Title

Leveraging knowledge engineering and machine learning for microbial bio-manufacturing

Permalink

https://escholarship.org/uc/item/7v29k7bv

Journal

Biotechnology Advances, 36(4)

ISSN

0734-9750

Authors

Oyetunde, Tolutola

Bao, Forrest Sheng

Chen, Jiung-Wen

et al.

Publication Date

2018-07-01

DOI

10.1016/j.biotechadv.2018.04.008

Peer reviewed

Leveraging knowledge engineering and machine learning for microbial bio- manufacturing

Tolutola Oyetunde[a], Forrest Sheng Bao[b], Jiung-Wen Chen[a], Hector Garcia Martin[c,d,e], Yinjie J. Tang[a,*]

[a] *Department of Energy, Environmental and Chemical Engineering, Washington University in Saint Louis, Saint Louis, MO 63130, USA*
[b] *Department of Computer Science, Iowa State University, Ames, IA 50011, USA*
[c] *DOE, Joint BioEnergy Institute, Emeryville, CA 94608, USA*
[d] *DOE, Agile BioFoundry, Emeryville, CA 94608, USA*
[e] *Biological Systems and Engineering Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA*

ABSTRACT

Genome scale modeling (GSM) predicts the performance of microbial workhorses and helps identify beneficial gene targets. GSM integrated with intracellular flux dynamics, omics, and thermodynamics have shown re- markable progress in both elucidating complex cellular phenomena and computational strain design (CSD). Nonetheless, these models still show high uncertainty due to a poor understanding of innate pathway regula- tions, metabolic burdens, and other factors (such as stress tolerance and metabolite channeling). Besides, the engineered hosts may have genetic mutations or non-genetic variations in bioreactor conditions and thus CSD rarely foresees fermentation rate and titer. Metabolic models play important role in design-build-test-learn cycles for strain improvement, and machine learning (ML) may provide a viable complementary approach for driving strain design and deciphering cellular processes. In order to develop quality ML models, knowledge engineering leverages and standardizes the wealth of information in literature (e.g., genomic/phenomic data, synthetic biology strategies, and bioprocess variables). Data driven frameworks can offer new constraints for mechanistic models to describe cellular regulations, to design pathways, to search gene targets, and to estimate fermentation titer/rate/yield under specified growth conditions (e.g., mixing, nutrients, and O_2). This review highlights the scope of information collections, database constructions, and machine learning techniques (such as deep learning and transfer learning), which may facilitate "Learn and Design" for strain development.

1. Introduction

Although synthetic biology has enabled powerful genome editing, construction of industrial viable hosts is still challenging (Chubukov et al., 2016). Traditional strain designs only look into the biosynthesis pathways followed by push-pull-power-block (3PB) principles (Liu et al., 2017). Although these intuitive approaches resolve obvious bottlenecks in upstream pathways, remove competing reactions, or in- crease cofactor availability, they do not guarantee high productivity. Therefore, modern CSD relies on stoichiometric models (e.g., genome scale models) and reaction thermodynamic information that search broader gene targets and predicts advantageous mutants. Meanwhile, omics data such as gene expressions, proteomics and ^{13}C-metabolic flux analysis are commonly leveraged to obtain insights into multiple levels of biological information (Ishii et al., 2007). More recent efforts have attempted to capture metabolic dynamics and decipher cellular regulatory mechanisms. After resolving parameter estimation and in- terpretability coupled with computational intensity of large dynamic systems, these mechanistic models can significantly improve CSD ap- plications.

Unlike typical models encoding fundamental laws (such as mass and energy balances), data driven algorithms (machine learning, ML) make predictions by deriving patterns from training sets comprising large amounts of experimental data. Since these models are black boxes de- riving predictive capabilities purely from experimental data, simula- tions do not require a complete mechanistic understanding of cell physiologies. Data mining, genome modeling, and big data techniques can leverage complex genetics, fermentation data and omics results for highlighting scenarios (such as different promoter strengths and in- duction characteristics) that may maximally yield metabolic outputs (Chen, 2016; Monk et al., 2016; Utrilla et al., 2016). Moreover, with rapid increase of published metabolic engineering studies and recent

[*] Corresponding author.

E-mail address: yinjie.tang@wustl.edu (Y.J. Tang).

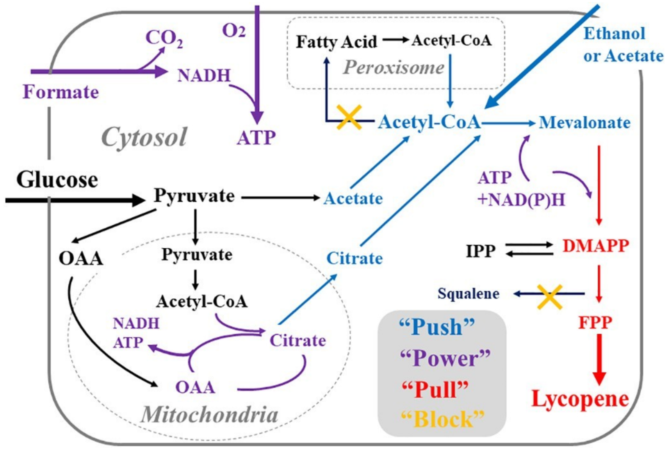


Fig. 1. Pathway-level strain design strategies: Lycopene production as a case study. To improve production in yeast, modifications are required including: (1) "Push" carbon flows towards the acetyl-CoA precursor, in which several acetyl- CoA routes (including acetyl-CoA synthase and citrate lyase reactions). (2) "Pull" carbon flow towards lycopene (i.e., overexpress mevalonate pathways).

(3) "Block" fluxes competing for mevalonate pathways (e.g., lipid synthesis);

(4) "Power" cell metabolism by engineering redox cofactor balances and pro- moting ATP production (i.e., increase oxidative phosphorylation).

advances in artificial intelligence research, the use of data driven ap- proaches may facilitate the understanding of cellular processes and assist mechanistic modeling for quality CSD.

