Metabolic Engineering Communications 11 (2020) e00149

Contents lists available at ScienceDirect

Metabolic Engineering Communications

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

Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering

Mohamed Helmy [a], Derek Smith [a], Kumar Selvarajoo [a,b,*]

[a] Singapore Institute of Food and Biotechnology Innovation (SIFBI), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[b] Synthetic Biology for Clinical and Technological Innovation (SynCTI), National University of Singapore (NUS), Singapore, Singapore

ARTICLE INFO

Keywords:
Systems biology
Artificial intelligence
Machine learning
Metabolic engineering
Food industry

ABSTRACT

Metabolic engineering aims to maximize the production of bio-economically important substances (compounds, enzymes, or other proteins) through the optimization of the genetics, cellular processes and growth conditions of microorganisms. This requires detailed understanding of underlying metabolic pathways involved in the production of the targeted substances, and how the cellular processes or growth conditions are regulated by the engineering. To achieve this goal, a large system of experimental techniques, compound libraries, computational methods and data resources, including multi-omics data, are used. The recent advent of multi-omics systems biology approaches significantly impacted the field by opening new avenues to perform dynamic and large-scale analyses that deepen our knowledge on the manipulations. However, with the enormous transcriptomics, proteomics and metabolomics available, it is a daunting task to integrate the data for a more holistic understanding. Novel data mining and analytics approaches, including Artificial Intelligence (AI), can provide breakthroughs where traditional low-throughput experiment-alone methods cannot easily achieve. Here, we review the latest attempts of combining systems biology and AI in metabolic engineering research, and highlight how this alliance can help overcome the current challenges facing industrial biotechnology, especially for food-related substances and compounds using microorganisms.

1. Introduction

With the growing population of our planet, food security remains a major challenge facing mankind. This is especially true for countries that do not possess large land spaces for agriculture, such as those in the Middle East (deserts), Japan (mostly mountainous), and Singapore (land scarce). Moreover, nature conservationists are mostly against the clearing of wild flora and fauna to feed the world. With the rapid phase of global population growth, food security has become even more important during the ongoing COVID-19 pandemic when countries have largely closed their borders, affecting the food import-export trade (Laborde et al., 2020). Furthermore, some countries have decided to stop food export until the end of the year (Laborde et al., 2020). Therefore, in times of crises such as pandemics, there is an imminent need to find alternative sources of food and food ingredients.

One possible way to supplement conventional food and ingredients stock is to adopt carefully engineered GRAS microorganisms, such as bacteria and yeast, for the production of food compounds (e.g. total protein production) or targeted substances (e.g. alcohol or vitamins) by optimizing the genetics and/or growth conditions (Nozzi et al., 2014; Xiao et al., 2015). In nature, however, the microorganisms explored or used often do not produce the needed amounts of the required proteins, substances or compounds. In such a situation, metabolic engineering can play a major role. The process involves identifying key pathways and enzymes that can be modified for the optimal production of the target molecule (Kallscheuer, 2018). This can be achieved through transcriptional and/or translational control, enzyme engineering (mutation and/or truncation) and growth optimization (García-Granados et al., 2019; Shukal et al., 2019). Furthermore, it is also an excellent platform to produce rare and economically valuable products such as taste substances, fragrance and cosmetic compounds. The overall increase in the yield has to be maximized so that the production can be economically viable. Therefore, metabolic engineering approaches aim to maximize the titres-rates-yields (TRYs) to be industrially competitive as compared to other methods such as chemical synthesis and extraction from natural substances (Zhang et al., 2020a; Comba et al., 2012). To achieve this

* Corresponding author. Singapore Institute of Food and Biotechnology Innovation (SIFBI), Agency for Science, Technology and Research (A*STAR), 61 Biopolis Drive, #04-14, 138673, Singapore
E-mail address: Kumar_Selvarajoo@sifbi.a-star.edu.sg (K. Selvarajoo).

https://doi.org/10.1016/j.mec.2020.e00149
Received 1 September 2020; Received in revised form 1 October 2020; Accepted 7 October 2020

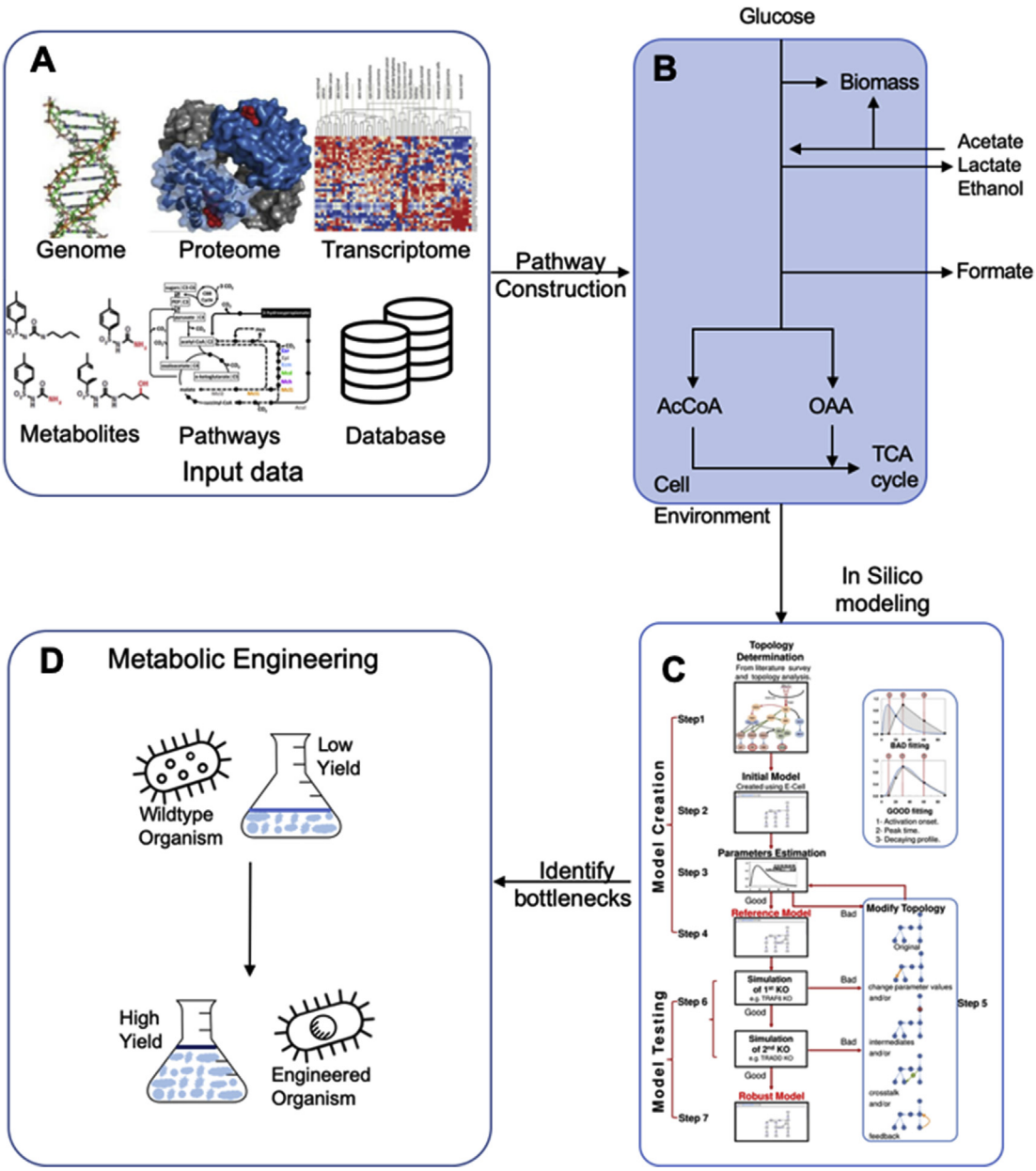

Fig. 1. Overview of the modeling strategies in the metabolic engineering research. (A) Data from different sources are used to construct (B) the metabolic pathways that produces the substance of interest. (C) An appropriate computational modeling approach is used to simulate the pathway response to a given perturbation *in silico*. The simulation results are analyzed to identity key regulatory steps, such as bottlenecks, which will then be tested using data from different conditions (e.g. gene knockouts or different growth conditions) (figure adapted from Helmy et al., 2009). (D) Finally, the model predictions are experimentally validated.

goal, limiting one's efforts only through experimental approaches may be insufficient. Interdisciplinary approaches linking mathematics, computational science and physics with metabolic engineering could most likely pave the best way forward (Fig. 1).

2. Modeling strategies in metabolic engineering

Computational modelling uses mathematical and statistical approaches built into computer algorithms that analyses experimental data to provide better understandings of the biological systems and/or predictions that guide subsequent lab work in an iterative manner (Helmy et al., 2009; Piras et al., 2014; Selvarajoo, 2017, 2018). It has become an integral and indispensable part of modern day biological research, especially when studying cellular networks through systems biology approaches (Kim et al., 2018). In the field of metabolic engineering, several types of computational models are employed and they provide new insights in identifying and tackling its challenges (Saa and Nielsen, 2017).

