

Evaluating the potential of applying machine learning tools to metabolic pathway optimization

Wenfa Ng

Citizen scientist, Singapore, Email: ngwenfa771@hotmail.com

Abstract

Successful engineering of a microbial host for efficient production of a target product from a given substrate can be viewed as an extensive optimization task. Such a task involves the selection of high activity enzymes as well as their gene expression regulatory control elements (i.e., promoters and ribosome binding sites). Finally, there is also the need to tune expression of multiple genes along a heterologous pathway to relieve constraints from rate-limiting step and help reduce metabolic burden on cells from unnecessary over-expression of high activity enzymes. While the aforementioned tasks could be performed through combinatorial experiments, such an approach incurs significant cost, time and effort, which is a handicap that can be relieved by application of modern machine learning tools. Such tools could attempt to predict high activity enzymes from sequence, but they are currently most usefully applied in classifying strong promoters from weaker ones as well as combinatorial tuning of expression of multiple genes. This perspective reviews the application of machine learning tools to aid metabolic pathway optimization through identifying challenges in metabolic engineering that could be overcome with the help of machine learning tools.

Keywords: pathway optimization, machine learning tools, enzyme activity prediction, promoter classification, expression tuning,

Highlights

- Optimizing a metabolic pathway entails the selection of high activity enzymes and gene regulatory elements that give appropriate expression level
- Some success has been achieved in correlating enzyme kinetic parameters and structure with activity level through neural network approaches
- Support vector machines and deep learning have found use in predicting promoter strength from sequence information
- Artificial neural network could be used to tune expression of individual gene in a multi-enzyme pathway

Introduction

Metabolic engineering sought to increase the production of small molecules using cellular metabolism that has been rewired through genetic engineering. Comprising a workflow that can be categorized into design-build-test-learn cycle, a typical metabolic engineering project would involve iterative design and test experiments aiming to improve the expression of desired genes and production of target metabolites. As such, many aspects of engineering a microbe for overproducing a metabolite can be characterized as an optimization problem. Specifically, a key concern in metabolic engineering has been the selection of enzymes with high activity under a broad range of conditions, and strong promoters and ribosome binding

sites, and the tuning of expression of multiple genes [1-4]. These goals could not be achieved through rational selection, but only with trial and error experimentation. The latter is not desirable considering the time and effort involved, and this has sowed the seeds for introducing data-driven algorithmic approaches to the traditional confines of metabolic engineering.

Although optimization can be performed by various algorithms, the latest trend has been the application of machine learning tools in optimization problems. In general, machine learning algorithms sought to identify patterns hidden in large datasets, and this enabling feature has been used in different aspects of metabolic engineering such as pathway optimization. Fundamentally, machine learning algorithms builds a model from input data using an iterative cycle of parameter fitting to a curvilinear description of the data. The obtained machine learning model could subsequently be used in predicting properties of enzymes or pathways using certain inputs. This thus arrives at the core enabling feature of machine learning: it is an automated search for a set of mathematic descriptions that describe particular sets of data. Usually thought to be data intensive, machine learning tools could also be applied to small datasets with or without data augmentation [5-7], and this latter feature dovetails with the inherently small scale nature of many biological datasets.

Thus far, machine learning has been applied to many but not all aspects of metabolic engineering and pathway optimization. For example, machine learning has been utilised in reconstruction of metabolic model of a species [8-12]. *De novo* pathway engineering is another aspect that has benefited from application of machine learning tools [13]. Machine learning tools have also enabled the deciphering of kinetic parameters of enzymes from metabolomics data [14,15]. In the same vein, correlations between expression level and design of various gene expression control elements (e.g., promoter and ribosome binding site) has been sought using the tools of machine learning [16].

This article sought to review areas where machine learning has informed pathway optimization. These include: (i) selecting enzymes with the highest activity for a pathway, (ii) selecting promoters and ribosome binding site of appropriate strength for particular genes, and (iii) tuning the expression of multiple genes in a pathway (Figure 1). But, the journey marched by machine learning in metabolic pathway optimization remains incomplete. For example, opportunities exist in applying machine learning to predict the regulatory motifs of enzymes and pathway dynamics from multiomics data [17], as well as assessing the performance of microbial cell factory [18].