2. Advances and limitations in computational strain design

Strain design requires an understanding of cellular metabolism and regulation, followed by identifications of genetic strategies to repro- gram cell metabolism for useful ends. In the past, strain improvement was achieved via random mutations or overexpression of a single bio- synthesis gene. With the advance of genome sequencing and synthetic biology, targeted modifications of multiple genes or pathways have become commonplace to redirect carbon flows to desired products (Lee and Kim, 2015; Parekh et al., 2000). 3PB principles have been in- tuitively used to manipulate cell performance. For example (Fig. 1), common strategies to optimize a yeast strain for the de novo production of lycopene include PUSH (increase the supply of the precursor cyto- plasmic acetyl-CoA), PULL (improve enzyme activities for lycopene synthesis), POWER (enhance ATP and NAD(P)H generation), and BLOCK (inhibit competing pathways). 3PB are not always effective because they may induce new bottlenecks after each genetic mod- ification, and thus design-build-test-learning cycle has to be performed. For an engineered host, many upstream pathways may place rate- limiting steps. Thus, genome-wide modifications must be performed after creation of proof-of-concept strains. In this context, GSM is useful to predict mutant physiologies, in which cell processes are inherently constrained by steady-state mass balances (i.e., flux balance analysis, FBA) (Orth et al., 2010). Such underdetermined systems are solved using objective functions (Schuetz et al., 2007). For example, biomass growth objective has shown decent accuracy to describe cultures in carbon limited conditions (Fong and Palsson, 2004). Marriage between GSM and optimization frameworks (e.g., OptForce and OptKnock) can search gene targets throughout metabolic networks (Burgard et al., 2003; Burgard et al., 2004; Pharkya and Maranas, 2006; Ranganathan et al., 2010). Typical CSD has the following layers: (1) construct genome scale models for the host, (2) search appropriate enzyme/ pathway targets, and (3) predict the behavior of the cell under given genetic and growth conditions. CSD aids engineers with a quantitative analysis of metabolic network and suggests a starting-point or a prior- itized intervention. It is particularly useful for combinatorial gene manipulations when the host production is limited by both carbon re- courses and energy molecules (ATP and NADPH). In a case study, a CiED (cipher of evolutionary design) framework successfully identified non-intuitive genetic perturbations that resulted in optimal phenotypes for the production of flavanone (Fowler et al., 2009). Moreover, Opt-Force algorithm has been successfully used to identify multiple gene targets in the pentose phosphate pathway for photosynthetic produc- tion of isoprenoids by cyanobacteria (Lin et al., 2017).

CSD paradigms with flux optimization theory still show high un- certainty due to cellular regulations and nonlinear metabolic responses to multiple genetic modifications. Besides, a typical GSM does not in- clude kinetic parameters, regulatory factors (such as transcriptional factor), or non-enzyme factors (e.g., product tolerance, cell stresses, and genetic stability). Particularly, stoichiometry-only CSD procedures cannot capture the effect of metabolite concentrations and their feed- back inhibitions. Current CSD still has poor capability to predict out- come of pathway overexpression or identify rate-limiting enzymes. New modeling tools leverage omics, kinetic, and thermodynamic informa- tion to improve both metabolic insights and CSD applications (Long et al., 2015). First, COBRA combined with transcriptomics and pro- teomics data has shown successes to predict strain performance based on the relationship between gene/protein profiles and the fluxome (e.g. tFBA (van Berlo et al., 2011), GIMME (Becker and Palsson, 2008), iMAT (Zur et al., 2010), ME-Models (O'Brien et al., 2013), and E-FLUX (Colijn et al., 2009)). Using gene data from high throughput sequencing tech- nique, GSM can not only narrow flux intervals, but also can identify genes that likely regulate microbial fluxes (Machado and Herrgård, 2014). Second, thermodynamics has been used to reduce modeling bias/uncertainty, including thermodynamics-based metabolic balance analysis (Henry et al., 2007), network-embedded thermodynamic ana- lysis (Kümmel et al., 2006), and energy balance analysis (Beard et al., 2002). Third, combining GSM with kinetic modeling has elucidated useful insights into cellular dynamics. For example, a genome-scale *E. coli* model k-ecoli457 (containing 457 model reactions) can describe dynamic flux data for multiple mutant strains as well as substrate-level regulatory interactions (Khodayari and Maranas, 2016). Similarly, k- OptForce software enables identification of a minimal set of interven- tions, comprised of both enzymatic kinetics and reaction flux changes, to achieve the overproduction of the target product. Such an approach can find non-intuitive interventions aiming at alleviating metabolite inhibition of key enzymes (Chowdhury et al., 2014).

The CSD still faces hurdles: cellular productivity is always much lower than model predictions. In general, increasing flux through a reaction is much more complicated to achieve than decreasing or eliminating it. Besides transcriptional regulations, several hidden fac- tors play roles in metabolic re-programming that is difficult to be in- cluded in GSM (Fig. 2). First, enzyme expressions and product synthesis consume cellular NAD(P)H, ATP and building blocks. Balancing be- tween carbon and energy fluxes can be complex if biosynthesis requires ATP and NAD(P)H. To reveal metabolic burdens from engineered pathways (Ceroni et al., 2015), both experiments and models have to be used to quantify the tradeoffs from synthetic biology components. Second, bio-productions can pose stresses or membrane damages on the host, which can further increase the ATP maintenance loss or other cellular stress responses (Hoehler and Jørgensen, 2013). Third, pathway engineering may induce metabolite changes within central metabolism, which drive flux adaptions at substrate level (such as Mi- chaelis-Menten kinetics, allosteric or feedback regulations) (Gerosa et al., 2015; Tummler et al., 2014). For example, cellular ATP/ADP/ AMP ratio strongly affects sugar catabolic rates. Fourth, the engineered hosts can be unstable in bioreactor conditions due to genetic mutations and non-genetic cell-to-cell variations. Cell behavior or genetic stability is closely related to nutrient supplies, bioreactor modes, and fermen- tation duration, or other process factors, which are often ignored in a modeling framework. Fifth, synthetic genetic components may have varied outcomes after introducing into a host. Even the order of genes