2.1. Dynamic and constraint-based metabolic modelling

There are several types of modelling approaches today, that can be largely grouped into i) parametric approaches such as dynamic modeling using ordinary differential equations (Kim et al., 2018), and ii) non-parametric models using Boolean logics, stoichiometric matrix and Bayesian inference algorithms (John et al., 2019; Toya and Shimizu, 2013). A dynamic model built using differential equations constructs an organism's metabolism step by step using known biochemical reactions and reaction kinetics from their genomic, enzymatic and biochemical information derived from experiments (Fig. 2A). Using this information, the models are used to predict metabolic outcomes for different *in silico*

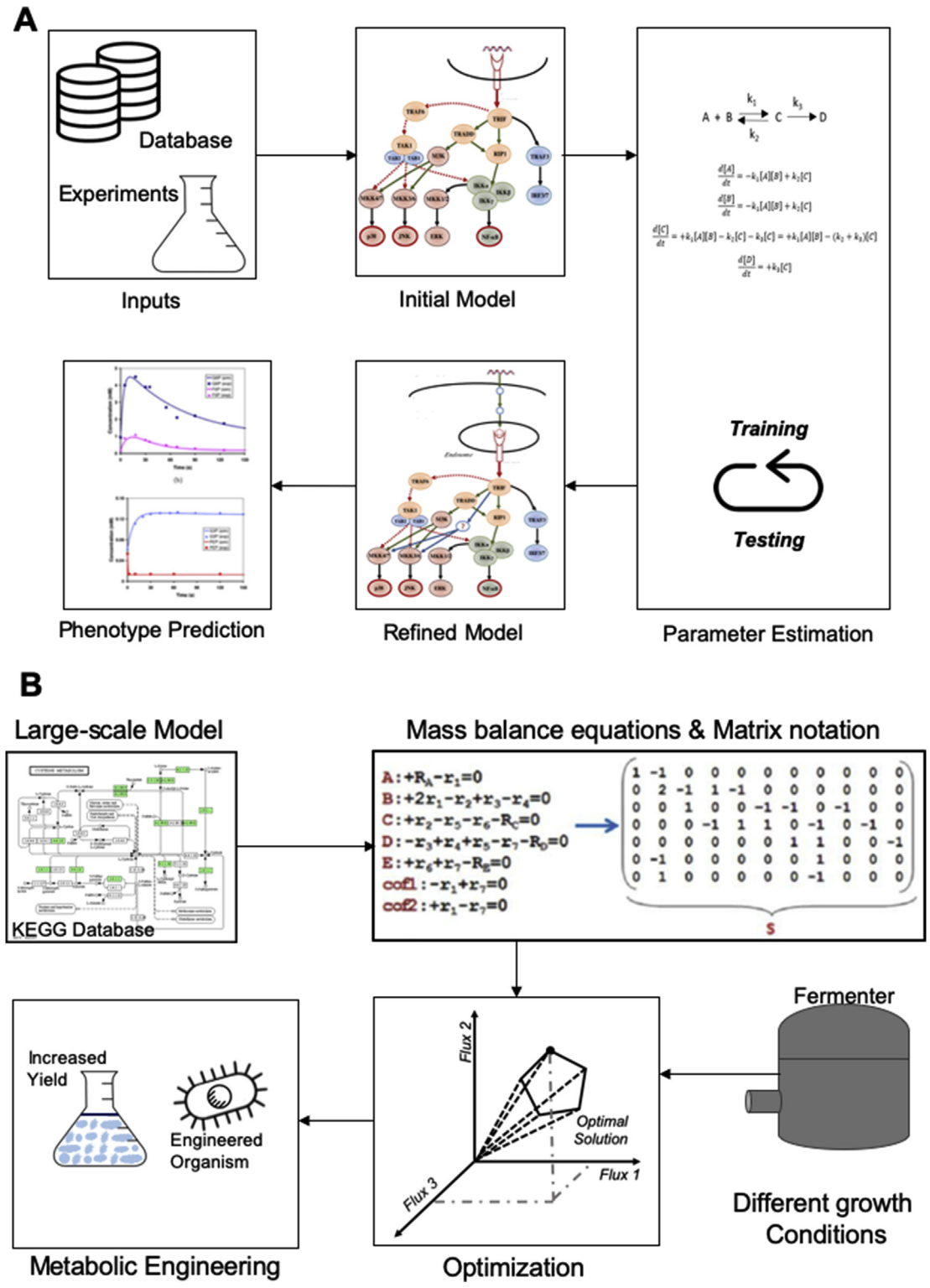

Fig. 2. Schematic representation of dynamic and constraint-based modeling approaches used in metabolic engineering. A) Dynamic kinetic modeling of metabolic pathways using differential equations. B) Flux balance analysis (FBA) modeling.

perturbations, or to understand the key regulatory mechanisms (such as bottlenecks) and flux distributions to a given perturbation (Selvarajoo et al., 2008; McCloskey et al., 2013). In other words, the dynamic models utilize *a priori* knowledge of metabolic pathways, enzymatic mechanisms and temporal experimental data to simulate the concentrations of metabolites over time. These models are usually referred to as kinetic models (Selvarajoo and Tomita, 2013).

Although kinetic models have been widely used and have proven their benefits (Saa and Nielsen, 2017), for large-scale modeling, such as genome-scale modeling, it is a daunting challenge to use dynamic modeling due to the absence of large-scale experimentally measured and reliable kinetics (Kim et al., 2018). To overcome this major challenge, as a trade-off, scientists use other types of modeling such as the parameter-less stoichiometric constraint-based modeling approaches.

Constraint-based models, have constraints for each decision that represent the minimum and maximum values of the decision (e.g. the minimum and maximum reaction rates) (Bordbar et al., 2014). A widely used constraint-based modeling is the flux balance analysis (FBA) (Orth et al., 2010). The FBA models thousands of metabolites and reactions with reasonable computational cost and prediction outcome (Fig. 2B). With FBA, the contribution of each individual gene to certain trait can be determined, and it can be used for the analysis, optimization and design of metabolic pathways (Skraly et al., 2018).

2.2. *Transcriptional control and ensemble modeling for metabolic pathway analysis*

Although numerous works have used metabolic regulation to control the production of targeted metabolites, recent works indicate that transcriptional and translation control can provide significant fold increase in the intended yield output (Shukal et al., 2019; Curran et al., 2014). The transcriptional control changes the way the gene of interest is regulated by manipulating its promoter region. This includes modifications such as mutating the ribosomal binding sites (RBS), the transcription factor binding sites (TFBS), designing and inserting shot sequencing (e.g. new binding sites), or designing an artificial promoter region (Curran et al., 2014). The transcriptional control requires deep understanding of how the gene of interest is regulated (activators, enhancers and suppressors) as well as the knowledge of its genomic structure around the binding sites, such as the nucleosome positions (Sharon et al., 2012) (Fig. 3A). Thus, modeling the transcriptional control remains a challenge as it requires complex data involving quantitative gene expression under each mutation condition to train a model that simulates the effect of each mutation and then use it to predict the impact on the new mutation. Nevertheless, statistical approaches such as the position weight matrix (PWM) modeling, which measures or scores aligned sequences that are likely functionally related, have shown promise for understanding the mutational impact on the transcriptional regulation in mammalian disease cells (Yiu Chan et al., 2019; Ji et al., 2018). Such methods could be explored in the future for controlling the transcriptional efficiency for metabolic engineering outcome.

Another modeling strategy that is used to model the metabolic pathways is ensemble modeling. Ensemble modeling is a strategy where multiple models with different modeling algorithms or multiple training sets are used to model and predict an outcome of a pathway. The prediction results of each base model are aggregated into one prediction

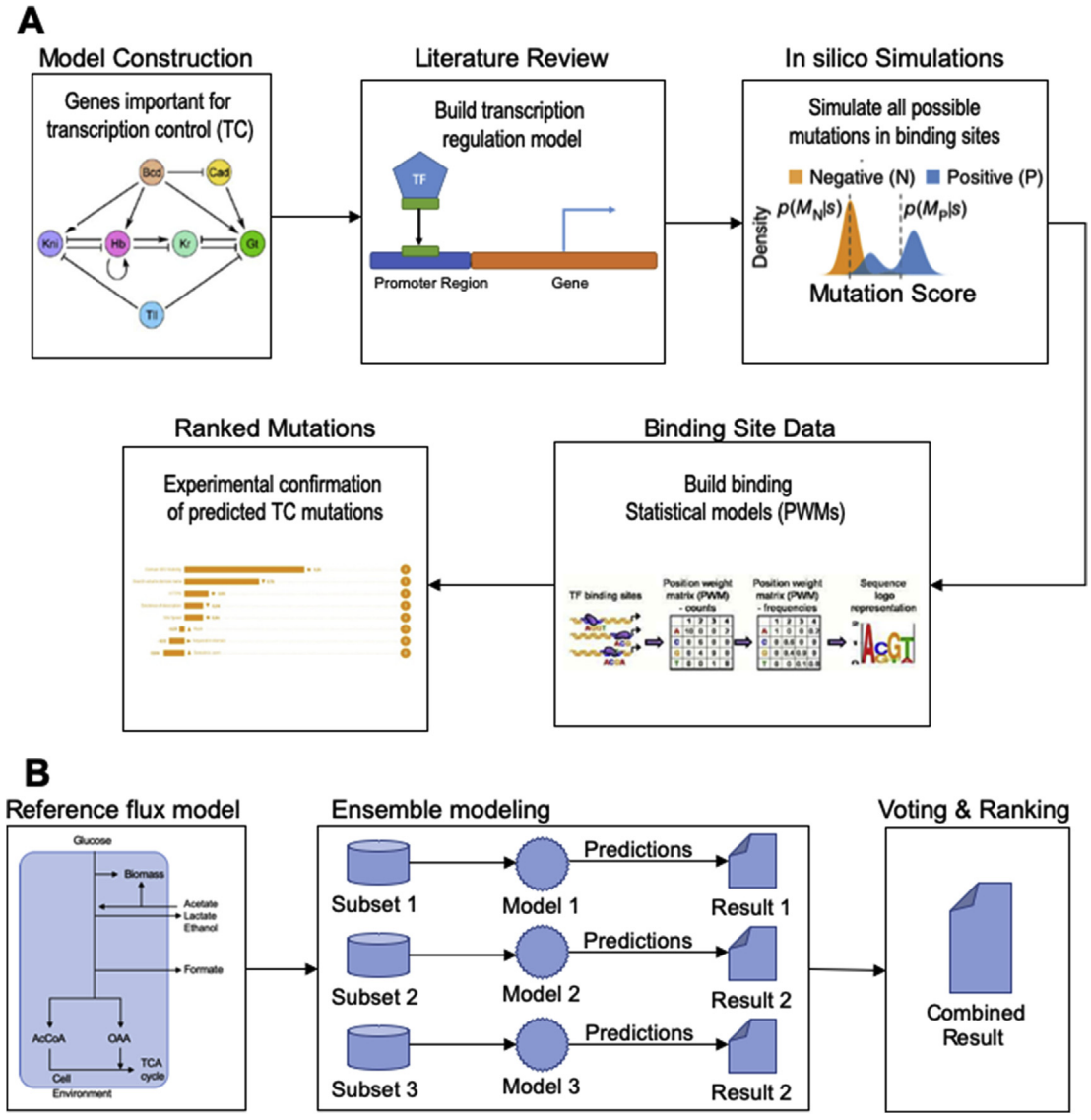

Fig. 3. Schematic representation of different modeling approaches used in metabolic engineering. Modeling steps of A) promoter-strength simulations using statistical models and mutations data, B) Ensemble modeling combining different sub-model simulations.

(Kotu and Deshpande, 2015) (Fig. 3B). This strategy has been employed in metabolic engineering to create large-scale models that predict the outcome of engineered pathways, by allowing the simulation of the network change upon perturbations, such as changes in growth conditions or in enzyme expression levels. This, therefore, waives the need of detailed kinetic parameters. The availability of perturbation data and its accuracy plays a crucial role in the development of the ensemble predictions, thereby, reducing the number of models to a smaller set (Tran et al., 2008).