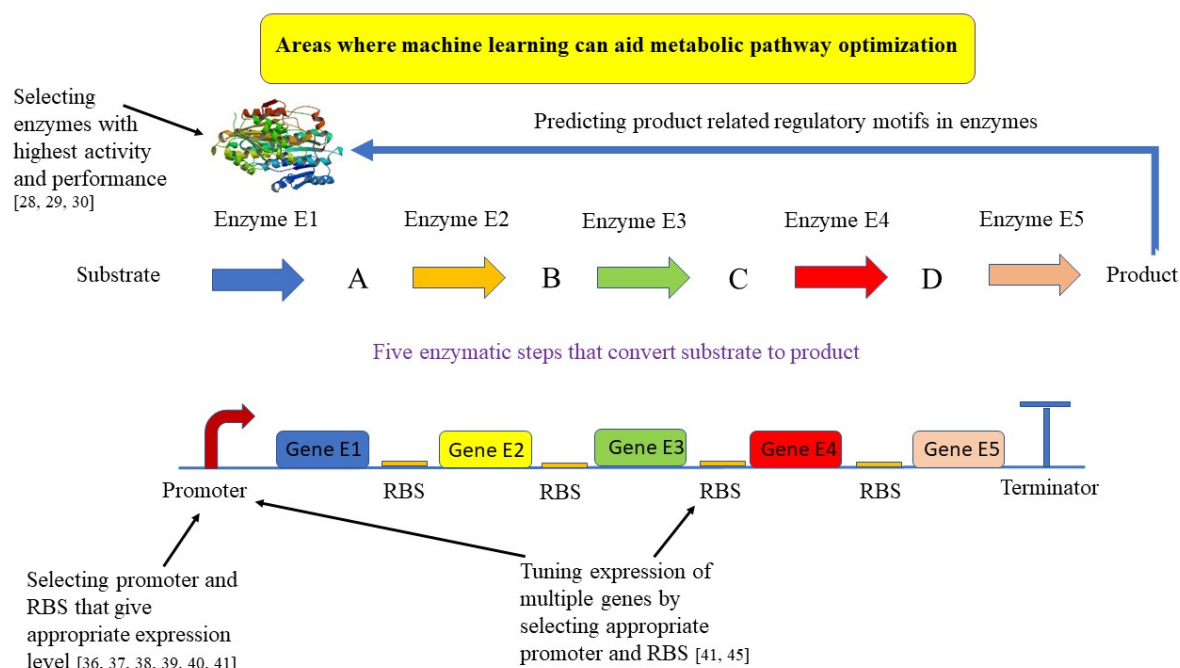


Figure 1: Areas where machine learning tools could aid metabolic pathway optimization at the enzyme selection, and promoter and ribosome binding site (RBS) tuning level. In particular, promoter and RBS tuning can be used in a combinatorial fashion to tune expression of multiple genes in a pathway. Discovery and prediction of enzyme regulatory motifs is an emerging area in which machine learning could aid metabolic pathway optimization.

Selecting the enzymes with highest activity for incorporating into a pathway

Essentially, a metabolic pathway comprises a set of reactions that transform a substrate into a product through a series of bond formation and cleavage. Both expert knowledge and retrobiosynthetic approaches could be used in developing this pathway [11]. In particular, machine learning methodologies have been successfully applied in retrobiosynthesis and is reviewed elsewhere [11,19]. With a set of coupled sequential reactions in mind, the next step is in selecting the appropriate enzymes for performing the respective reactions. Recently, machine learning has helped refine gene annotation through better recognition of genomic signals such as polyadenylation signals and translational start site [20]. In particular, deep learning approaches have played an important role in dissecting the often convoluted signals from the genome in assigning gene function to sequence information [21-25], and is poised to help identify more enzyme candidates with suitable functions in a metabolic engineering project. The latter comes about due to enzyme promiscuity where some enzymes could be repurposed for other functions [26,27]. Usually, enzymes with the highest activities and performance are desired. But, given the plethora of similar enzymes in different species, how does one select the best performing enzyme for an application? Can machine learning help rule out some candidates that are unlikely to work? In addition, which performance measure should be the basis for optimization? Since enzyme performance can be described by turnover number, inhibitory concentration (K_i) and binding affinity between substrate and enzyme (K_m), multiple parameters could be used in machine learning tasks for predicting enzyme performance with amino acid sequence as input. But, the challenge lies in relative lack of full set of

characterization data for different enzymes in varied species. Such data are incomplete given the effort and resources needed to perform detailed biochemical assays for each substrate.