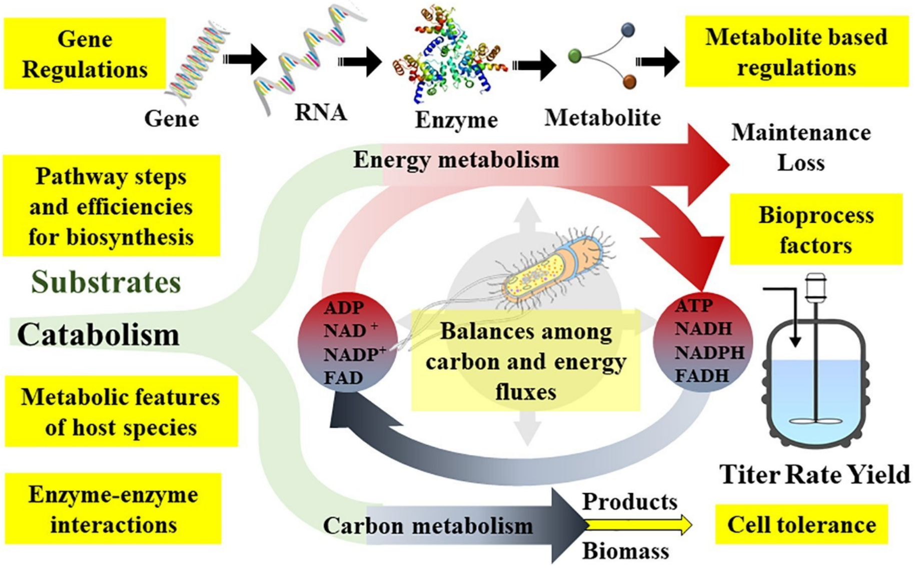
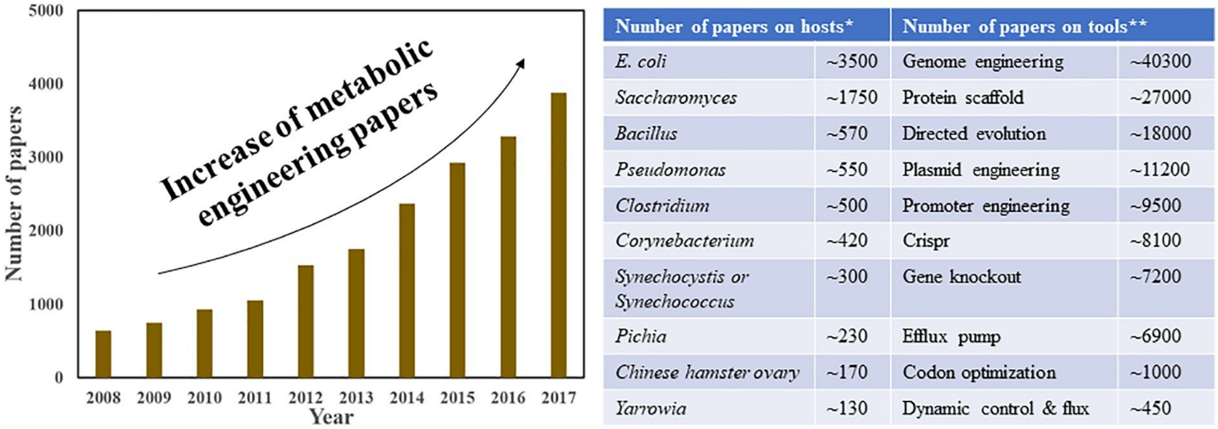


Fig. 2. Basic schematic of microbial metabolism and strain design showing the interplay between carbon and energy processes subject to regulation/influential factors (highlighted in yellow boxes). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Number of papers on hosts*		Number of papers on tools**	
E. coli	~3500	Genome engineering	~40300
Saccharomyces	~1750	Protein scaffold	~27000
Bacillus	~570	Directed evolution	~18000
Pseudomonas	~550	Plasmid engineering	~11200
Clostridium	~500	Promoter engineering	~9500
Corynebacterium	~420	Crispr	~8100
Synechocystis or Synechococcus	~300	Gene knockout	~7200
Pichia	~230	Efflux pump	~6900
Chinese hamster ovary	~170	Codon optimization	~1000
Yarrowia	~130	Dynamic control & flux	~450

Fig. 3. Rapid increase of metabolic engineering data (information was based on PubMed key words search on Jan 25, 2018).

in a pathway may change productivity of a heterologous pathway due to unknown expression balance of cascade enzymes (Nishizaki et al., 2007). Finally, innate enzyme channels (i.e., co-localize cascade enzymes to shuttle metabolites) is another hidden factor that may confound pathway regulations (Poshyvailo et al., 2017). In summary, metabolic pathways are orchestrated by overlapping regulatory mechanisms, affecting thousands of molecular components. Modeling of genetic inputs and their nontrivial interactions/tradeoffs, as a whole, presents a formidable challenge.

3. Machine learning and knowledge engineering

Currently, genomics data at different cellular levels are still insufficient to determine holistic metabolic regulations (Kochanowski et al., 2013). Prediction of fluxes based on metabolite concentrations or enzyme abundance is still inaccessible for the majority of metabolic reactions (Gerosa et al., 2015; Hackett et al., 2016). Due to these limitations, data driven approaches may be used in conjunction to mechanistic models to simulate complex cellular behavior by transforming both accountable and unaccountable influential variables (Fig. 2). Machine learning (ML) is a branch of artificial intelligence that train computers to perform tasks by gaining the capability from 'experience' (data) rather than being specifically programmed to do so. ML studies are broadly classified into supervised, unsupervised and reinforced learning. In supervised learning, the computer develops an input-output model from sets of inputs and ‘correct’ (i.e., labeled) outputs. In un- supervised learning (e.g., cluster analysis), hidden patterns and struc- tures can be uncovered from the data. ML has many applications such as finance, personalized medicine, cancer diagnosis, computer vision, and energy forecasting (Jordan and Mitchell, 2015; Libbrecht and Noble, 2015). ML techniques have gained widespread use in compu- tational biology (Razavian, 2004; Sommer and Gerlich, 2013; Tarca et al., 2007). Traditional applications include analysis of gene and protein networks. More recent applications have sought to guide design of microbial cell factories. For example, a PCA-based (Principal Com- ponent Analysis) framework improved yield using fermentation, gene expression and proteomics datasets (Alonso-Gutierrez et al., 2015). Other studies have elucidated useful insights into cellular metabolism by combining ML with metabolic modeling. For example, knowledge about kinetic parameters or metabolite concentrations involves high uncertainty. Through the use of ML principles (e.g., support vector machine), researchers can obtain a narrow range of kinetic parameters or metabolite concentrations (Andreozzi et al., 2016; Yang et al., 2017).

Rapid growth of synthetic biology in the past decade has generated a large amount of literature and experimental databases (Fig. 3). However, every case study uses different conditions and the number of variables is very large. In the ML field, better organized data always trump better algorithms. Thus, it is necessary to standardize the data- sets and build databases by extracting and clustering published data (i.e., Knowledge Engineering) (Sowa, 2000; Studer et al., 1998). To date, there are many databases that focus on documenting known knowledge about cellular networks (genomic, transcriptomic, meta- bolic, and regulatory networks) and the interactions between them (Jing et al., 2014). These include KEGG (Kanehisa, 2002; Kanehisa et al., 2016), BiGG (King et al., 2016), Rhea (Alcántara et al., 2011), CecaFDB (Zhang et al., 2014), MetaCyc (Caspi et al., 2016), and BioCyc (Karp et al., 2017). While such databases can potentially provide con- siderable insight into cellular metabolism and its regulation, they have limitations since they do not contain information about performance of engineered strains (yield, titer, and production rate) nor parameters related to bioprocess conditions (such as reactor configurations and growth medium). Recently, a number of efforts have focused on cur- ating experimental metabolic information from published literature (Winkler et al., 2015; Wu et al., 2016). Frameworks like Experimental Data Depot (EDD) (Morrell et al., 2017), LASER (Winkler et al., 2015), and OMERO (Allan et al., 2012) have been developed to standardize documentation and integration of biological experimental information. Frameworks for specific microorganisms have also been developed (Maarleveld et al., 2014). These frameworks also enable basic data visualization as well as a suite of tools for data manipulation/analyses. Other frameworks like KBase (Arkin et al., 2016) focus on integrating not only data but computational methods for enhanced predictive fi- delity of biological functions.