An example of ensemble modeling was performed for two non-native central pathways for carbon conservation, the non-oxidative glycolysis (NOG) and the reverse glyoxylate cycle (rGC) pathways, using ensemble modeling robustness analysis (EMRA). EMRA successfully determined the probability of system failure and identified possible targets for flux improvement (Lee et al., 2014). In another study, ensemble modeling was used to help in developing a L-lysine-producing strain in *E. coli* (Contador et al., 2009). Nevertheless, ensemble modeling come with some major challenges. Building an ensemble with different modeling algorithms is more difficult than using any standard modeling strategy. The requirement of perturbation-response data makes it similar to many other data-dependant modeling strategies that perform poorly in the absence of reliable data, and the difficulty in interpreting its overall results. These limitations hinder the utility of this powerful modeling approach.

2.3. Protein modeling for metabolic engineering

Another widely used modeling approach for metabolic engineering is *in silico* three-dimensional (3D) molecular modeling for the study of receptor/enzyme–ligand docking and protein homology design (Wang et al., 2018). It has a wide range of applications in drug design and metabolism research, and therapeutic antibodies design and molecular interactions research (protein-protein and protein-DNA interactions). In metabolic engineering, 3D modeling is used to design, and simulate engineered enzymes that are indispensable for the optimization process of the microorganism's metabolism (Fisher et al., 2014). In protein engineering, where no structural data is available, molecular modelling is used to model the 3D-structures of enzymes, and coupled with enzyme-substrate docking studies, can be used to target regions of interest to improve various attributes, such as specificity, activity and stability under a given environment. This has been used to great effect for single enzymes as *in vitro* industrial biocatalysts (e.g. sitagliptin (Savile et al., 2010)), as well as for entire enzyme cascades (e.g. islatravir (Huffman et al., 2019)) for the production of active pharmaceutical ingredients.

2.4. Limitations of current modeling strategies

Dynamic modeling strategies, as mentioned above, often depend on the parameters that are used to build the model. The parameters (such as reaction kinetics or flux ranges) can be determined using bottom-up or top-down approaches (Cuperlovic-Culf, 2018). The bottom-up approach is highly dependent on experiments (such as *in vitro* enzymatic assays) since it requires information on the reaction kinetics of each enzyme, which is highly challenging to determine for all the enzymes in a pathway or network. Furthermore, even if information is obtained from *in vitro* experiments, the data are often several orders of magnitude different from actual *in vivo* experiments (Selvarajoo et al., 2009). Moreover, modeling usually requires data (kinetics or flux rates) for multiple conditions or time points to train the model and test its accuracy or applicability, which requires iterative experimental work (Helmy et al., 2009). Despite the fact that the bottom-up modeling approaches often use optimization algorithms to estimate the model parameters, such as the genetic algorithm, the complex and non-linear nature of the relationships between metabolites limit the usefulness of the model fitting algorithms (Cuperlovic-Culf, 2018; Srinivasan et al., 2015).

Another aspect of limitations is the scale of the model. Since the bottom-up approach requires detailed experimental measurements, it is more suitable for small-scale models. Extending the model size requires either more experiments (higher cost and longer time) or more computational estimation reliance of the parameter values (lower accuracy). Thus, an accurate dynamic model based on a bottom-up approach is difficult to establish due to the extended level of uncertainty in the kinetic properties of the enzymes and their reactions (Andreozzi et al., 2016). Ensemble modeling helps in building large-scale models, however, it also suffers from major limitations as mentioned earlier.

On the other hand, top-down approaches utilize time series metabolomic data to indirectly infer the kinetics, flux rates or concentrations of metabolites, through the establishment of correlation and causation networks between metabolites (Cuperlovic-Culf, 2018). The causation network establishes the cause-effect relationships between the metabolites in the networks and is usually built using time series metabolomic data, while the correlation network uses mathematical and statistical methods to determine the probable relation between the enzymes and metabolites in the network (Srinivasan et al., 2015). Most of the top-down methods utilize optimization algorithms, such as genetic algorithms and evolutionary programming, to estimate the model parameters based on the available experimental data. However, the complex (and usually non-linear) relationships, within the metabolic models and the heterogeneous nature of its parameters (e.g., kinetic parameters, concentrations) limits the capacities of the fitting algorithms (Cuperlovic-Culf, 2018).

Nevertheless, the top-down approach has shown notable success in analyzing cellular pathways with simple linear response or mass-action kinetic models with little parameter sensitivity (Selvarajoo, 2011, Selvarajoo et al. 2009).

For the comparative 3D protein modelling, it is most commonly performed using template-based methods, where homologous protein structures are used to generate models using stand-alone programs such as MODELER (Šali and Blundell, 1993) or through online servers such as ROBETTA, which incorporates the RosettaCM method (Song et al., 2013), HHPRED (Zimmermann et al., 2018), and ITASSER (Yang et al., 2014). These methods produce useful models where good templates are available, but many protein sequences of interest have limited template information, and so poor-quality models are common which hinders their practical applications in guiding protein engineering works.

Most of the above-mentioned modeling strategies require the availability of sufficient and high-quality experimental data. The data includes metabolite concentrations, and their chemical structures, properties, pathways, reaction rates, genomic sequences, genome annotations, transcriptome sequence, gene expression data and many other types of data, as required for their respective modeling strategies. Fortunately, a large number of bioinformatics databases and servers are now freely available with most of these data. Many of them are meta-databases that collect and aggregate data from multiple sources such as KEGG and MetaCyc (Kanehisa, 2004; Caspi et al., 2020). Despite the benefits of these bioinformatics resources, the challenge is in finding the correct dataset and modeling/analytical approaches to take advantage of this wealth of data. This, therefore, raises the need of the involvement of novel data mining and data analytics approaches, such as artificial intelligence (AI).

3. Integrating artificial intelligence in metabolic engineering research

Artificial intelligence (AI) provides computers the ability to make decisions based on analyzing the data independently by following predetermined rules or pattern recognition models. In the biomedical and biotechnology fields in particular, AI is heavily employed in addressing certain research challenges while being under-utilized in other aspects. The drug and vaccine discovery fields, for instance, are employing AI to address the challenges of developing new drugs, repurposing existing

drugs, understanding drug mechanisms, designing and optimizing clinical trials and identifying biomarkers (Smith, 2020a). Recent surveys show that more than 40 pharma companies and 230 startup companies are employing AI in different aspects of drug discovery (Smith, 2020b, 2020c). This has resulted in the development of over one hundred drugs that are in different development phases in the fields of oncology, neurology and infectious diseases (Smith, 2020d). Furthermore, the research on COVID-19 drugs and vaccination development is employing AI, and this has resulted in dozens of promising drug lead compounds and vaccines in such a short period of time (Regulatory Affairs Professionals Society, 2020; Ledford, 2020). AI is also employed in the fields of genomics, protein-protein interaction prediction, signaling pathways prediction and analysis, protein-DNA binding, cancer diagnosis, and genomic mutation variant calling among several other applications (Alipanahi et al., 2015; Hui et al., 2013; Poplin et al., 2018).

On the other hand, AI is not similarly utilized in the fields of metabolomics and metabolic engineering, especially for food applications. Although the idea of combining systems biology and AI (machine learning in particular) to study metabolism is relatively old (Zelezniak et al., 2018), the applications of it is still under explored.

Machine learning (ML) is the field of AI that is interested in developing computer programs that learn and improve its performance automatically based on experience and without explicitly being. In the last few years, ML research and techniques have improved as large datasets generated by modern analytical lab instruments become available. Therefore, in recent reports we are starting to see ML-based research in identifying weight loss biomarkers (Dias-Audibert et al., 2020), the discovery of food identity markers (Erban et al., 2019) farm animal metabolism (Ghaffari et al., 2019) and many other applications in untargeted metabolomics (Heinemann, 2019; Liebal et al., 2020). In metabolic engineering, several recent articles review the application of ML in the biosystems design and microbial bio-manufacturing (Volk et al., 2020; Choi et al., 2019). In the following sections, we review the advantage of ML and systems biology integration in pathways discovery and analysis, identifying essential enzymes, modeling of metabolisms and growth, genome annotation, the analysis of multi-omics datasets and 3D protein modeling.

3.1. Machine learning for pathway discovery

Pathways identification and analysis is very crucial for metabolic engineering. It is common that the biochemical pathway of a targeted substance (e.g. enzyme or compound) is unknown or poorly studied. Furthermore, in many cases, the gene(s) or gene cluster that is responsible for producing the targeted substance needs to be transferred to a model organism so that it can be easily manipulated and optimized (García-Granados et al., 2019). As mentioned above, the different modeling techniques have their limitations. On the other hand, when combining omics data and using standard data analysis approaches for pathways the final predictions come with its uncertainty (Cheng et al., 2015).

ML can be utilized to identify the pathways upstream of the substance. For instance, ML model that used naive Bayes, decision trees, logistic regression and pathway information of many organisms were used in MetaCyc to predict the presence of a novel metabolic pathway in a newly-sequenced organism. The analysis of the model performance showed that most of the information about the presence of a pathway in an organism is contained in a small set of used features. Mainly, the number of reactions along the path from input to output compound was the most informative feature (Cuperlovic-Culf, 2018). In general, the ML models used for pathway prediction showed better performance than the standard mathematical and statistical methods (Quest et al., 2010). Nevertheless, pathway discovery is still heavily relying on traditional approaches such as gene sequence similarity and network analysis. Thus, better ML algorithms/methods for pathways discovery are needed.

3.2. ML for identifying essential enzymes

ML can be invaluable for the identification of important genes or enzymes in the pathways of interest. ML classifiers, such as support vector machine, logistic regression and decision tree-based models, have been instrumental in predicting gene essentiality within metabolic pathways through training and testing models (by using labeled data of essential and non-essential genes) (Nandi et al., 2017). It was also used in finding new drug targets by determining the essential enzymes in a metabolic network of each enzyme by its local network topology, co-expression and gene homologies, and flux balance analyses (Plaimas et al., 2008). Plaimas *et al* used an ML model that was trained to distinguish between essential and non-essential reactions, which followed an experimental validation using the phenotypic outcome of single knockout mutants of *E. coli* (KEIO collection). The model was used for error detection to validate experimental data. When the predictions contradict the KEIO collection, they indicate errors in the experimental data. Subsequently, the model prediction were experimentally validated (Plaimas et al., 2008).