One example of applying machine learning to predicting enzyme activity is in using an ensemble of enzyme characteristics such as biochemical parameters and structure to inform enzyme catalytic turnover number, which is a proxy parameter for enzyme activity [28] (Figure 2). Correlations between catalytic turnover number and enzyme structure elucidated by the machine learning tools hold important implications for how structural biology could inform enzyme biochemistry [28]. Indeed, other studies have also corroborated that enzyme conformation can be reliably correlated with enzyme activity level [29,30]. Furthermore, it has been shown that sequence alone could not accurately describe enzyme activity [31]. However, a study has shown that combination of sequence information and structural descriptors of enzyme-substrate recognition is useful for predicting enzyme function and activity [31]. Hence, the current state-of-the-art in enzyme activity prediction remains firmly in the realm of structure-activity correlate, with attempts at extending the correlation to the sequence level meeting challenges at our inability to resolve the protein folding problem. But future advances in using machine learning to circumvent the protein structure prediction problem may ultimately tie the link between enzyme sequence and activity level.

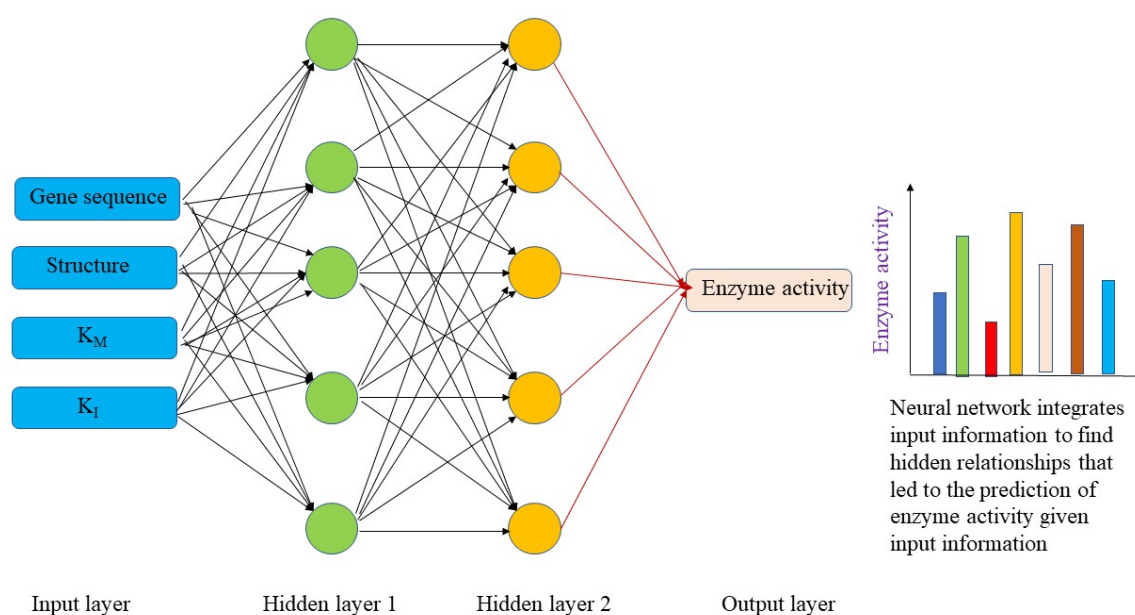


Figure 2: Neural network is a common machine learning tool for processing and mining multivariate complex input information. Complexity of the neural network depends on the number of hidden layers that process input information from the previous layer. Through processing by hidden layers, different facets of the input information are effectively mined to glean hidden mathematical relationships between variables. Shown here is the approach for mining hidden enzyme characteristics (sequence, structure and kinetic parameters) and enzyme activity correlate through neural network machine learning.

Optimization of gene expression regulatory elements

Gene expression regulatory elements such as promoters and ribosome binding site (RBS) controls the level at which the heterologous genes could be expressed. To facilitate selection of appropriate promoter and RBS for tuning the expression of heterologous genes, a need exists to build a predictive model able to correlate promoter or RBS sequence with expression level.