Knowledge databases will benefit data standardization and pave the way for artificial intelligence to boost CSD and automation of strain development. Detailed information (including fermentation process variables, omics data, genetic tools or components) is valuable for ML to make predictions. Frameworks like LASER and EDD provide tem- plates for such information to be gathered and standardized. Typical mechanistic models need to simplify complex biological systems, while ML can estimate strain physiological responses under diverse biopro- cess (such as nutrients and bio-reactor modes) and genetic factors (e.g., metabolic burdens from gene overexpression or other synthetic biology parts) without understanding cellular processes. Particularly, the deep learning (DL), a recent powerful class of ML techniques, capable of handling massive datasets and mining complicated patterns hidden in data, will prove useful towards this end (Angermueller et al., 2016). Nonetheless, DL algorithms require much larger amounts of quality data than traditional ML approaches, which can be practical only after significant progresses in knowledge engineering.

4. Paradigms of machine learning techniques in bio- manufacturing

Both bioprocessing and systems biology have widely employed ML, which can play an important role in design-build-test-learn cycle for strain improvement and fermentation optimizations (Fig. 4). Table 1 gives published ML applications to predict metabolic outcomes. Most of these applications follow a similar workflow: (1) identification of output variables (like yield, titer, or rate); (2) iterative feature selection to identify input factors that are most influential on performance me- trics; (3) model selection depending on data availability; and (4) model training and validation. Data driven model provide complementary information to GSM. The later focuses on predicting biosynthesis yields, while production rates and titers are determined by the synergistic impact of product yields, bioprocesses, strain tolerance, and biomass growth. ML could take into account the genetic design of the microbial host system and the “suboptimal” conditions under which the fermen- tation process occurs. The hybrid of ML-GSM may identify effective metabolic strategies or targets and qualitatively benchmark various performances of engineered production platforms.

Fig. 4 shows possible paradigms for utilizing data-driven techniques in systems metabolic engineering. The earlier applications of ML in fermentation processes usually involved data from bioprocess studies. These studies aim to link influential factors (e.g., bioreactor conditions) to cell productivity via linear/nonlinear regressions or neural network (paradigm #1). Most of the applications listed in Table 1 are of this kind. The advantage of this scheme is that the data formats of inputs/ outputs are relatively simple (usually from one set of study). Because the dataset size is usually small, model scope is fairly limited. Another type of efforts has sought to decode complexity in cellular networks by using omics dataset as well as details of synthetic biology constructs (paradigm #2). These frameworks learn system behaviors at different regulation layers and decode key genes that control desired cellular functions, which enable design-build-test-learn cycle during strain im- provements (Gill et al., 2016). They can also improve the fidelity of metabolic network reconstructions used for genome scale modeling (Oyetunde et al., 2016). A limitation of such frameworks is that they do not usually consider the bioprocess conditions or engineering strategies. Researchers may potentially combine the benefits of the first two paradigms. Via knowledge engineering to generate a database that contains structured input (species, nutrient, culture conditions, genetic tools, strain tolerance and stability) and outputs (yield/rate/titer), ML can capture microbial physiologies in response to various genetic and fermentation conditions. For example, ML models were developed for a priori estimation of chemical productivity from engineered *E. coli* and *S. cerevisiae*, given a set of model inputs (e.g., biosynthesis steps, nutrient supplementation, bioreactor modes) (Colletti et al., 2011; Varman et al., 2011). Such models via linear regressions correctly predict that the product synthesis using long pathways unavoidably gives poor production yield and titer. These models are useful for manufacturers to decide whether a product should be produced via engineered microbial cell factories or via a chemical synthesis route.

The advantages of GSM/FBA over ML lie in their interpretable and biologically meaningful solutions. On the contrary, ML models rely purely on statistics, thus may generate predictions that violate some biological constraints or lie out of reasonable ranges. In this regard, ML models are expected to gain great improvement when combined with GSM/FBA models. GSMs can help identify whether ML outcomes are biologically feasible, within biological reasonable ranges, or directly place upper bounds for ML outcomes. ML, FBA algorithm and constraint logic programming can be integrated to offer an expressive way to re- present knowledge that involves statistics, constraints (usually on in- tegers or real numbers) and logics (paradigm #3). Such hybrid models take into account the metabolic network, genetic design of the micro- bial host system, and the “suboptimal” conditions under which the fermentation process occurs. For example, supervised learning methods and FBA have been used together to predict bacterial central metabo- lism (Wu et al., 2016). In that study, experimental data of 37 bacteria species from over 100 ^{13}C-MFA papers were extracted and converted into structured data. Three supervised algorithms, Support Vector Ma- chine (SVM), k-Nearest Neighbors (kNN), and Decision Tree were em- ployed to train regressors to predict fluxes using features (substrate types, genetic modifications, and cultivation methods). The ML can generate reasonable flux boundaries for FBA models to reduce solution space during flux predictions of nonmodel microbial species. In sum- mary, paradigm #3 binds the ML predictions with the GSM optimiza- tions, which can not only predict production metrics (like yield, titer and rate) but also can suggest optimal genetic engineering strategies to employ (like what kind of plasmid to use, promoter strength, etc.) during design-build-test-learn cycle.

Finally, metabolic engineering is a rapid-developing field. The new high-throughput technologies can quickly generate large amount of data, such as high throughput mass spectrometry (Fuhrer and Zamboni, 2015) and microfluidics (Heinemann et al., 2017a, 2017b). These data allow extensive validation of ML platforms and parameter estimations. Even those failed experimental data are valuable for training ML. For example, combinatorial synthesis and screening approaches create vast