In an earlier study, the side effects of drugs on the metabolic network were investigated by predicting an enzyme inhibitory effect through building an ML model. The model used network topology, functional classes of inhibitors and enzymes as background knowledge, with logic-based representation and a combination of abduction and induction methods to predict drug inhibitory side effects. The abduction was used to generate hypotheses based on ground facts about the inhibited enzymes (ground hypotheses), while the induction process is to learn general rules of enzyme inhibition (non-ground hypotheses). The model simulations show that in the presence of sufficient training data, the non-ground hypotheses show better predictive accuracy (Tamaddoni-Nezhad et al., 2006).

3.3. ML for genome annotation

Newly sequenced genomes undergo two types of annotations; structural annotation and functional annotation. The structural annotation is the process of identifying the genome components and their structures (e.g. identifying genes, their exons, introns and UTRs or their regulatory regions), while the functional annotation identifies the functions of the genes and their products. Both types of annotation are important for metabolic engineering research; the structural annotation identifies the genes, their sequences, length and structure and, therefore, helps in finding alternative organisms where the same gene, pathways or gene clusters exist. The functional annotation helps in identifying organisms that produce the same substance or tolerate the same growth conditions. Comparative genomics, network biology and traditional bioinformatics methods, such as sequence alignment, are usually utilized in this process (Bradbury et al., 2013; Ikeda et al., 2014).

The rapid advancements in the genome sequencing technologies and the significant drop in its cost in the last decade raised the advantage for fast and accurate annotation methods (El-Metwally et al., 2014). This resulted in the development of several new annotation methods that analyse the newly sequenced genomes from different sequencing platforms that addressed many of the challenges, however, many other challenges remain such as missing short genes and erroneous exon start and end annotation (Armstrong et al., 2019; Li et al., 2019). Thus, several other methods were introduced with the idea of combining multi-omics data in the process of the genome annotation and, in particular, the proteomic and transcriptomic data (Armengaud, 2009; Ang et al., 2019; Helmy et al., 2012). Despite these efforts, over 20% of the sequenced genomes in the genome online database (GOLD) are still awaiting annotation (Mukherjee et al., 2017).

The high-volume and multi-dimensional nature of the genome sequencing data makes it very suitable for applications of machine learning algorithms (Yip et al., 2013). The ML model will be trained using annotated genomes to identify genome structures, e.g. genes or

regulatory regions, using their features to identify the same structures in the newly sequenced genomes (Alpaydin, 2020). Yip *et al* reviewed over 15 different ML methods developed to identify several types of structural components such as protein-coding genes, non-coding RNAs (ncRNAs), microRNAs (miRNAs), regulatory elements and protein-binding sites/-motifs (Yip et al., 2013). More recent reports show the utilization of ML algorithms in genome annotation process by including multi-omics data; building successful large-scale models became possible through the incremental expansion of the model architecture, the iterative training process and the richness of the data, which allow some relaxation in the initial restrictions in the model parameters (Borodovsky, 2019).

Amin *et al* demonstrated the potential of deep learning in genome annotation by using Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to develop DeepAnnotator, an annotation tool that outperformed the NCBI annotation pipeline in RNA genes annotation (Amin et al., 2018). The new versions of the annotation tool GeneMarks for annotation prokaryotic genome (GeneMarkS2+) and the eukaryotic self-training gene finder (GeneMark-EP+) both are utilizing ML algorithms in the annotation process (Borodovsky, 2019; Lomsadze et al., 2018). Deep convolutional neural networks were used to annotate gene-start sites in different species by training the model using the sites from one species as the positive sample and random sequences from the same species as the negative sample. The model was able to identify gene-start sites in other species (Khodabandelou et al., 2020).

Although the idea of employing ML in functional annotation started relatively early, it is still underutilized in functional annotation compared to structure annotation. An early attempt of using ML in genes functional annotation from biomedical literature utilized Hierarchical Text Categorization (HTC) (Kiritchenko et al., 2020), while Tetko *et al* provided a high-quality curated functional annotation data as a benchmark dataset for the developers of machine ML-based functional annotation methods for bacterial genomes (Tetko et al., 2005). The recent reports show the applications of ML-based methods in a wide variety of functional annotations such as the discovery of missing or wrong protein function annotations (Nakano et al., 2019), predicting gene functions in plant (Mahood et al., 2020), controlling the false discovery rate (FDR), increase the accuracy of protein functional predictions (Hong et al., 2020), and genome-wide functional annotation of splice-variants in eukaryotes (Panwar et al., 2016).

3.4. ML of multi-omics datasets

The advancements of -omics technologies have resulted in a huge accumulation of data (genomics, transcriptomics, proteomics and metabolomics) that is estimated to grow in size to exceed astronomical levels by 2025 (Stephens et al., 2015). This enormous amount of data has shifted scientific research more towards data-driven approaches such as ML (Cuperlovic-Culf, 2018). Combining ML methods with omics data is a typical systems biology approach to address several biomedical challenges. An ML approach was used to replace the traditional kinetic models in estimating the metabolite concentrations over time by combining ML models, proteomic and metabolomic time series data. This ML approach leverages arbitrary chunks of new data systematically to improve predictions without assuming particular interactions, instead it chooses the most predictive ones. This new approach produces qualitative and quantitative predictions that outperformed the classical kinetic model (Costello and Martin, 2018). Also, proteomic and metabolomic data of yeast were combined under several perturbation conditions (97 kinase knockouts), and ML was used to predict the yeast metabolome using the enzyme expression proteome of each kinase-deficient condition. The ML quantifies the role of enzyme abundance through mapping the regulatory enzyme expression patterns then utilizing them in predicting the metabolome under the knockout condition (Zelezniak et al., 2018).

The availability of transcriptome data and the ability of ML methods to deal with big data led to the development of several genome-scale methods to predict the phenotype using ML models. To take advantage of the accumulated transcriptome data, a biology-guided deep learning system named DeepMetabolism was developed. DeepMetabolism uses transcriptomics data to predict cell phenotypes. It integrates unsupervised pre-training with supervised training to predict the phenotype with high accuracy and high speed (Guo et al., 2017). On the other hand, Jervis *et al* implemented an ML algorithm to model the bacterial ribosome binding sites (RBSs) sequence-phenotype relationship and accurately predicted the optimal high-producers, an approach that directly apply on a wide range of metabolic engineering applications (Jervis et al., 2019).

Similar to the proven utility of ML in the analysis of the transcriptomic data, it is also used with the accumulated fluxomic data that descripts the complete set of metabolic fluxes in a living entity. For instance, MFlux is a tool that predicts the bacterial central metabolism utilized machine learning to for mining the existing fluxomic data to identify the hidden relationships between environmental and genetic factors and metabolic fluxes. To study the complex relationship between controlling factors and metabolic fluxes, Mflux used three machine learning methods, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Tree (DT). The tool provides predictive models that significantly accelerate flux quantification and a web-based platform that predicts the bacterial central metabolism via machine learning via graphical user interface (GUI) (Wu et al., 2016a). ML approaches were also integrated with transitional genome-scale modeling to for evaluate the microbial factory performance. The model used manually curated databases of over 1,200 experimentally realized *E. coli* cell factories enriched with genetic modifications and bioprocess variables. The simulations from a standard genome-scale metabolic models were used to add additional features. Consequently, ensemble modelling was used to improve dealing with data challenges. This hybrid approach predicted the *E.coli* factory performance with high cross-validation accuracy (Oyetunde et al., 2019).

3.5. ML in protein modeling

In the field of 3D protein modeling, several AI-based advances are also noted. The most recent Critical Assessment of protein Structure Prediction (CASP) meeting in 2018 saw AI methods come of age. The program AlphaFold (Senior et al., 2020) used a neural net to extract covariant residue pairs from sequence alignments, coupled with estimated distances between them (from 2-20Å), and then used the ROSETTA energy function (Alford et al., 2017) to fold the protein based on these AI-derived restraints. AlphaFold performed exceptionally well in the competition, giving high-accuracy models with template-modelling scores of 0.7 or higher for 24 out of 43 domains (as compared with 14/43 for the next best method). This has been developed into a lab-based version called ProSPr (Billings et al., 2019). Yang *et al* used a similar protocol, but with added estimation of relative residue orientations, resulting in trROSETTA (Yang et al., 2020), which improved predictions still further. These 3D modelling methods may be implemented into a comprehensive metabolic engineering platform.

One area that could be addressed in the improvement of 3D protein modelling methods is the inclusion of cofactors. Many enzymes are often folded around cofactors; small-to-large organic molecules which form part of the catalytic machinery, such as flavin adenine dinucleotide (FAD) or haem. These molecules are often removed in template-based modelling (both manual and automated versions), yet their presence is often important for the correct folding of the enzyme (Higgins et al., 2005). This has the effect of lowering the quality of the model due to the removal of key restraints from the structure, requiring extra docking or structure manipulation to reinsert the cofactor after modelling. It should be possible to include the presence of cofactors through a survey of the Protein Data Bank (Berman et al., 2002), where ML methods can be used to identify key determinants of cofactor binding, coupled with identification of these determinants within a target sequence, and application of

a combined sequence-and-template-based optimization protocol inclusive of these structural features.

An extension of this might also be used for identification of substrates for enzymes within a metabolic pathway or unnatural substrates which is particularly valuable for the development of synthetic biosynthetic pathways. One input would be enzyme sequence alignments of known function, as well as structural information for both enzyme families and substrates. A neural network could be used to identify common patterns of binding pocket residues across multiple families of enzymes for different substrates, and identify potential sequences that would be suitable for inclusion in a particular metabolic pathway, inclusive of sequence determinants for ease of inclusion into heterologous expression systems. Also, if no sequence is available that produces a required product, it might be possible to predict the binding pocket residues that might be mutated to give that product. Predictions made can then be experimentally tested, and results fed back into the model.

3.6. Application of ML models for engineering bio-economy strains

In recent years, the importance of harnessing natural and food ingredients from diverse sources is increasingly realized, such as using engineered microbes or synthetically derived as highlighted in the introduction section. These approaches provide several benefits for producing a more sustainable bio-based economy that relies less on precious land or limited livestock. Nevertheless, the bioengineering processes utilized still remain suboptimal, due to the complexity of living systems' emergent behaviors (such as feedback/feedforward inhibition, cofactor imbalances, toxicity of intermediates, bioreactor heterogeneity) that tend to reduce the overall effect of any internal modifications such as adding or engineering a metabolic pathway (Yadav et al., 2012; Lim and Kim, 2019). Thus, achieving economically viable large-scale production of microbial-derived metabolites or compounds requires appropriately optimized production strains that generate high yields. Until today, however, metabolic engineering efforts mainly serve for broadening and further reducing the cost of those molecules of commercial interests.

