Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Machine learning applications in genome-scale metabolic modeling

Yeji Kim[1,2,3], Gi Bae Kim[1,2,3] and Sang Yup Lee[1,2,3]

Abstract

Genome-scale metabolic modeling and simulation have been widely employed in biological studies and biotechnological applications due to their powerful capabilities of estimating metabolic fluxes at the systems level. In recent years, machine learning (ML) has been beginning to be applied to the reconstruction and analysis of genome-scale metabolic models (GEMs) to improve their quality. Also, ML has been used to diversify the utilization of information derived from genome-scale metabolic modeling and simulation. Recent studies have shown that machine learning can improve predictive performance and data coverage of GEMs. Also, genome-scale metabolic modeling and simulation provide interpretability of ML applications. Although many biological data still need to be made suitable for ML applications, it is expected that ML will be increasingly applied to GEMs to further improve the practical use and find new applications of GEMs.

Addresses

[1] Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea

[2] Systems Metabolic Engineering and Systems Healthcare Cross-Generation Collaborative Laboratory, KAIST, Daejeon 34141, Republic of Korea

[3] KAIST Institute for the BioCentury, KAIST Institute for Artificial Intelligence, BioProcess Engineering Research Center and BioInformatics Research Center, KAIST, Daejeon, 34141, Republic of Korea

Corresponding author: Lee, Sang Yup (leesy@kaist.ac.kr)

Current Opinion in Systems Biology 2021, **25**:42–49

This review comes from a themed issue on **Mathematical Modelling (2021)**

Edited by **Stacey D. Finley** and **Vassily Hatzimanikatis**

For complete overview of the section, please refer the article collection - Mathematical Modelling (2021)

Available online 11 March 2021

https://doi.org/10.1016/j.coisb.2021.03.001

Keywords

Genome-scale metabolic model, Genotype-phenotype association, Machine learning, Metabolic network, Omics data.

Introduction

High-throughput omics technologies employed widely over the last two decades have generated an unprecedentedly large amount of biological data of various types spanning all domains of life. As in other engineering and science disciplines, big data in biology is opening up the possibility of data-driven science toward understanding complex biological systems and phenomena [1]. However, the vast amount of biological data often generated under a wide range of different experimental conditions makes data interpretation and utilization more difficult. In this regard, machine learning (ML) has emerged as a powerful approach to handle biological omics data with diverse purposes of prediction, classification, and discovery [2].

Biological big data are allowing us to study biological networks (e.g., gene regulatory networks, metabolic networks, and protein–protein interaction networks) at the systems level [3]. Among various studies, the reconstruction of genome-scale metabolic models (GEMs) has arisen as a major biological network modeling approach at the systems level based on the availability of numerous complete genome sequences. GEMs, which delineate gene–protein–reaction associations for an entire set of metabolic genes, can be utilized to predict metabolic fluxes and serve as a platform to integrate multiple types of omics and kinetic data [4]. GEMs have been widely used in various applications, including understanding cellular phenotypes under genotypically and environmentally perturbed conditions, developing metabolic engineering strategies, and identifying therapeutic targets, to name a few.

In recent years, research on applying ML to GEM studies has been actively conducted, and this has greatly improved the fidelity of the genotype-phenotype association by combining the contextual knowledge of the biological system inferred by GEM studies and the predictive capability of ML [5–8]. Here we review recent trends in implementing ML in GEM studies focusing on two aspects: ML applications in constraint-based reconstruction and analysis (COBRA) of GEMs and ML applications using information-derived from GEMs.

ML applications in constraint-based reconstruction and analysis of GEMs

The COBRA of GEMs has been widely conducted to understand genotype-phenotype associations, metabolic

fluxes, and characteristics in diverse organisms. A general workflow of COBRA consists of genome annotation, functional annotation of metabolic genes, a mathematical representation of the metabolic network, and analysis of the reconstructed GEM with constraint-based simulations [2,4,9]. In order to alleviate the intensive reconstruction process, several computational tools have been developed, including automatic GEM reconstruction pipelines [4,10–12]. More recently, the ML methods have also begun to be applied in COBRA.