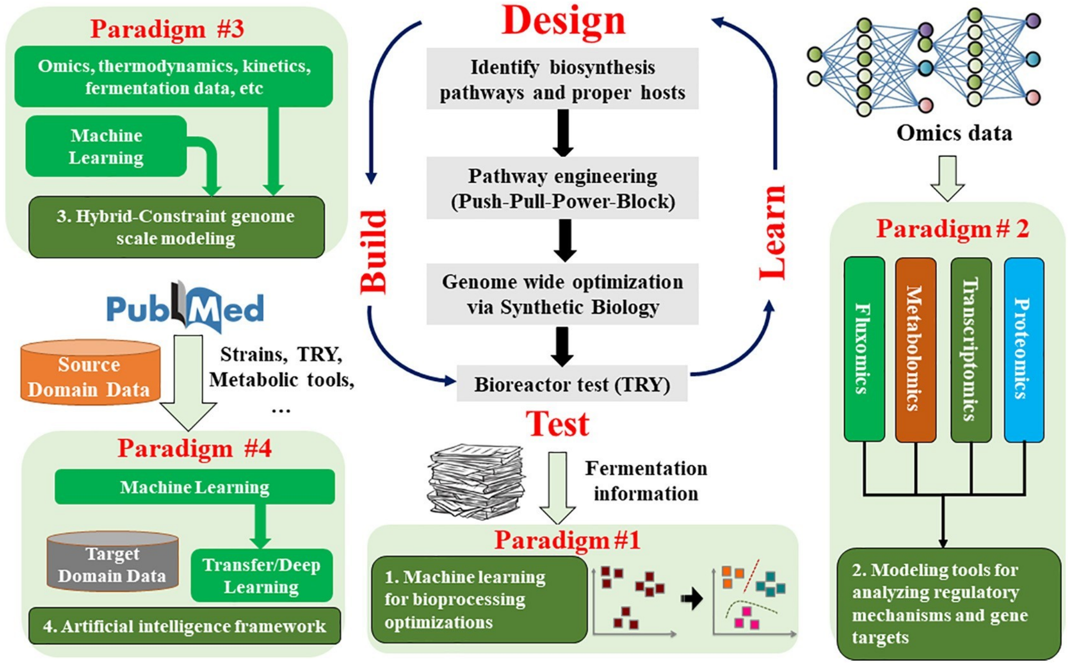

Paradigm #3
Omics, thermodynamics, kinetics, fermentation data, etc
Machine Learning
3. Hybrid-Constraint genome scale modeling
PubMed
Source Domain Data
Strains, TRY, Metabolic tools, …
Paradigm #4
Machine Learning
Target Domain Data
Transfer/Deep Learning
4. Artificial intelligence framework
Design
Identify biosynthesis pathways and proper hosts
Pathway engineering (Push-Pull-Power-Block)
Genome wide optimization via Synthetic Biology
Bioreactor test (TRY)
Build
Learn
Test
Fermentation information
Paradigm #1
1. Machine learning for bioprocessing optimizations
Omics data
Paradigm # 2
Fluxomics
Metabolomics
Transcriptomics
Proteomics
2. Modeling tools for analyzing regulatory mechanisms and gene targets

Fig. 4. Paradigms of data-driven techniques in systems metabolic engineering.

numbers of off-target phenotypes that can be used to study engineered metabolism by supervised learning. On the other hand, many input/ output variables are not continuous or complete among different datasets. Advanced Deep Learning (DL) can investigate noisy but large biological data (Chicco et al., 2014; Leung et al., 2014). Due to its nonlinear mapping power, DL can unify incomplete inputs/outputs. Small dataset sizes (which are usually the case for metabolic engineering data) can be tackled by strategies such as unsupervised pretraining (Bengio et al., 2013). During the learning process, noisy and incomplete data will be automatically “flattened” in their new representation space. Furthermore, DL can solve one system and apply the knowledge gained to a different but related new system (Dai et al., 2007; Raina et al., 2007), which may offer systems design or a priori estimation of broad-scope microbial factories. Subsequently, advanced mechanistic models, knowledge engineering, and machine learning lead to ever-improving artificial intelligence framework that relies less and less on the intuition of human engineers (paradigm #4).

5. Hindrances to successful application of machine learning techniques

Despite the promise of ML for synthetic biology and metabolic engineering, several hurdles still need to be tackled. A key challenge for applying ML is the lack of formatted, high-quality, and high quantity data. For example, DL will need ~10,000 conditions to be effective. Large research groups are devoting increasingly time and manpower to establish and standardize systems biology database that will facilitate the validation and improvement of ML frameworks in the near future (Arkin et al., 2016). However, most existing publications contain data with no unified format and these datasets have to be manually curated from non-standardized reports. It is quite challenging to extract the information from a large amount of publications, because the data could be noisy and each paper contains large amounts of variables. Errors can arise from the original authors of the paper or researchers attempting to extract the information. This opens up the need of automatic and semi-automatic tools for collecting experimental data from literature. Natural language processing (NLP) may enable the automatic extraction of relevant data from thousands of publications, which can perform text summarization, evaluate paper quality, and minimize the impact and occurrence of human errors. On the other hand, transfer learning is a ML technique which alleviates the data insufficiency problem by transferring knowledge in one domain (typically with lots of data) to another domain where data are scarce (LeCun et al., 2015) (paradigm #4). For example, data and models on *E. coli* are relatively abundant. This knowledge can be transferred to the non-model microbial platforms, which have few available data by well-tuned transfer learning algorithms. Such practices will not only facilitate the specific task of microbial prediction, but also build a unified viewpoint of re- presentation learning and domain adaptation through the study on practical biological data (Pan and Yang, 2010).

Another major concern is the fact ML models do not generalize well to data points representing conditions not present in the training data. For instance, the training datasets are enormous to identify gene targets for engineering a new host while optimizing bioreactor conditions for typical fermentations requires far less data. This challenge underscores the importance of creating hybrid data-driven and mechanistic models. The success of such hybrid frameworks has been demonstrated in recent efforts (Kogadeeva and Zamboni, 2016; Khodayari and Maranas, 2016; Wu et al., 2016). One study showed the possibility of using data-driven approaches to guide future developments of mechanistic-based models (King et al., 2017). Furthermore, there has been a rapid increase in metabolic engineering data, while the influential factors (e.g., genetic tools, basic microbial pathways and hosts) have remained fairly limited. Specifically, the variability of key upstream pathways towards bio- synthesis is unchanged (Fig. 5), and most bio-manufacturing comes from a few precursors (such as PEP, acetyl-CoA and pyruvate). Proper feature extraction from existing metabolic engineering data might re- sult in rather robust coverage of possible conditions. Therefore, the number of model parameters may not increase as the size of the training database grows, which ensures the predictive fidelity of the ML plat- form.

In conclusion, metabolic engineering field has accumulated large set of data that are beyond the capability of traditional data analytics. ML presents a frontier research to gain new understanding of microbial metabolism, to improve the reproducibility of experimental work, to enable the rapid design of efficient and robust strains, and to inform commercial decisions. By combining traditional mechanistic models with knowledge engineering and data-driven techniques, a new strain

Table 1
Application of data-driven techniques in metabolic engineering fields.