To address these issues, Brunk *et al* engineered eight *E. coli* lab strains that produced three commercially important biofuels: isopentenol, limonene, and bisabolene (Brunk et al., 2016). To understand the key regulatory or emergent bottleneck scenarios that limit their industrial applicability, they undertook a large scale -omics based systems biology approach where they performed time-series proteomics and metabolomics measurements, and analyzed the resultant high-throughput data using statistical analytics and genome-scale modeling. The integrated approach revealed several novel key findings. For example, they elucidated time-dependent regulation of gene, protein and metabolic pathways related to the TCA cycle and Pentose-Phosphate pathway, and the resultant coupling of the pathways that affected NADPH metabolism. These emergent responses were collectively implicated to downregulate the expected biofuel production. The findings, subsequently, led them to identify a crucial gene (*ydbK*) whose removal led to a 2-fold increase in the production of isopentenol in one of the *E. coli* strains (Brunk et al., 2016).

Despite their success on one strain (out of eight), the overall dynamic changes of metabolic pathways at the different stages of growth for all strains were not understood, as they employed a steady-state genome-scale model, which provided a qualitative, rather than quantitative, inference. This, as mentioned earlier (in *Dynamic Modeling*), is due to the lack of kinetic parameter values that are required to develop and test a dynamic model for each strain. To overcome this difficulty, Costello and Martin (2018) used the same time-series proteomics and metabolomics data of Brunk *et al* and developed a ML model to effectively predict pathway dynamics in an automated fashion (Costello and Martin, 2018). Their model produced both qualitative and quantitative predictions that had better predictions compared to a traditional kinetic model side-by-side. Basically, their ML model derived a mapping function between the proteomics and metabolomics dataset with the aid of regression techniques and neural networks onto a training data, and finally verifying the prediction on a test data. Apart from better accuracy in the dynamic profiles of the metabolites predicted, the model also did not require detailed understanding of the regulatory steps, which is a major weakness for all modeling approaches. However, their ML model was short in predicting effective regulator(s) for enhanced production of any of the biofuels (isopentenol, limonene, and bisabolene), nor was there any experimental verification. Although this is a major weakness in current systems metabolic engineering approaches, ML-based modeling has the future potential to productively guide bioengineering strains without knowing complete metabolic regulatory processes, which are very challenging to obtain.

One interesting and popular area of industrially relevant metabolic engineering product in the food and consumer care industries are the terpenes and terpenoids; secondary metabolites or organic compounds naturally found in diverse living species, especially in plants. Due to their high commercial values, numerous researches have focused on producing them or their derivatives at industrial scale using microbes (Caputi and Aprea, 2011; Zhang et al., 2020b). Although several hundreds, or even thousands, of fold increase has been achieved at test tube or flask level by engineering microbes (Czajka et al., 2018), the achievement at large industrial scale bioreactors are far from reality. It is our opinion that ML models can help to uncover the relations between output and input more accurately, and identify sweet spots for carefully targeted steps for generating bioreactor scale targeted output. Although there is no current workable evidence for this, we believe the future looks promising for this front, provided large investments are made to generate biological data that are required by dynamic or ML models to effectively be predictive.

4. Challenges and future projections

Integrating systems biology and ML holds a great promise for improving the way we study and understand metabolism (Fig. 4), as well as to improve and engineer alternative food sources that are healthier, affordable and nutritious. However, as reviewed in this article, this integration faces several challenges and limitations in order to fully utilize the power of both systems biology and ML.

A major challenge that faces the application of systems biology and ML in food-grade or GRAS metabolic engineering is the lack of data. Systems biology requires high throughput data from multi-omics levels (genomic, transcriptomic, proteomic and metabolomic), and such data is only available for a small subset of microorganisms in general, and significantly lacking for the food-grade or GRAS strains, in particular. The availability of such data is necessary for more holistic studying of the organism and helps in discovering new pathways or proteins, simpler, shorter directed pathways or new enzymes with better production rate (García-Granados et al., 2019). This information will also help in choosing the most appropriate organism to be used for the engineering and production projects. Usually, certain model organisms called "chassis" such as yeast and *E. coli* are used in these projects where the gene(s) or pathways of the substance of interest is transferred from the donor organism. However, the availability of sufficient information about both the donor organism and the chassis help choosing the correct chasse and avoid facing unexpected qualities such as resistance to certain conditions or missing of important pathways (Khoury et al., 2014).

In addition to the need of large-scale -omics data for building ML models, another data problem is facing the application of ML in the metabolic engineering research. Training an ML model for metabolic engineering requires sufficient quantitative data for multiple conditions. The multiple conditions can be multiple knockouts, perturbations or growth conditions. For instance, to build an ML model that predict the required engineering (e.g. knockouts) to improve the promoter strength, we need to train the model using quantitative data of the downstream gene expression under multiple knockouts or mutants. This becomes more challenging with the differences observed in different cultivation scales (Zhang et al., 2015). Usually, the behavior of the microorganisms

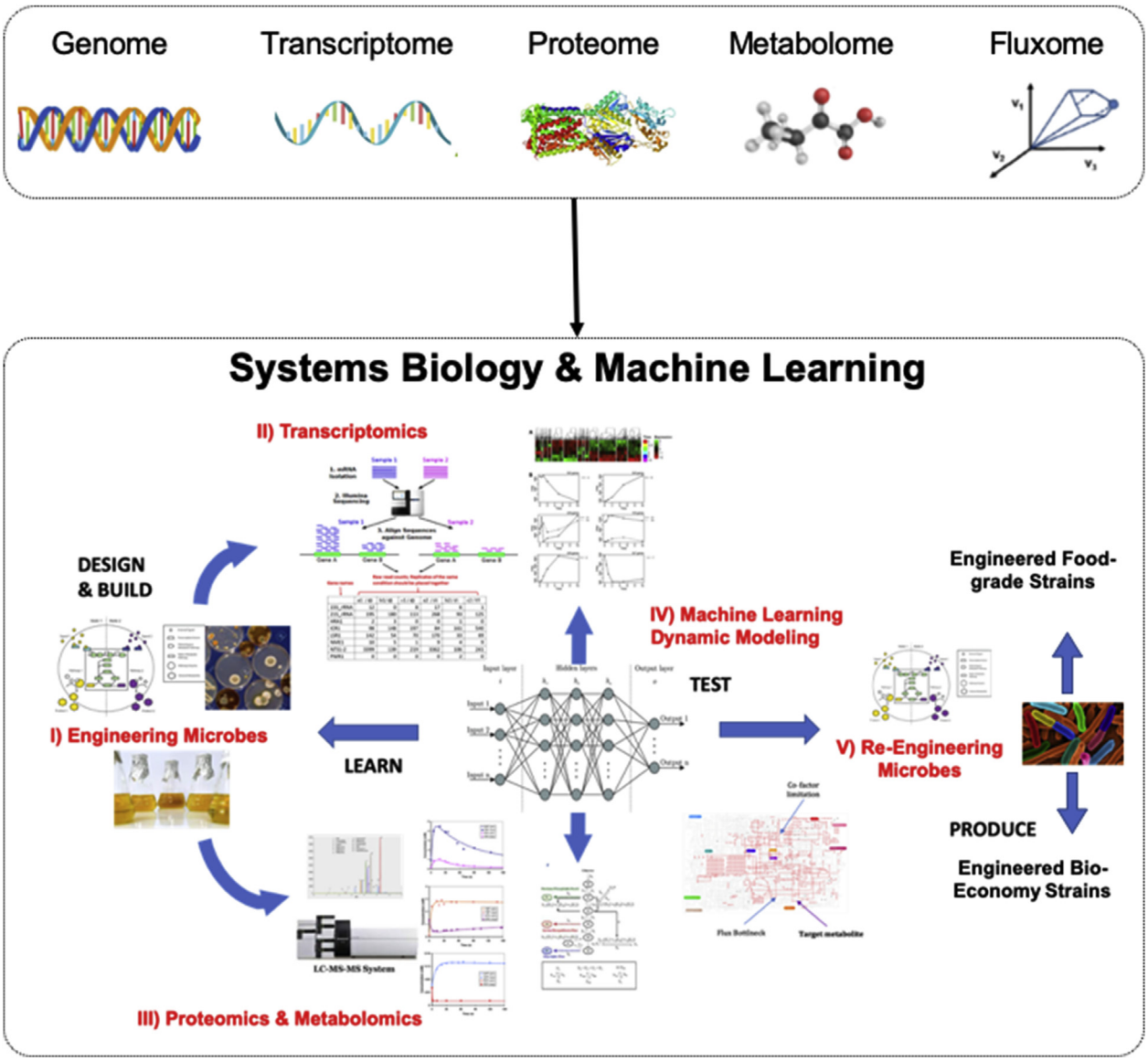

Fig. 4. Integrating systems biology and machine learning in metabolic engineering research. Systems biology and ML approaches are highly suitable for processing and analyzing multi-omics data with massive sizes and features. Starting from an initial strain and design (i), transcriptomics (ii), proteomics and metabolomics data generation (iii) provide multitudes of data which require integration by data analytics, modelling and machine learning (iv). This will help provide targets for re-design/-engineering which need to be experimentally tested (v). This will lead to an enhanced engineering process to generate engineered microbes that can be used in the modern bio-industries such as food industry.

changes when scaling up the cultivation form the lab setup (small bioreactors usually 1 L or less) to the industrial scale bioreactors (up to 2, 000 L). These differences come from the differences between the lab and industrial bioreactors in several factors including the lighting systems, mixing, gas transfer, and the bubbles hydrodynamics (Brennan and Owende, 2010; Johnson et al., 2018). Thus, data generated from lab setup will not be suitable for modeling the dynamics of large-scale cultivation setup. Similarly, the predictive ML models investigating the translation control, transcription factor binding sites, ribosomal binding sites, enzyme engineering (mutation or truncation) and growth optimization require high quality quantitative data in multiple conditions. The same data can also be used in building different mathematical and statistical models, which allows the development of more integrated methods.

It is notable that several systems biology databases for AI application are available (such as the LASER database, jQMM database and Kbase) (Winkler et al., 2015; Birkel et al., 2017; Arkin et al., 2018), however, it is still difficult to find a suitable data for building AI models for metabolic engineering online and it needs to be created for each project. We need more research that focus on the generation of high-quality quantitative data, and on building online resources, such as meta databases, that collect and combine these data to make it available for the community. The data should also include details on the experimental conditions and follow the standards of biological databases design (Helmy et al., 2016). Many of the current online resources are missing these details which prevent or limit the application of data miming and machine learning methods on their contents (Wu et al., 2016b).