Theoretically, building a machine learning model capable of predicting promoter or RBS strength from sequence information does not necessarily require accurate definition of promoter or RBS sequence, which remains a research topic [32]. But, if promoter sequences in a training dataset are accurate, this would reduce the noise in the model and afford more accurate prediction. Hence, the computational challenge in applying machine learning to promoter strength prediction lies in the identification of small snippets of nucleotide sequence that strongly correlates with expression level [33]. Currently, a commonly used method for extracting sequence features is position weight matrix [34], but the approach may not be transferable to different species [33]. Another problem with promoter strength prediction is the relative lack of data, particularly in cases where machine learning is applied to experimentally characterized promoters [35,36]. But, use of genome-wide RNA-seq data may provide sufficient data that significantly improves machine learning based predictions of promoter strength.

Typically, the input data for training are promoter or RBS sequence and expression level as measured by protein or mRNA transcripts abundance. Such data could be modelled by support vector machine algorithms [36,37] (Figure 3), but recently, deep learning methods have also been applied to the problem and have shown promising results [38,39]. One approach uses pseudo-dinucleotide composition coupled to CNN for both promoter identification and strength prediction in prokaryotic organisms [39]. The method demonstrated good performance compared to state-of-the-art methods, but it is still limited to classifying promoters into strong or weak promoters, which does not provide metabolic engineers with the ability to achieve fine-grained control over gene expression. Another approach took into consideration evolutionarily relationships between orthologous genes and showed that such a methodology provided better predictions of mRNA abundance from DNA sequence [40]. Overall, neural network-based approaches may not be the only way forward in promoter strength prediction, particularly in cases with small datasets. For example, kernel-based approaches such as support vector machines have provided better performance compared to artificial neural network in some instances [36].

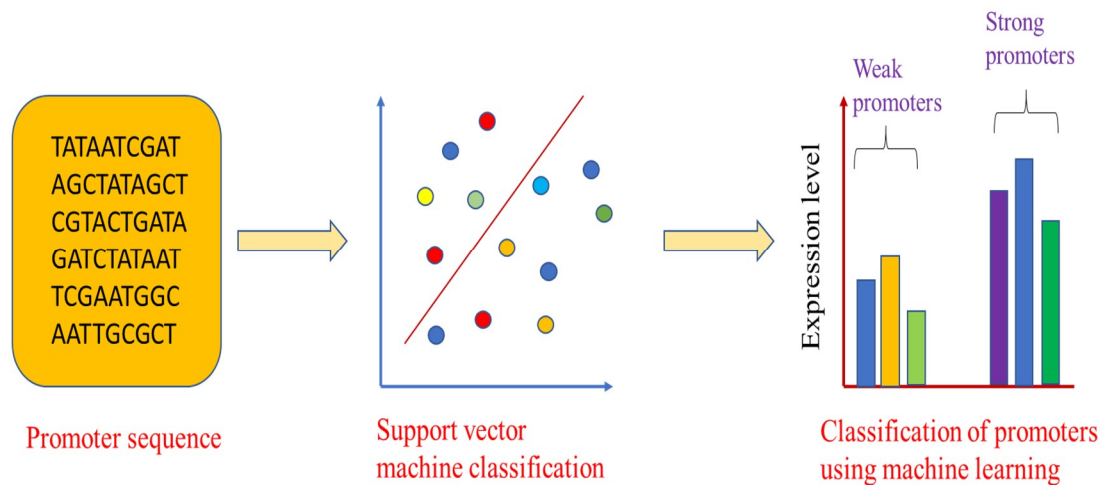


Figure 3: Kernel-based approaches such as support vector machine (SVM) glean hidden relationships in input data through discerning an imaginary plane that classifies data elements into different groups. Such SVM models could serve as classifier of data such as classifying promoters of a given sequence into strong or weak promoters. But, with better and more data, SVM could also build regression models that uses a mathematical relationship to describe the correlation between promoter sequence and expression level.