ML technique	Application	Comment	Reference
Neural networks experimental	Improve the yield of target protein	Used NN technique to build predictive model from results and stochastic sampling. Discovered experimental conditions that give ~350% improvement of yield	Caschera et al. (2011)
Naïve Bayes, kNN, decision trees, logistic regression	Metabolic pathway prediction	The ML methods performed as well as the well-designed and refined algorithm (PathoLogic). Besides, ML methods have the advantage of easily adding new features to test and further optimize the performance.	Dale et al. (2010)
Multiple kernel learning, transfer learning	Predicting protein interactions in fungal secretion pathways	They predicted the protein-protein interaction in the cross-species *T. reesei* by the learning features obtained from *S. cerevisiae*.	Kludas et al. (2016)
Transfer learning substrate	Predict the matrix metalloprotease (MMPs) cleavage sites	They learn the knowledge from the source domain (MMP-9 and MMP-12) to improve the prediction of cleavage sites of other MMPs (MMP-2, -3, -7, and -8) in the target domain.	Wang et al. (2017)
Neural networks condition	Use NN to investigate the effect of process (e.g. time, temperature, pH, etc.) on xylitol production	In this study, a multilayer perceptron (MLP) based feed forward neural network model with Levenberg-Marquardt back propagation (BP-MLP) algorithm was trained with 339 experimental data points. The model could predict the optimal harvest time in xylitol production.	Pappu and Gummadi (2016)
Neural networks cyclodextrin	Optimize the fermentation process of glycosyltransferase production.	They first found the key influential factors using Plackett-Burman Design (PBD) and then optimized by NN. The NN contains one hidden layer.	Amiri et al. (2015)
Neural networks rapamycin	Optimization of fermentation parameters of production by *Streptomyces hygroscopicus* NRRL 5491	The authors applied Plackett–Burman design (PBD) method, artificial neural networks (ANN), and genetic algorithms (GA). The ANN was used to further optimize the key factors found in PBD method.	Sinha et al. (2014)
SVM, Neural networks fermentation	Predict the yield of glutamic acid from process parameters (pH, temperature, carbon source concentration, aeration)	They choose SVM method because it is suitable for small datasets (which is usually the case for production data). They also determined that SVM was more accurate in predicting yield than NN.	Wang et al. (2016)
Gaussian process model, SVM	Estimate the probability of a given enzyme to catalyze a given reaction	The authors created a semi-supervised Gaussian model to predict if a given enzyme is able to catalyze the desired reaction. Furthermore, the Michaelis constant was also predicted by Gaussian progress regression to quantify the affinity between enzyme and the reaction. The results shows the ML can be a powerful tool to speed up the application of synthetic biology.	Mellor et al. (2016)
Decision tree	Develop a data-driven model to accurately design CRISPR-based transcription regulator.	The authors used pairwise datasets of guideRNAs and gene expression to build a predictive model	Sheng et al. (2017)
SVM feature	Predict the essential genes in *E. coli* metabolism	The authors proposed a strategy of data curation and selection to improve the performance of SVM model. Instead of performing flux balance analysis, which are condition specific, to obtain flux features, they applied flux coupling analysis to get the higher sensitivity and specificity of the model.	Nandi et al. (2017)
PCA	Identify specific enzymes that limiting the production of target molecules in a pathway	Based on the PCA distribution, they manipulated the gene expression level of mevalonate pathway enzymes in *E. coli* to improve the production of limonene up to 40%.	Alonso-Gutierrez et al. (2015)

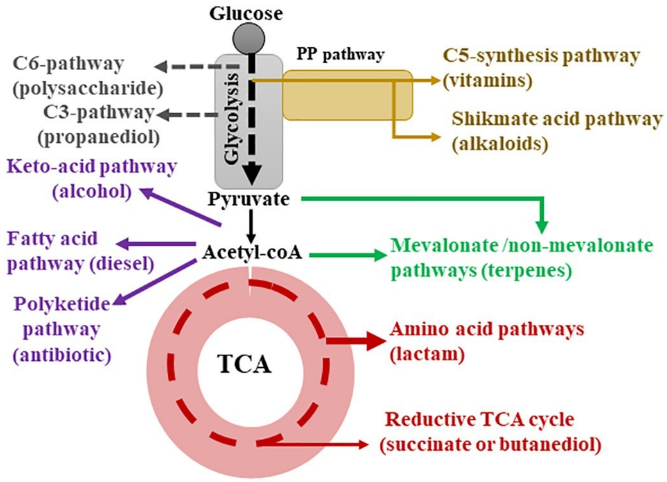


Fig. 5. Common biosynthesis pathways from the central metabolic network.

design framework will ultimately automate synthetic biology and bio- manufacturing.

Acknowledgements

This work is supported by NSF (MCB 1616619) and DOE (DESC0018324). This work was also part of the Joint BioEnergy Institute (http://www.jbei.org) and part of the Agile BioFoundry (http://agilebiofoundry.org) supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, through contract DE-AC02-05CH11231.

References

Alcántara, R., Axelsen, K.B., Morgat, A., Belda, E., Coudert, E., Bridge, A., Cao, H., De Matos, P., Ennis, M., Turner, S., 2011. Rhea—a manually curated resource of bio- chemical reactions. Nucleic Acids Res. 40, D754–D760.

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W.J., Neves, C., Patterson, A., Porter, M., Tarkowska, A., Loranger, B., Avondo, J., Lagerstedt, I., Lianas, L., Leo, S., Hands, K., Hay, R.T., Patwardhan, A., Best, C., Kleywegt, G.J., Zanetti, G., Swedlow, J.R., 2012. OMERO: flexible, model- driven data management for experimental biology. Nat. Methods 9, 245–253.

Alonso-Gutierrez, J., Kim, E.-M., Batth, T.S., Cho, N., Hu, Q., Chan, L.J.G., Petzold, C.J., Hillson, N.J., Adams, P.D., Keasling, J.D., Garcia Martin, H., Lee, T.S., 2015. Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. Metab. Eng. 28, 123–133.

Amiri, A., Mohamad, R., Rahim, R.A., Illias, R.M., Namvar, F., Tan, J.S., Abbasiliasi, S.,

2015. Cyclodextrin glycosyltransferase biosynthesis improvement by recombinant Lactococcus lactis NZ: NSP: CGT: medium formulation and culture condition opti- mization. Biotechnol. Biotechnol. Equip. 29, 555–563.

Andreozzi, S., Miskovic, L., Hatzimanikatis, V., 2016. iSCHRUNK–in silico approach to characterization and reduction of uncertainty in the kinetic models of genome-scale
metabolic networks. Metab. Eng. 33, 158–168.

Angermueller, C., Pärnamaa, T., Parts, L., Oliver, S., 2016. Deep learning for computa- tional biology. Mol. Syst. Biol. 878.

Arkin, A.P., Stevens, R.L., Cottingham, R.W., Maslov, S., Henry, C.S., Dehal, P., Ware, D.,
Perez, F., Harris, N.L., Canon, S., 2016. The DOE Systems Biology Knowledgebase (KBase). bioRxiv, pp. 96354.

Beard, D.A., Liang, S.D., Qian, H., 2002. Energy balance for analysis of complex metabolic networks. Biophys. J. 83 (1), 79–86.

Becker, S.a., Palsson, B.O., 2008. Context-specific metabolic networks are consistent with experiments. PLoS Comput. Biol. 4.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.

van Berlo, R.J.P., de Ridder, D., Daran, J.-M., Daran-Lapujade, P.A.S., Teusink, B.,
Reinders, M.J.T., 2011. Predicting metabolic fluxes using gene expression differences as constraints. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 206–216.

Burgard, A.P., Pharkya, P., Maranas, C.D., 2003. Optknock: a bilevel programming fra- mework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84 (6), 647–657.