Another major challenge in the ML field is what is known as “the black box problem”. The black box problem of AI techniques in general is defined as the difficulty of understanding how they work and how and why they give these results (Zednik, 2019). This causes the end user of the technique to be uncertain about the quality of the output, and the often biologically unfamiliar modeler will not be able to intervene to improve the performance as well as raising some legal concerns (Rieder and Simon, 2017; MT Ribeiro, 2016; Burrell, 2016). For example, in the application of ML in 3D structural modelling, as well as enzyme-substrate

identification, the newer AI-based modelling methods are showing some promising results, however, due to the nature of neural nets, it is very difficult to interpret exactly what the programs are learning about the protein-folding problem. We can predict a structure, but without understanding the underlying model for folding. If a way could be found to capture this information, it would be of great use to the community for further study. To address the black box problem, scientists in the field of AI developed a group of AI methods called explainable artificial intelligence (XAI) that aim to make the results of AI methods understandable to humans. Although this is still new, it holds potential to solve the problems that prevent the systematic performance improvement of AI models (Zednik, 2019; Dosilovic et al., 2018).

On the other hand, although genome annotation, both structural and functional, affects most of the biomedical research aspects, it has a special impact on metabolic engineering in general and applications in food industry in particular. The food-grade or GRAS microorganisms are a small subset of all organisms, and many of them are either not well-studied, or have not been studied at all. Hence, there is a big challenge in using these species in ML-based metabolic engineering, as many of them are either not sequenced or sequenced with draft annotation or with no annotation. The annotations are usually automated using standard pipelines which identify the common genes that they share with other microorganisms and can miss the organism-specific features that need deeper attention. These features are exactly what make those organisms suitable for metabolic engineering and food industry. Improved ML-based genome annotation methods will help improving the annotation of the food-safe and GRAS genomes which will directly impact the research in this area.

Another area that needs special attention is pathways prediction in the absence of genome sequence or genome annotation. Since many of the food-safe and GRAS microorganisms are not sequenced yet, methods that predict the pathways for important substances using different -omics data is required. It is easy now to perform whole- or phospho-proteomics, and transcriptomics across different growth conditions or different life stages of an organism. These -omics data can be used, in the absence of genome sequence, to predict the endogenous or biosynthetic pathways of the substance of interest. Here, ML methods can be used instead of the traditional pathway prediction approach due its better suitability to the nature and size of the data.

Overall, despite the challenges and limitations of AI or ML techniques in dealing with biological datasets, there is no better time than now to explore the full potential of these techniques and to further develop novel methods to overcome the many challenges, including “the black box problem”. This is especially so since we know living systems are highly complex, and using physical or biochemical theories alone may not be sufficient to explore all complexities. Thus, heuristic approaches and ML can and will play a crucial to support all future systems biology efforts. In parallel, the improvements to the data collection from -omics technologies in time will help to narrow the gap of uncertainty or ambiguity for future systems biology and ML integration for optimal metabolic engineering strategies.

Acknowledgment

The authors thank Simon Zhang Congqiang for critical comments, and the Singapore Institute of Food and Biotechnology Innovation (SIFBI, A*STAR) for supporting this work (under IAFPP3 - H20H6a0028).

References

Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., Labonte, J.W., Pacella, M.S., Bonneau, R., Bradley, P., Dunbrack, R.L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., Gray, J.J., 2017. The rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theor. Comput. 13, 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838. https://doi.org/10.1038/nbt.3300.

Alpaydin, E., 2020. Introduction to Machine Learning. MIT press.

Amin, M.R., Yurovsky, A., Tian, Y., Skiena, S., 2018. DeepAnnotator: genome annotation with deep learning. Comput. Biol. Heal. Informatics. 18 https://doi.org/10.1145/3233547.3233577.

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. https://doi.org/10.1016/j.ymben.2015.10.002.

Ang, M.Y., Low, T.Y., Lee, P.Y., Wan Mohamad Nazarie, W.F., Guryev, V., Jamal, R., 2019. Proteogenomics: from next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. Clin. Chim. Acta 498, 38–46. https://doi.org/10.1016/j.cca.2019.08.010.

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., Sneddon, M.W., Henderson, M.L., Riehl, W.J., Murphy-Olson, D., Chan, S.Y., Kamimura, R.T., Kumari, S., Drake, M.M., Brettin, T.S., Glass, E.M., Chivian, D., Gunter, D., Weston, D.J., Allen, B.H., Baumohl, J., Best, A.A., Bowen, B., Brenner, S.E., Bun, C.C., Chandonia, J.M., Chia, J.M., Colasanti, R., Conrad, N., Davis, J.J., Davison, B.H., Dejongh, M., Devoid, S., Dietrich, E., Dubchak, I., Edirisinghe, J.N., Fang, G., Faria, J.P., Frybarger, P.M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J., Haun, H.L., He, F., Jain, R., Joachimiak, M.P., Keegan, K.P., Kondo, S., Kumar, V., Land, M.L., Meyer, F., Mills, M., Novichkov, P.S., Oh, T., Olsen, G.J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S.S., Price, G.A., Ramakrishnan, S., Ranjan, P., Ronald, P.C., Schatz, M.C., Seaver, S.M.D., Shukla, M., Sutormin, R.A., Syed, M.H., Thomason, J., Tintle, N.L., Wang, D., Xia, F., Yoo, H., Yoo, S., Yu, D., 2018. KBase: the United States department of energy systems biology knowledgebase. Nat. Biotechnol. 36, 566–569. https://doi.org/10.1038/nbt.4163.

Armengaud, J., 2009. A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr. Opin. Microbiol. 12, 292–300. https://doi.org/10.1016/j.mib.2009.03.005.

Armstrong, J., Fiddes, I.T., Diekhans, M., Paten, B., 2019. Whole-genome alignment and comparative annotation. Annu. Rev. Anim. Biosci. 7, 41–64. https://doi.org/10.1146/annurev-animal-020518-115005.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C., 2002. The protein data bank. Acta Crystallogr. Sect. D Biol. Crystallogr. 58, 899–907. https://doi.org/10.1107/S0907444902003451.

Billings, W.M., Hedelius, B., Millecam, T., Wingate, D., Della Corte, D., 2019. ProSPr: democratized implementation of alphafold protein distance prediction network. BioRxiv 830273. https://doi.org/10.1101/830273.

Birkel, G.W., Ghosh, A., Kumar, V.S., Weaver, D., Ando, D., Backman, T.W.H., Arkin, A.P., Keasling, J.D., Martín, H.G., 2017. The JBEI quantitative metabolic modeling library (jQMM): a python library for modeling microbial metabolism. BMC Bioinformatics 18, 205. https://doi.org/10.1186/s12859-017-1615-y.

Bordbar, A., Monk, J.M., King, Z.A., Palsson, B.O., 2014. Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15, 107–120. https://doi.org/10.1038/nrg3643.

Borodovsky, M., 2019. New machine learning algorithms for genome annotation. In: Proc. 10th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics. Association for Computing Machinery (ACM), New York, NY, USA. https://doi.org/10.1145/3307339.3342128, 1–1.

Bradbury, L.M.T., Niehaus, T.D., Hanson, A.D., 2013. Comparative genomics approaches to understanding and manipulating plant metabolism. Curr. Opin. Biotechnol. 24, 278–284. https://doi.org/10.1016/j.copbio.2012.07.005.

Brennan, L., Owende, P., 2010. Biofuels from microalgae-A review of technologies for production, processing, and extractions of biofuels and co-products. Renew. Sustain. Energy Rev. 14, 557–577. https://doi.org/10.1016/j.rser.2009.10.009.

Brunk, E., George, K.W., Alonso-Gutierrez, J., Thompson, M., Baidoo, E., Wang, G., Petzold, C.J., McCloskey, D., Monk, J., Yang, L., O'Brien, E.J., Batth, T.S., Martin, H.G., Feist, A., Adams, P.D., Keasling, J.D., Palsson, B.O., Lee, T.S., 2016. Characterizing strain variation in engineered E. coli using a multi-omics-based workflow. Cell Syst 2, 335–346. https://doi.org/10.1016/j.cels.2016.04.004.

Burrell, J., 2016. How the machine ‘thinks’: understanding opacity in machine learning algorithms. Big Data Soc 3. https://doi.org/10.1177/2053951715622512, 205395171562251.

Caputi, L., Aprea, E., 2011. Use of terpenoids as natural flavouring compounds in food industry. Recent Pat. Food, Nutr. Agric. 3, 9–16. https://doi.org/10.2174/2212798411103010009.

Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D., 2020. The MetaCyc database of metabolic pathways and enzymes-a 2019 update. Nucleic Acids Res. 48, D445–D453.

Cheng, Y., Liu, J., Zhang, H., Wang, J., Zhao, Y., Geng, W., 2015. Transcriptome analysis and gene expression profiling of abortive and developing ovules during fruit development in hazelnut. PloS One 10, e0122072. https://doi.org/10.1371/journal.pone.0122072.

Choi, K.R., Jang, W.D., Yang, D., Cho, J.S., Park, D., Lee, S.Y., 2019. Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering. Trends Biotechnol. 37, 817–837. https://doi.org/10.1016/j.tibtech.2019.01.003.

Comba, S., Arabolaza, A., Gramajo, H., 2012. Emerging engineering principles for yield improvement in microbial cell design. Comput. Struct. Biotechnol. J. 3, e201210016 https://doi.org/10.5936/csbj.201210016.

Contador, C.A., Rizk, M.L., Asenjo, J.A., Liao, J.C., 2009. Ensemble modeling for strain development of l-lysine-producing Escherichia coli. Metab. Eng. 11, 221–233. https://doi.org/10.1016/j.ymben.2009.04.002.

Costello, Z., Martin, H.G., 2018. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. Npj Syst. Biol. Appl. 4, 19. https://doi.org/10.1038/s41540-018-0054-3.

Cuperlovic-Culf, M., 2018. Machine learning methods for analysis of metabolic data and metabolic pathway modeling. Metabolites 8, 4. https://doi.org/10.3390/metabo8010004.

Curran, K.A., Crook, N.C., Karim, A.S., Gupta, A., Wagman, A.M., Alper, H.S., 2014. Design of synthetic yeast promoters via tuning of nucleosome architecture. Nat. Commun. 5, 1–8. https://doi.org/10.1038/ncomms5002.