In comparison to promoters, ribosome binding sites are more well-defined. This comes about due to the structure of gene regulatory region where ribosome binding sites (RBS) are downstream of the transcriptional start site (TSS), which could be experimentally defined by RNA-seq data [33]. Similar to promoters, RBS are important modulators of gene expression level given that it governs the strength of binding between the small subunit (SSU) of ribosome with the mRNA transcript obtained after transcription. In a recent study, machine learning tools were used in defining the RBS sequence-phenotype relationship, which forms the basis for predicting optimal RBS sequences for multi-gene pathway. Computational predictions were validated through experiments and demonstrated the approach's utility in enabling screening of a large combination of RBS sequences for multi-gene pathway [41]. But, in general, correlation between RBS sequence and expression level may not be easily discernible by machine learning tools. For example, a recent study did not find strong correlation between experimental protein expression data and predicted RBS strength [42], thereby, indicating room for improvement in the application of machine learning to RBS strength prediction. One major hurdle in RBS strength prediction comes from the relatively small sequence space of these gene regulatory elements as RBS are inherently shorter than promoters. Lack of sufficient variability in expression level from the small RBS sequence set would thus severely hamper prediction of protein expression level from RBS sequence.

Tuning expression of multiple genes using machine learning

Expression of a heterologous gene in a cell incurs a metabolic burden. For long pathway comprising multiple genes, such metabolic burden may have a detrimental effect on cell growth. In other situations, there may be excessive expression of enzyme for a particular step of the pathway that may lead to depletion of an intermediate metabolite needed to maintain other critical pathways of the cell. Hence, a need exists in tuning the expression of individual gene in a pathway to ensure that only sufficient enzymes are expressed to enable proper

functioning of the pathway and deliver higher yields, and preventing metabolic choke points from emerging. Tools available for tuning the expression of multiple genes in a pathway would be promoter and ribosome binding site. Since heterologous genes are typically expressed in an operon in prokaryotic hosts, ribosome binding site tuning are more often used in prokaryotes. On the other hand, need for individual promoter for each gene of the pathway in eukaryotic hosts meant that promoter tuning is as important as ribosome binding site tuning in eukaryotes. Combinatorial tuning of promoter and RBS may thus afford fine-grained control over gene expression in eukaryotes.

Conceptually, the problem of optimizing expression of individual genes in a pathway can be depicted as a search for optimal levels of individual enzyme in a gene expression landscape. Statistical design of experiments approach has been put to use in this endeavour, yielding promising results that reduce experiment effort [43,44]. However, such search for the optimal combination of promoter and RBS usually will not arrive at the global optimum. In particular, extent in which the gene expression landscape is sampled determines the likelihood in which an optimal could be obtained. As the number of tunable parameters (promoter and RBS) increases with each additional gene in the pathway, the optimization problem could quickly escalate in complexity and size beyond the search capability of conventional optimization algorithms.

One way to circumvent the problem is through employing machine learning to detect hidden mathematical relationships between different sampled points on the gene expression landscape where combinatorial pathway optimization experimentation help deliver the data points that feeds the machine learning algorithms (Figure 4). In a recent example, artificial neural networks are employed to glean relationships between product titer of different strains with different promoters in a combinatorial optimization exercise. Predictions from the machine learning algorithms were verified experimentally, thereby, demonstrating the utility of the approach [45]. Besides neural networks, support vector machines and other kernel-based approaches may also be useful for such applications. However, how well machine learning performs critically depends on the characteristics of the input dataset and its size. Better predictions would naturally come from a larger dataset, which places greater demand on experimentation in combinatorial tuning of expression of multiple genes. In addition, input data should also cover a wide range in order to achieve a large dynamic range for corresponding predictions of product yield, titer, and productivity. At a more fundamental level, much room exists for the utilization of machine learning approaches in combinatorial pathway optimization since most studies in the field still relies on statistical design of experiment or construction of smart libraries to expedite search for optimal gene expression level of a pathway [46,47]. Developing better methods to efficiently and cost-effectively generate the input data for training various machine learning models remain an important research topic.

