Li, C., Chen, B., Coley, C.W., Yang, Y., Wu, R., 2022. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nat. Commun. 13, 3342.
Zimmermann, J., Kaleta, C., Waschina, S., 2021. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biol. 22, 81.