Czajka, J.J., Nathenson, J.A., Benites, V.T., Baidoo, E.E.K., Cheng, Q., Wang, Y., Tang, Y.J., 2018. Engineering the oleaginous yeast Yarrowia lipolytica to produce the aroma compound β-ionone. Microb. Cell Factories 17, 136. https://doi.org/10.1186/s12934-018-0984-x.

Dias-Audibert, F.L., Navarro, L.C., de Oliveira, D.N., Delafiori, J., Melo, C.F.O.R., Guerreiro, T.M., Rosa, F.T., Petenuci, D.L., Watanabe, M.A.E., Velloso, L.A., Rocha, A.R., Catharino, R.R., 2020. Combining machine learning and metabolomics to identify weight gain biomarkers. Front. Bioeng. Biotechnol. 8, 6. https://doi.org/10.3389/fbioe.2020.00006.

Dosilovic, F.K., Brcic, M., Hlupic, N., 2018. Explainable artificial intelligence: a survey. In: 2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc. Institute of Electrical and Electronics Engineers Inc., pp. 210–215. https://doi.org/10.23919/MIPRO.2018.8400040

El-Metwally, S., Ouda, O.M., Helmy, M., 2014. Challenges in the Next-Generation Sequencing Field. Springer, New York, NY, pp. 45–49. https://doi.org/10.1007/978-1-4939-0715-1_5.

Erban, A., Fehrle, I., Martinez-Seidel, F., Brigante, F., Más, A.L., Baroni, V., Wunderlin, D., Kopka, J., 2019. Discovery of food identity markers by metabolomics and machine learning technology. Sci. Rep. 9, 1–19. https://doi.org/10.1038/s41598-019-46113-y.

Fisher, A.K., Freedman, B.G., Bevan, D.R., Senger, R.S., 2014. A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories. Comput. Struct. Biotechnol. J. 11, 91–99. https://doi.org/10.1016/j.csbj.2014.08.010.

García-Granados, R., Lerma-Escalera, J.A., Morones-Ramírez, J.R., 2019. Metabolic engineering and synthetic biology: synergies, future, and challenges. Front. Bioeng. Biotechnol. 7, 36. https://doi.org/10.3389/fbioe.2019.00036.

Ghaffari, M.H., Jahanbekam, A., Sadri, H., Schuh, K., Dusel, G., Prehn, C., Adamski, J., Koch, C., Sauerwein, H., 2019. Metabolomics meets machine learning: longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis. J. Dairy Sci. 102, 11561–11585. https://doi.org/10.3168/jds.2019-17114.

Guo, W., Xu, Y., Feng, X., 2017. DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing. http://arxiv.org/abs/1705.03094. (Accessed 24 July 2020).

Heinemann, J., 2019. Machine learning in untargeted metabolomics experiments. In: Methods Mol. Biol. Humana Press Inc., pp. 287–299. https://doi.org/10.1007/978-1-4939-8757-3_17

Helmy, M., Gohda, J., Inoue, J., Tomita, M., Tsuchiya, M., Selvarajoo, K., 2009. Predicting novel features of toll-like receptor 3 signaling in macrophages. PloS One 4, e4661. https://doi.org/10.1371/journal.pone.0004661.

Helmy, M., Tomita, M., Ishihama, Y., 2012. Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics Metabolomics View project chi sequence View project. Genes, Genomes Genomics 6, 76–85. https://www.researchgate.net/publication/269574159. (Accessed 8 August 2020).

Helmy, M., Crits-Christoph, A., Bader, G.D., 2016. Ten simple rules for developing public biological databases. PLoS Comput. Biol. 12, e1005128 https://doi.org/10.1371/journal.pcbi.1005128.

Higgins, C., Muralidhara, B., Wittung-Stafshede, P., 2005. How do cofactors modulate protein folding? Protein Pept. Lett. 12, 165–170. https://doi.org/10.2174/0929866053005782.

Hong, J., Luo, Y., Zhang, Y., Ying, J., Xue, W., Xie, T., Tao, L., Zhu, F., 2020. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief. Bioinform. 21, 1437–1447.

Huffman, M.A., Fryszkowska, A., Alvizo, O., Borra-Garske, M., Campos, K.R., Canada, K.A., Devine, P.N., Duan, D., Forstater, J.H., Grosser, S.T., Halsey, H.M., Hughes, G.J., Jo, J., Joyce, L.A., Kolev, J.N., Liang, J., Maloney, K.M., Mann, B.F., Marshall, N.M., McLaughlin, M., Moore, J.C., Murphy, G.S., Nawrat, C.C., Nazor, J., Novick, S., Patel, N.R., Rodriguez-Granillo, A., Robaire, S.A., Sherer, E.C., Truppo, M.D., Whittaker, A.M., Verma, D., Xiao, L., Xu, Y., Yang, H., 2019. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. Science 366, 1255–1259. https://doi.org/10.1126/science.aay8484, 80-.

Hui, S., Xing, X., Bader, G.D., 2013. Predicting PDZ domain mediated protein interactions from structure. BMC Bioinformatics 14. https://doi.org/10.1186/1471-2105-14-27.

Ikeda, H., Shin-Ya, K., Omura, S., 2014. Genome mining of the Streptomyces avermitilis genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. J. Ind. Microbiol. Biotechnol. 41, 233–250. https://doi.org/10.1007/s10295-013-1327-x.

Jervis, A.J., Carbonell, P., Vinaixa, M., Dunstan, M.S., Hollywood, K.A., Robinson, C.J., Rattray, N.J.W., Yan, C., Swainston, N., Currin, A., Sung, R., Toogood, H., Taylor, S., Faulon, J.L., Breitling, R., Takano, E., Scrutton, N.S., 2019. Machine learning of designed translational control allows predictive pathway optimization in Escherichia coli. ACS Synth. Biol. 8, 127–136. https://doi.org/10.1021/acssynbio.8b00398.

Ji, Z., He, L., Rotem, A., Janzer, A., Cheng, C.S., Regev, A., Struhl, K., 2018. Genome-scale identification of transcription factors that mediate an inflammatory network during breast cellular transformation. Nat. Commun. 9 https://doi.org/10.1038/s41467-018-04406-2.

John, P.C. St, 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 https://doi.org/10.1371/journal.pcbi.1007424.

Johnson, T.J., Katuwal, S., Anderson, G.A., Gu, L., Zhou, R., Gibbons, W.R., 2018. Photobioreactor cultivation strategies for microalgae and cyanobacteria. Biotechnol. Prog. 34, 811–827. https://doi.org/10.1002/btpr.2628.

Kallscheuer, N., 2018. Engineered microorganisms for the production of food additives approved by the European Union-A systematic analysis. Front. Microbiol. 9, 1746. https://doi.org/10.3389/fmicb.2018.01746.

Kanehisa, M., 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32, 277D–280. https://doi.org/10.1093/nar/gkh063.

Khodabandelou, G., Routhier, E., Mozziconacci, J., 2020. Genome annotation across species using deep convolutional neural networks. PeerJ Comput. Sci. 6, e278. https://doi.org/10.7717/peerj-cs.278.

Khoury, G.A., Smadbeck, J., Kieslich, C.A., Floudas, C.A., 2014. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol. 32, 99–109. https://doi.org/10.1016/j.tibtech.2013.10.008.

Kim, O.D., Rocha, M., Maia, P., 2018. A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering. Front. Microbiol. 9, 1690. https://doi.org/10.3389/fmicb.2018.01690.

Kiritchenko, S., Matwin, S., Famili, A.F.. Functional annotation of genes using hierarchical text categorization. n.d. http://www.pdg.cnb.uam.es/BioLINK/. (Accessed 9 August 2020).

Kotu, V., Deshpande, B., 2015. Data mining process. In: Predict. Anal. Data Min. Elsevier, pp. 17–36. https://doi.org/10.1016/b978-0-12-801460-8.00002-1.

Laborde, D., Martin, W., Swinnen, J., Vos, R., 2020. COVID-19 risks to global food security. Science 369, 500–502, 80-.

Ledford, H., 2020. Dozens of coronavirus drugs are in development - what happens next? Nature 581, 247–248. https://doi.org/10.1038/d41586-020-01367-9.

Lee, Y., Lafontaine Rivera, J.G., Liao, J.C., 2014. Ensemble Modeling for Robustness Analysis in engineering non-native metabolic pathways. Metab. Eng. 25, 63–71. https://doi.org/10.1016/j.ymben.2014.06.006.

Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., Walters, J.R., 2019. Insect genomes: progress and challenges. Insect Mol. Biol. 28, 739–758. https://doi.org/10.1111/imb.12599.

Liebal, U.W., Phan, A.N.T., Sudhakar, M., Raman, K., Blank, L.M., 2020. Machine learning applications for mass spectrometry-based metabolomics. Metabolites 10, 243. https://doi.org/10.3390/metabo10060243.

Lim, H.J., Kim, D.-M., 2019. Cell-free metabolic engineering: recent developments and future prospects. Methods Protoc 2, 33. https://doi.org/10.3390/mps2020033.

Lomsadze, A., Gemayel, K., Tang, S., Borodovsky, M., 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. 28, 1079–1089. https://doi.org/10.1101/gr.230615.117.

Mahood, E.H., Kruse, L.H., Moghe, G.D., 2020. Machine learning: a powerful tool for gene function prediction in plants. Appl. Plant Sci. 8 https://doi.org/10.1002/aps3.11376.

McCloskey, D., Palsson, B., Feist, A.M., 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol. 9 https://doi.org/10.1038/msb.2013.18.

Mt Ribeiro, S.S.C.G., 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier, ArXiv. 1602, 04938v3.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., Thomas, A.D., Ali, R., Sharma, K., Kyrpides, N.C., Reddy, T.B.K., 2017. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res. 45, D446–D456. https://doi.org/10.1093/nar/gkw992.

Nakano, F.K., Lietaert, M., Vens, C., 2019. Machine learning for discovering missing or wrong protein function annotations. BMC Bioinformatics 20, 1–32. https://doi.org/10.1186/s12859-019-3060-6.

Nandi, S., Subramanian, A., Sarkar, R.R., 2017. An integrative machine learning strategy for improved prediction of essential genes in Escherichia coli metabolism using flux-coupled features. Mol. Biosyst. 13, 1584–1596. https://doi.org/10.1039/c7mb00234c.

Nozzi, N.E., Desai, S.H., Case, A.E., Atsumi, S., 2014. Metabolic engineering for higher alcohol production. Metab. Eng. 25, 174–182. https://doi.org/10.1016/j.ymben.2014.07.007.