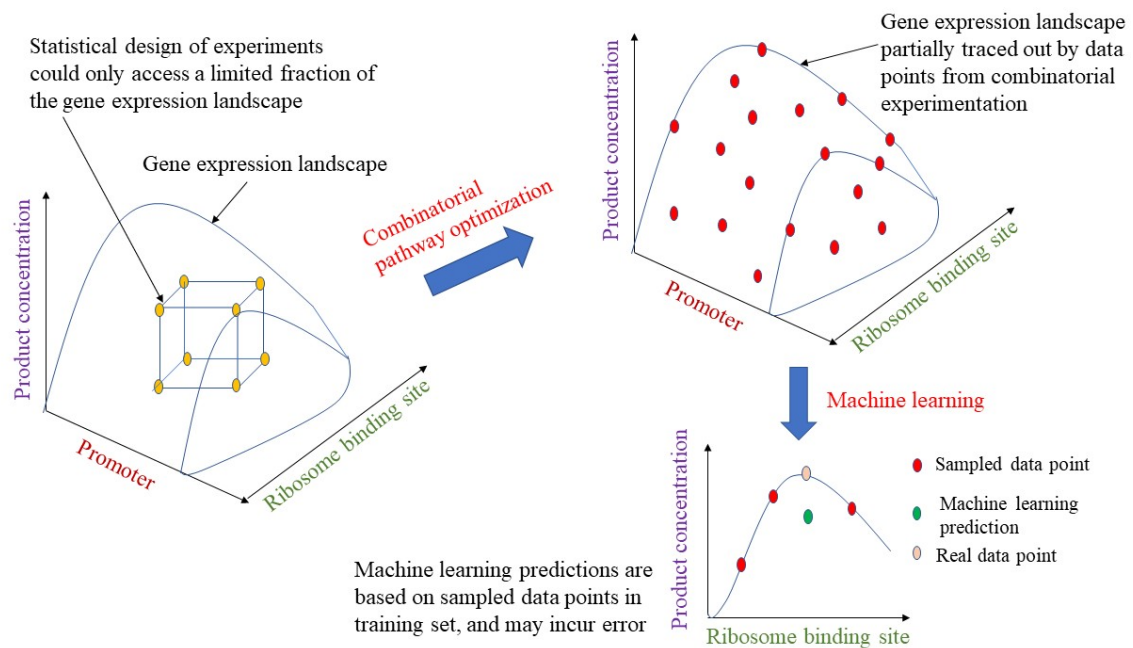


Figure 4: Progressive refinement in our ability to sample a complex multi-dimensional gene expression landscape. Specifically, traditional statistical design of experiments could only sample a limited fraction of the gene expression landscape. This situation is partially ameliorated with combinatorial pathway optimization that afford sampling of a larger fraction of the landscape. Machine learning could theoretically build upon combinatorial pathway optimization by using its data points as training set to impute values between experimental data points on a curve, but errors remain inevitable given the relative lack of data points in experimental biological dataset.

Conclusions

Machine learning has been applied to many facets of metabolic engineering and pathway optimization. From selection of enzymes to tuning of gene regulatory elements, machine learning's greatest strength has been the gleaning of hidden patterns in complex dataset to help offer solutions in new situations through building a predictive mathematical model. Such automated tools significantly ease the burden on metabolic engineers in making critical decisions such as gene selection and promoter choice during pathway optimization. But, application of machine learning tools to metabolic engineering remain a significant challenge for the novice researcher. This is made even harder by the cryptic nature of machine learning algorithms. Thus, more resources may be provided to enable researchers to begin integrating these tools across the pathway development process.