Burgard, A.P., Nikolaev, E.V., Schilling, C.H., Maranas, C.D., 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. 14 (2), 301–312. Caschera, F., Bedau, M.A., Buchanan, A., Cawse, J., de Lucrezia, D., Gazzola, G., Hanczyc,
M.M., Packard, N.H., 2011. Coping with complexity: machine learning optimization of cell-free protein synthesis. Biotechnol. Bioeng. 108, 2218–2228.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D., 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44.

Ceroni, F., Algar, R., Stan, G.B., Ellis, T., 2015. Quantifying cellular capacity identifies gene expression designs with reduced burden. Nat. Methods 12 (5), 415–418.

Chen, G.Q., 2016. Omics meets metabolic pathway engineering. Cell Syst. 2, 362–363. Chicco, D., Sadowski, P., Baldi, P., 2014. Deep autoencoder neural networks for gene ontology annotation predictions. In: Proceedings of the 5th ACM Conference on
Bioinformatics, Computational Biology, and Health Informatics. ACM, pp. 533–540.

Chowdhury, A., Zomorrodi, A.R., Maranas, C.D., 2014. k-OptForce: integrating kinetics with flux balance analysis for strain design. PLoS Comput. Biol. 10.

Chubukov, V., Mukhopadhyay, A., Petzold, C.J., Keasling, J.D., Martín, H.G., 2016.
Synthetic and systems biology for microbial production of commodity chemicals. NPJ Syst. Biol. Appl. 2, 16009.

Colijn, C., Brandes, A., Zucker, J., Lun, D.S., Weiner, B., Farhat, M.R., Cheng, T.-Y., Moody, D.B., Murray, M., Galagan, J.E., 2009. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid pro- duction. PLoS Comput. Biol. 5, e1000489.

Colletti, P.F., Goyal, Y., Varman, A.M., Feng, X., Wu, B., Tang, Y.J., 2011. Evaluating factors that influence microbial synthesis yields by linear regression with numerical and ordinal variables. Biotechnol. Bioeng. 108, 893–901.

Dai, W., Yang, Q., Xue, G.-R., Yu, Y., 2007. Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning. ACM, pp. 193–200. Dale, J.M., Popescu, L., Karp, P.D., 2010. Machine learning methods for metabolic pathway prediction. BMC Bioinforma. 11, 15.

Fong, S.S., Palsson, B.Ø., 2004. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. Nat. Genet. 36, 1056–1058.

Fowler, Z.L., Gikandi, W.W., Koffas, M.A.G., 2009. Increased malonyl coenzyme A bio- synthesis by tuning the *Escherichia coli* metabolic network and its application to flavanone production. Appl. Environ. Microbiol. 75, 5831–5839.

Fuhrer, T., Zamboni, N., 2015. High-throughput discovery metabolomics. Curr. Opin.
Biotechnol. 31, 73–78.

Gerosa, L., Haverkorn Van Rijsewijk, B.R.B., Christodoulou, D., Kochanowski, K., Schmidt, T.S.B., Noor, E., Sauer, U., 2015. Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. Cell Syst. 1, 270– 282.

Gill, R.T., Halweg-Edwards, A.L., Clauset, A., Way, S.F., 2016. Synthesis aided design: the biological design-build-test engineering paradigm? Biotechnol. Bioeng. 113, 7–10.

Hackett, S.R., Zanotelli, V.R.T., Xu, W., Goya, J., Park, J.O., Perlman, D.H., Gibney, P.A., Botstein, D., Storey, J.D., Rabinowitz, J.D., 2016. Systems-level analysis of me- chanisms regulating yeast metabolic flux. Science 354 (6311). http://dx.doi.org/10. 1126/science.aaf2786.

Heinemann, J., Deng, K., Shih, S.C.C., Gao, J., Adams, P.D., Singh, A.K., Northen, T.R., 2017a. On-chip integration of droplet microfluidics and nanostructure-initiator mass
spectrometry for enzyme screening. Lab Chip 17, 323–331.

Heinemann, J., Noon, B., Willems, D., Budeski, K., Bothner, B., 2017b. Analysis of raw biofluids by mass spectrometry using microfluidic diffusion-based separation. Anal. Methods 9, 385–392.

Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2007. Thermodynamics-based metabolic
flux analysis. Biophys. J. 92 (5), 1792–1805.

Hoehler, T.M., Jørgensen, B.B., 2013. Microbial life under extreme energy limitation. Nat.
Rev. Microbiol. 11, 83–94.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M.,
Hirai, K., Hoque, A., Ho, P.Y., 2007. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 316 (5824), 593–597.

Jing, L.S., Shah, F.F.M., Mohamad, M.S., Hamran, N.L., Salleh, A.H.M., Deris, S., Alashwal, H., 2014. Database and tools for metabolic network analysis. Biotechnol. Bioprocess Eng. 19, 568–585.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245).

Kanehisa, M., 2002. The KEGG database. In: 'In silico' Simulation of Biological Processes. 247. pp. 91–103.

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

Karp, P.D., Billington, R., Caspi, R., Fulcher, C.A., Latendresse, M., Kothari, A., Keseler, I.M., Krummenacker, M., Midford, P.E., Ong, Q., 2017. The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform. 1–9. http://dx.doi. org/10.1093/bib/bbx085.

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.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., Lewis, N.E., 2016. BiGG Models: a platform for integrating, standar- dizing and sharing genome-scale models. Nucleic Acids Res. 44, D515–D522.

King, Z.A., O'Brien, E.J., Feist, A.M., Palsson, B.O., 2017. Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion. Metab. Eng. 39, 220–227.

Kludas, J., Arvas, M., Castillo, S., Pakula, T., Oja, M., Brouard, C., Jäntti, J., Penttilä, M., Rousu, J., 2016. Machine learning of protein interactions in fungal secretory path- ways. PLoS One 11, 1–20.

Kochanowski, K., Sauer, U., Chubukov, V., 2013. Somewhat in control-the role of tran- scription in regulating microbial metabolic fluxes. Curr. Opin. Biotechnol. 24 (6), 987–993.

Kogadeeva, M., Zamboni, N., 2016. SUMOFLUX: a generalized method for targeted 13C metabolic flux ratio analysis. PLoS Comput. Biol. 12 (9), e1005109.

Kümmel, A., Panke, S., Heinemann, M., 2006. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol. Syst. Biol. 2 (1).

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. http://dx. doi.org/10.1038/nature14539.

Lee, S.Y., Kim, H.U., 2015. Systems strategies for developing industrial microbial strains. Nat. Biotechnol. 33.

Leung, M.K.K., Xiong, H.Y., Lee, L.J., Frey, B.J., 2014. Deep learning of the tissue-regu- lated splicing code. Bioinformatics 30, i121–i129.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332.

Lin, P.C., Saha, R., Zhang, F., Pakrasi, H.B., 2017. Metabolic engineering of the pentose phosphate pathway for enhanced limonene production in the cyanobacterium *Synechocysti* s sp. PCC 6803. Sci. Rep. 7 (1), 17503.