Orth, J.D., Thiele, I., Palsson, B.O., 2010. What is flux balance analysis? Nat. Biotechnol. 28, 245–248. https://doi.org/10.1038/nbt.1614.

Oyetunde, T., Liu, D., Martin, H.G., Tang, Y.J., 2019. Machine learning framework for assessment of microbial factory performance. PloS One 14, e0210558. https://doi.org/10.1371/journal.pone.0210558.

Panwar, B., Menon, R., Eksi, R., Li, H.D., Omenn, G.S., Guan, Y., 2016. Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning. J. Proteome Res. 15, 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883.

Piras, V., Tomita, M., Selvarajoo, K., 2014. Transcriptome-wide variability in single embryonic development cells. Sci. Rep. 4, 1–9. https://doi.org/10.1038/srep07137.

Plaimas, K., Mallm, J.P., Oswald, M., Svara, F., Sourjik, V., Eils, R., König, R., 2008. Machine learning based analyses on metabolic networks supports high-throughput knockout screens. BMC Syst. Biol. 2 https://doi.org/10.1186/1752-0509-2-67.

Poplin, R., Chang, P.C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., Depristo, M.A., 2018. A universal snp and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983. https://doi.org/10.1038/nbt.4235.

Quest, D.J., Land, M.L., Brettin, T.S., Cottingham, R.W., 2010. Next generation models for storage and representation of microbial biological annotation. BMC Bioinformatics 11. https://doi.org/10.1186/1471-2105-11-15.

Regulatory Affairs Professionals Society, 2020. COVID-19 Vaccine Tracker. RAPS. https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker. (Accessed 24 July 2020).

Rieder, G., Simon, J., 2017. Big Data: a New Empiricism and its Epistemic and Socio-Political Consequences, pp. 85–105.

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. https://doi.org/10.1016/j.biotechadv.2017.09.005.

Šali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815. https://doi.org/10.1006/jmbi.1993.1626.

Savile, C.K., Janey, J.M., Mundorff, E.C., Moore, J.C., Tam, S., Jarvis, W.R., Colbeck, J.C., Krebber, A., Fleitz, F.J., Brands, J., Devine, P.N., Huisman, G.W., Hughes, G.J., 2010. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science 329, 305–309. https://doi.org/10.1126/science.1188934, 80-.

Selvarajoo, K., 2011. Macroscopic law of conservation revealed in the population dynamics of Toll-like receptor signaling. Cell Commun. Signal. 9, 9. https://doi.org/10.1186/1478-811X-9-9.

Selvarajoo, K., 2017. A systems biology approach to overcome TRAIL resistance in cancer treatment. Prog. Biophys. Mol. Biol. 128, 142–154. https://doi.org/10.1016/j.pbiomolbio.2017.02.009.

Selvarajoo, K., 2018. Order parameter in bacterial biofilm adaptive response. Front. Microbiol. 9, 1721. https://doi.org/10.3389/fmicb.2018.01721.

Selvarajoo, K., Tomita, M., 2013. Physical laws shape biology. Science 339, 646. https://doi.org/10.1126/science.339.6120.646-a, 80-.

Selvarajoo, K., Takada, Y., Gohda, J., Helmy, M., Akira, S., Tomita, M., Tsuchiya, M., Inoue, J., Matsuo, K., 2008. Signaling flux redistribution at toll-like receptor pathway junctions. PloS One 3, e3430. https://doi.org/10.1371/journal.pone.0003430.

Selvarajoo, K., Tomita, M., Tsuchiya, M., 2009. Can complex cellular processes be governed by simple linear rules? J. Bioinf. Comput. Biol. 7, 243–268. https://doi.org/10.1142/S0219720009003947.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., Segal, E., 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. 30, 521–530. https://doi.org/10.1038/nbt.2205.

Shukal, S., Chen, X., Zhang, C., S, S., X, C., C, Z., 2019. Systematic engineering for high-yield production of viridiflorol and amorphadiene in auxotrophic Escherichia coli. Metab. Eng. 55, 170–178. https://pubmed.ncbi.nlm.nih.gov/31326469/. (Accessed 24 July 2020).

Skraly, F.A., Ambavaram, M.M.R., Peoples, O., Snell, K.D., 2018. Metabolic engineering to increase crop yield: from concept to execution. Plant Sci. 273, 23–32. https://doi.org/10.1016/j.plantsci.2018.03.011.

Smith, S., 2020a. 9 Artificial Intelligence in Drug Discovery Trends and Statistics. https://blog.benchsci.com/artificial-intelligence-in-drug-discovery-trends-and-statistics. (Accessed 24 July 2020).

Smith, S., 2020b. 43 Pharma Companies Using Artificial Intelligence in Drug Discovery. https://blog.benchsci.com/pharma-companies-using-artificial-intelligence-in-drug-discovery. (Accessed 24 July 2020).

Smith, S., 2020c. 230 Startups Using Artificial Intelligence in Drug Discovery. https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery. (Accessed 24 July 2020).

Smith, S., 2020d. 116 Drugs in the Artificial Intelligence in Drug Discovery Pipeline. https://blog.benchsci.com/drugs-in-the-artificial-intelligence-in-drug-discovery-pipeline. (Accessed 24 July 2020).

Song, Y., Dimaio, F., Wang, R.Y.R., Kim, D., Miles, C., Brunette, T., Thompson, J., Baker, D., 2013. High-resolution comparative modeling with RosettaCM. Structure 21, 1735–1742. https://doi.org/10.1016/j.str.2013.08.005.

Srinivasan, S., Cluett, W.R., Mahadevan, R., 2015. Constructing kinetic models of metabolism at genome-scales: a review. Biotechnol. J. 10, 1345–1359. https://doi.org/10.1002/biot.201400522.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E., 2015. Big data: astronomical or genomical? PLoS Biol. 13, e1002195 https://doi.org/10.1371/journal.pbio.1002195.

Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Muggleton, S., 2006. Application of abductive ILP to learning metabolic network inhibition from temporal data. In: Mach. Learn. Springer, pp. 209–230. https://doi.org/10.1007/s10994-006-8988-x.

Tetko, I.V., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Fobo, G., Ruepp, A., Antonov, A.V., Surmeli, D., Mewes, H.-W., 2005. MIPS bacterial genomes functional annotation benchmark dataset. Bioinformatics 21, 2520–2521.

Toya, Y., Shimizu, H., 2013. Flux analysis and metabolomics for systematic metabolic engineering of microorganisms. Biotechnol. Adv. 31, 818–826. https://doi.org/10.1016/j.biotechadv.2013.05.002.

Tran, L.M., Rizk, M.L., Liao, J.C., 2008. Ensemble modeling of metabolic networks. Biophys. J. 95, 5606–5617. https://doi.org/10.1529/biophysj.108.135442.

Volk, M.J., Lourentzou, I., Mishra, S., Vo, L.T., Zhai, C., Zhao, H., 2020. Biosystems design by machine learning. ACS Synth. Biol. 9, 1514–1533. https://doi.org/10.1021/acssynbio.0c00129.

Wang, X., Song, K., Li, L., Chen, L., 2018. Structure-based drug design strategies and challenges. Curr. Top. Med. Chem. 18, 998–1006. https://doi.org/10.2174/1568026618666180813152921.

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. https://doi.org/10.1016/j.meteno.2015.06.003.

Wu, S.G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R., Zhang, X., Shimizu, K., Tang, Y.J., Bao, F.S., 2016a. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS Comput. Biol. 12, 1004838. https://doi.org/10.1371/journal.pcbi.1004838.

Wu, S.G., Shimizu, K., Tang, J.K.-H., Tang, Y.J., 2016b. Facilitate collaborations among synthetic biology, metabolic engineering and machine learning. ChemBioEng Rev 3, 45–54. https://doi.org/10.1002/cben.201500024.

Xiao, L., Xiao, Y., Wang, Z., Tan, H., Tang, K., Zhang, L., 2015. Metabolic engineering of vitamin C production in Arabidopsis. Biotechnol. Bioproc. Eng. 20, 677–684. https://doi.org/10.1007/s12257-015-0090-4.

Yadav, V.G., De Mey, M., Giaw Lim, C., Kumaran Ajikumar, P., Stephanopoulos, G., 2012. The future of metabolic engineering and synthetic biology: towards a systematic practice. Metab. Eng. 14, 233–241. https://doi.org/10.1016/j.ymben.2012.02.001.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y., 2014. The I-TASSER suite: protein structure and function prediction. Nat. Methods 12, 7–8. https://doi.org/10.1038/nmeth.3213.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D., 2020. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U.S.A. 117, 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Yip, K.Y., Cheng, C., Gerstein, M., 2013. Machine learning and genome annotation: a match meant to be? Genome Biol. 14, 205. https://doi.org/10.1186/gb-2013-14-5-205.

Yiu Chan, C.W., Gu, Z., Bieg, M., Eils, R., Herrmann, C., 2019. Impact of cancer mutational signatures on transcription factor motifs in the human genome. BMC Med. Genom. 12, 64. https://doi.org/10.1186/s12920-019-0525-4.

Zednik, C., 2019. Solving the black box problem: a normative framework for explainable artificial intelligence. Philos. Technol. 1–24. https://doi.org/10.1007/s13347-019-00382-7.

Zelezniak, A., Vowinckel, J., Capuano, F., Messner, C.B., Demichev, V., Polowsky, N., Mülleder, M., Kamrad, S., Klaus, B., Keller, M.A., Ralser, M., 2018. Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts. Cell Syst 7, 269–283. https://doi.org/10.1016/j.cels.2018.08.001 e6.

Zhang, D., Dechatiwongse, P., del Rio-Chanona, E.A., Maitland, G.C., Hellgardt, K., Vassiliadis, V.S., 2015. Dynamic modelling of high biomass density cultivation and biohydrogen production in different scales of flat plate photobioreactors. Biotechnol. Bioeng. 112, 2429–2438. https://doi.org/10.1002/bit.25661.

Zhang, C., Chen, X., Too, H.P., 2020a. Microbial astaxanthin biosynthesis: recent achievements, challenges, and commercialization outlook. Appl. Microbiol. Biotechnol. 104, 5725–5737. https://doi.org/10.1007/s00253-020-10648-2.

Zhang, C., Chen, X., Orban, A., Shukal, S., Birk, F., Too, H.P., Rühl, M., 2020b. Agrocybe aegerita serves as a gateway for identifying sesquiterpene biosynthetic enzymes in higher fungi. ACS Chem. Biol. 15, 1268–1277. https://doi.org/10.1021/acschembio.0c00155.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. 430, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.