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work,

References

1. Santos-Merino M, Singh AK, Ducat DC: **New Applications of Synthetic Biology Tools for Cyanobacterial Metabolic Engineering.** *Frontiers in Bioengineering and Biotechnology* 2019, **7**.
2. Smirnov N: **Engineering of Metabolic Pathways Using Synthetic Enzyme Complexes.** *Plant Physiology* 2019, **179**:918.
3. Alper HS, Avalos JL: **Metabolic pathway engineering.** *Synthetic and Systems Biotechnology* 2018, **3**:1-2.
4. Choi KR, Jang WD, Yang D, Cho JS, Park D, Lee SY: **Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering.** *Trends in Biotechnology* 2019, **37**:817-837.
5. Shaikhina T, Khovanova NA: **Handling limited datasets with neural networks in medical applications: A small-data approach.** *Artificial Intelligence in Medicine* 2017, **75**:51-63.
6. Deist TM, Patti A, Wang Z, Krane D, Sorenson T, Craft D: **Simulation-assisted machine learning.** *Bioinformatics* 2019, **35**:4072-4080.
7. Zhang Y, Ling C: **A strategy to apply machine learning to small datasets in materials science.** *npj Computational Materials* 2018, **4**:25.
8. Zampieri G, Vijayakumar S, Yaneske E, Angione C: **Machine and deep learning meet genome-scale metabolic modeling.** *PLOS Computational Biology* 2019, **15**:e1007084.
9. Toubiana D, Puzis R, Wen L, Sikron N, Kurmanbayeva A, Soltabayeva A, del Mar Rubio Wilhelmi M, Sade N, Fait A, Sagi M, et al.: **Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data.** *Communications Biology* 2019, **2**:214.
10. Baranwal M, Magner A, Elvati P, Saldinger J, Violi A, Hero AO: **A deep learning architecture for metabolic pathway prediction.** *Bioinformatics* 2019.
11. Wang L, Dash S, Ng CY, Maranas CD: **A review of computational tools for design and reconstruction of metabolic pathways.** *Synthetic and Systems Biotechnology* 2017, **2**:243-252.
12. Boudelloua I, Saidi R, Hoehndorf R, Martin MJ, Solovyev V: **Prediction of Metabolic Pathway Involvement in Prokaryotic UniProtKB Data by Association Rule Mining.** *PLOS ONE* 2016, **11**:e0158896.
13. Presnell KV, Alper HS: **Systems Metabolic Engineering Meets Machine Learning: A New Era for Data-Driven Metabolic Engineering.** *Biotechnology Journal* 2019, **14**:1800416.
14. St. John PC, Strutz J, Broadbelt LJ, Tyo KEJ, Bomble YJ: **Bayesian inference of metabolic kinetics from genome-scale multiomics data.** *PLOS Computational Biology* 2019, **15**:e1007424.
15. St. John PC, Bomble YJ: **Approaches to Computational Strain Design in the Multiomics Era.** *Frontiers in Microbiology* 2019, **10**.
16. de Jongh RPH, van Dijk ADJ, Julsing MK, Schaap PJ, de Ridder D: **Designing Eukaryotic Gene Expression Regulation Using Machine Learning.** *Trends in Biotechnology* 2020, **38**:191-201.
17. Costello Z, Martin HG: **A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data.** *npj Systems Biology and Applications* 2018, **4**:19.
18. Oyetunde T, Liu D, Martin HG, Tang YJ: **Machine learning framework for assessment of microbial factory performance.** *PLOS ONE* 2019, **14**:e0210558.

19. Lin G-M, Warden-Rothman R, Voigt CA: **Retrosynthetic design of metabolic pathways to chemicals not found in nature.** *Current Opinion in Systems Biology* 2019, **14**:82-107.
20. Kalkatawi M, Magana-Mora A, Jankovic B, Bajic VB: **DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions.** *Bioinformatics* 2018, **35**:1125-1132.
21. Ryu JY, Kim HU, Lee SY: **Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers.** *Proceedings of the National Academy of Sciences* 2019, **116**:13996.
22. Dalkiran A, Rifaioğlu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T: **ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature.** *BMC Bioinformatics* 2018, **19**:334.
23. Gao R, Wang M, Zhou J, Fu Y, Liang M, Guo D, Nie J: **Prediction of Enzyme Function Based on Three Parallel Deep CNN and Amino Acid Mutation.** *International Journal of Molecular Sciences* 2019, **20**:2845.
24. Li Y, Wang S, Umarov R, Xie B, Fan M, Li L, Gao X: **DEEPre: sequence-based enzyme EC number prediction by deep learning.** *Bioinformatics* 2017, **34**:760-769.
25. Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H: **Robust predictions of specialized metabolism genes through machine learning.** *Proceedings of the National Academy of Sciences* 2019, **116**:2344.
26. Chen W, Yao J, Meng J, Han W, Tao Y, Chen Y, Guo Y, Shi G, He Y, Jin J-M, et al.: **Promiscuous enzymatic activity-aided multiple-pathway network design for metabolic flux rearrangement in hydroxytyrosol biosynthesis.** *Nature Communications* 2019, **10**:960.
27. Glasner ME, Truong DP, Morse BC: **How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation.** *The FEBS Journal* 2019, **n/a**.
28. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO: **Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models.** *Nature Communications* 2018, **9**:5252.
29. Suleymanov Y: **Enzyme activity from machine learning.** *Science* 2019, **364**:448.
30. Bonk BM, Weis JW, Tidor B: **Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis.** *Journal of the American Chemical Society* 2019, **141**:4108-4118.
31. Yang M, Fehl C, Lees KV, Lim E-K, Offen WA, Davies GJ, Bowles DJ, Davidson MG, Roberts SJ, Davis BG: **Functional and informatics analysis enables glycosyltransferase activity prediction.** *Nature Chemical Biology* 2018, **14**:1109-1117.
32. Abbas MM, Mohie-Eldin MM, El-Manzalawy Y: **Assessing the Effects of Data Selection and Representation on the Development of Reliable E. coli Sigma 70 Promoter Region Predictors.** *PLOS ONE* 2015, **10**:e0119721.
33. Wang S, Cheng X, Li Y, Wu M, Zhao Y: **Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns.** *Scientific Reports* 2018, **8**:17695.
34. Bharanikumar R, Premkumar KAR, Palaniappan A: **PromoterPredict: sequence-based modelling of Escherichia coli σ 70 promoter strength yields logarithmic dependence between promoter strength and sequence.** *PeerJ* 2018, **6**:e5862.