Liu, D., Wan, N., Zhang, F., Tang, Y.J., Wu, S.G., 2017. Enhancing fatty acid production in *Escherichia coli* by Vitreoscilla hemoglobin overexpression. Biotechnol. Bioeng. 114, 463–467.

Long, M.R., Ong, W.K., Reed, J.L., 2015. Computational methods in metabolic en- gineering for strain design. Curr. Opin. Biotechnol. 34, 135–141.

Maarleveld, T.R., Boele, J., Bruggeman, F.J., Teusink, B., 2014. A data integration and visualization resource for the metabolic network of *Synechocystis* sp. PCC 6803. Plant Physiol. 113.

Machado, D., Herrgård, M., 2014. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput. Biol. 10, e1003580.

Mellor, J., Grigoras, I., Carbonell, P., Faulon, J.L., 2016. Semisupervised Gaussian process for automated enzyme search. ACS Synth. Biol. 5, 518–528.

Monk, J.M., Koza, A., Campodonico, M.A., Machado, D., Seoane, J.M., Palsson, B.O., Herrgard, M.J., Feist, A.M., 2016. Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. Cell Syst. 3, 238–251 (e12).

Morrell, W., Birkel, G., Forrer, M., Lopez, T., Backman, T., Dussault, M., Petzold, C.J., Baidoo, E.E.K., Costello, Z., Ando, D., Alonso Gutierrez, J., George, K., Mukhopadhyay, A., Vaino, I., Keasling, J.D., Adams, P.D., Hillson, N.J., Garcia Martin, H., 2017. The Experiment Data Depot: a web-based software tool for biolo- gical experimental data storage, sharing, and visualization. ACS Synth. Biol. http:// dx.doi.org/10.1021/acssynbio.7b00204.

Nandi, S., Subramanian, A., Sarkar, R., 2017. An integrative machine learning strategy for improved prediction of essential genes in *Escherichia coli* metabolism using flux- coupled features. Mol. BioSyst. 1584–1596.

Nishizaki, T., Tsuge, K., Itaya, M., Doi, N., Yanagawa, H., 2007. Metabolic engineering of carotenoid biosynthesis in *Escherichia coli* by ordered gene assembly in *Bacillus sub- tilis*. Appl. Environ. Microbiol. 73, 1355–1361.

O'Brien, E.J., Lerman, J. a, Chang, R.L., Hyduke, D.R., Palsson, B.Ø., 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. 9, 693.

Orth, J.D., Thiele, I., Palsson, B.Ø., 2010. What is flux balance analysis? Nat. Biotechnol. 28, 245–248.

Oyetunde, T., Zhang, M., Chen, Y., Tang, Y., Lo, C., 2016. BoostGAPFILL: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods. Bioinformatics 33 (4), 608–611.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.

Pappu, J.S.M., Gummadi, S.N., 2016. Modeling and simulation of xylitol production in bioreactor by *Debaryomyces nepalensis* NCYC 3413 using unstructured and artificial neural network models. Bioresour. Technol. 220, 490–499.

Parekh, S., Vinci, V. a, Strobel, R.J., 2000. Improvement of microbial strains and fermentation processes. Appl. Microbiol. Biotechnol. 54, 287–301.

Pharkya, P., Maranas, C.D., 2006. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab. Eng. 8 (1), 1–13.

Poshyvailo, L., von Lieres, E., Kondrat, S., 2017. Does metabolite channeling accelerate enzyme-catalyzed cascade reactions? PLoS One 12, e0172673.

Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y., 2007. Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning. ACM, pp. 759–766.

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 (4), e1000744.

Razavian, N., 2004. Applications of Machine Learning in Computational Biology.

Schuetz, R., Kuepfer, L., Sauer, U., 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol. Syst. Biol. 3, 119.

Sheng, J., Guo, W., Ash, C., Freitas, B., Paoletti, M., Feng, X., 2017. Data-Driven Prediction of CRISPR-Based Transcription Regulation for Programmable Control of Metabolic Flux. arXiv Prepr (arXiv1704.03027).

Sinha, R., Singh, S., Srivastava, P., 2014. Studies on process optimization methods for rapamycin production using *Streptomyces hygroscopicus* ATCC 29253. Bioprocess Biosyst. Eng. 37, 829–840.

Sommer, C., Gerlich, D.W., 2013. Machine learning in cell biology – teaching computers to recognize phenotypes. J. Cell Sci. 126, 5529–5539.

Sowa, J.F., 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations. MIT Press.

Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge engineering: principles and methods. Data Knowl. Eng. 25, 161–197.

Tarca, A.L., Carey, V.J., Chen, X., Romero, R., Drăghici, S., 2007. Machine learning and its applications to biology. PLoS Comput. Biol. 3, e116.

Tummler, K., Lubitz, T., Schelker, M., Klipp, E., 2014. New types of experimental data shape the use of enzyme kinetics for dynamic network modeling. FEBS J. 281, 549–571.

Utrilla, J., O'Brien, E.J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A.M., Palsson, B.O., 2016. Global rebalancing of cellular resources by pleiotropic point mutations illustrates a multi-scale mechanism of adaptive evolution. Cell Syst. 2, 260–271.

Varman, A.M., Xiao, Y., Leonard, E., Tang, Y.J., 2011. Statistics-based model for pre- diction of chemical biosynthesis yield from *Saccharomyces cerevisiae*. Microb. Cell Factories 10, 45.

Wang, G., Xu, B., Jiang, W., 2016. SVM Modeling for Glutamic Acid Fermentation Process. pp. 5551–5555.

Wang, Y., Song, J., Marquez-lago, T.T., Leier, A., Li, C., Webb, G.I., Shen, H., 2017. Knowledge-Transfer Learning for Prediction of Matrix Metalloprotease Substrate-Cleavage Sites. pp. 1–15. http://dx.doi.org/10.1038/s41598-017-06219-7.

Winkler, J.D., Halweg-Edwards, A.L., Gill, R.T., 2015. The LASER database: formalizing design rules for metabolic engineering. Metab. Eng. Commun. 2, 30–38.

Wu, S.G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R., Zhang, X., Shimizu, K., Tang, Y.J., Bao, F.S., 2016. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS Comput. Biol. 12, e1004838.

Yang, H.F., Zhang, X.N., Li, Y., Zhang, Y.H., Xu, Q., Wei, D.Q., 2017. Theoretical Studies of Intracellular Concentration of Micro-organisms' Metabolites. Sci. Rep. 7 (1), 9048. Zhang, Z., Shen, T., Rui, B., Zhou, W., Zhou, X., Shang, C., Xin, C., Liu, X., Li, G., Jiang, J., Li, C., Li, R., Han, M., You, S., Yu, G., Yi, Y., Wen, H., Liu, Z., Xie, X., 2014. CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13C-fluxomics. Nucleic Acids Res. 43, D549–D557.

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