35. Gilman J, Singleton C, Tennant RK, James P, Howard TP, Lux T, Parker DA, Love J: **Rapid, Heuristic Discovery and Design of Promoter Collections in Non-Model Microbes for Industrial Applications.** *ACS Synthetic Biology* 2019, **8**:1175-1186.
36. Meng H, Ma Y, Mai G, Wang Y, Liu C: **Construction of precise support vector machine based models for predicting promoter strength.** *Quantitative Biology* 2017, **5**:90-98.
37. Lai H-Y, Zhang Z-Y, Su Z-D, Su W, Ding H, Chen W, Lin H: **iProEP: A Computational Predictor for Predicting Promoter.** *Molecular Therapy - Nucleic Acids* 2019, **17**:337-346.
38. Oubounyt M, Louadi Z, Tayara H, Chong KT: **DeePromoter: Robust Promoter Predictor Using Deep Learning.** *Frontiers in Genetics* 2019, **10**.
39. Tayara H, Tahir M, Chong KT: **Identification of prokaryotic promoters and their strength by integrating heterogeneous features.** *Genomics* 2020, **112**:1396-1403.
40. Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H: **Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence.** *Proceedings of the National Academy of Sciences* 2019, **116**:5542.
41. Jervis AJ, Carbonell P, Vinaixa M, Dunstan MS, Hollywood KA, Robinson CJ, Rattray NJW, Yan C, Swainston N, Currin A, et al.: **Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in Escherichia coli.** *ACS Synthetic Biology* 2019, **8**:127-136.
42. Opgenorth P, Costello Z, Okada T, Goyal G, Chen Y, Gin J, Benites V, de Raad M, Northen TR, Deng K, et al.: **Lessons from Two Design–Build–Test–Learn Cycles of Dodecanol Production in Escherichia coli Aided by Machine Learning.** *ACS Synthetic Biology* 2019, **8**:1337-1351.
43. Xu P, Rizzoni EA, Sul S-Y, Stephanopoulos G: **Improving Metabolic Pathway Efficiency by Statistical Model-Based Multivariate Regulatory Metabolic Engineering.** *ACS Synthetic Biology* 2017, **6**:148-158.
44. Heinsch SC, Das SR, Smanski MJ: **Simulation Modeling to Compare High-Throughput, Low-Iteration Optimization Strategies for Metabolic Engineering.** *Frontiers in Microbiology* 2018, **9**.
45. Zhou Y, Li G, Dong J, Xing X-h, Dai J, Zhang C: **MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae.** *Metabolic Engineering* 2018, **47**:294-302.
46. Jeschek M, Gerngross D, Panke S: **Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort.** *Nature Communications* 2016, **7**:11163.
47. Jeschek M, Gerngross D, Panke S: **Combinatorial pathway optimization for streamlined metabolic engineering.** *Current Opinion in Biotechnology* 2017, **47**:142-151.