At the early stage of COBRA, metabolic genes are annotated by identifying the functions of the corresponding enzymes. For the high-throughput reconstruction of GEMs, enzyme functions need to be predicted accurately and quickly. DeepEC, a deep learning-based computational framework, was developed to predict enzyme commission (EC) numbers in a high-throughput manner [13]. DeepEC uses three convolutional neural networks (CNNs) to extract latent features of enzymes from protein sequences (Figure 1a). The first CNN predicts whether an input protein sequence is an enzyme, while the second and the third CNNs predict EC numbers of the enzyme up to three and four digits, respectively. The functions of enzymes are assigned to the predicted EC numbers only if the predicted three-digit and four-digit EC numbers are consistent. Otherwise, DeepEC performs homology analysis to assign the EC numbers of homologous proteins. Since the deep learning-based functional annotation of enzymes is fast and accurate, it will be very useful in reconstructing not only the GEM of a particular organism but also community GEMs and pan-GEMs in a high-throughput manner.

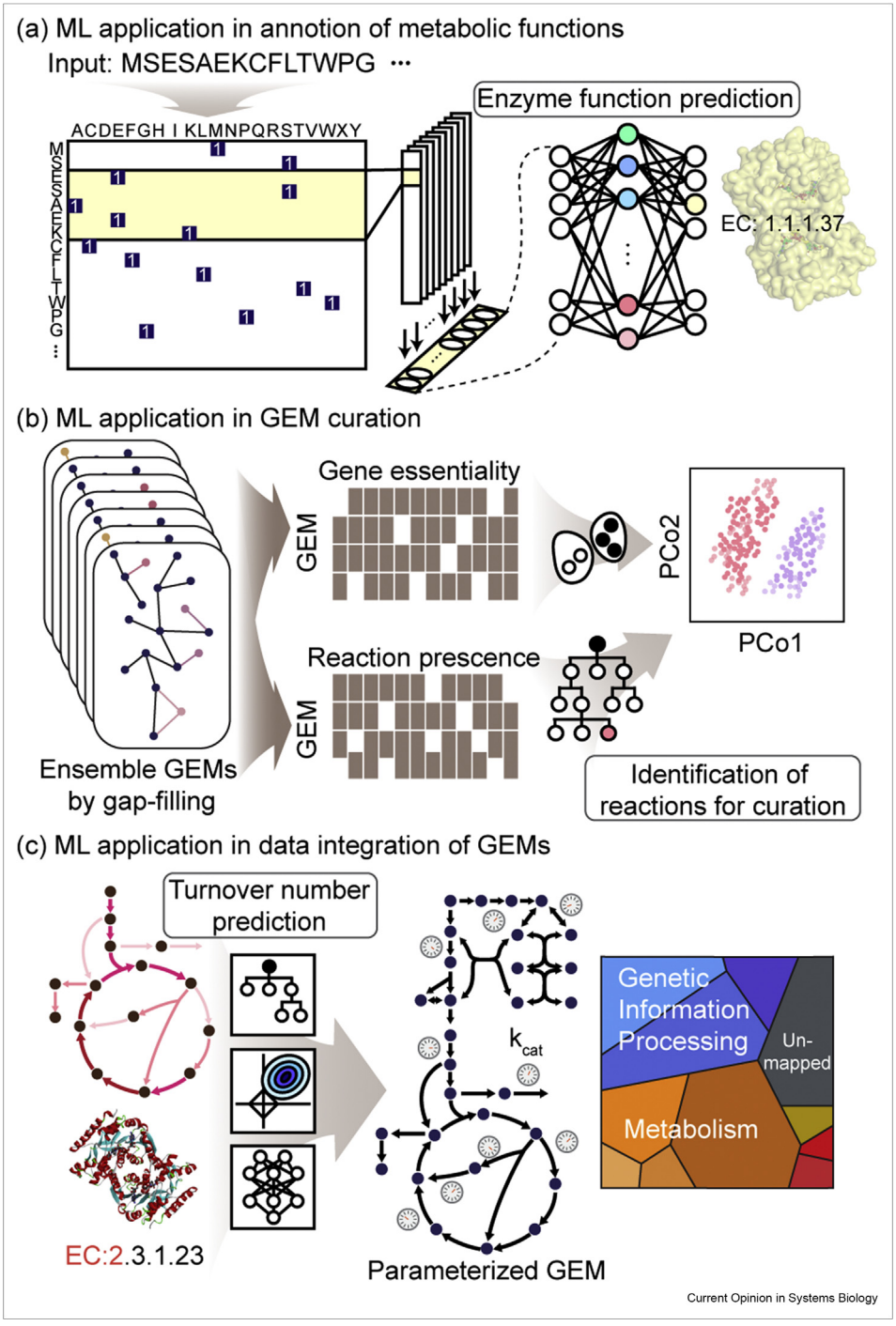
Although various automatic tools have enabled the high-throughput reconstruction of draft GEMs, high-quality GEM reconstruction still requires intensive manual curation of metabolic networks [14]. Recently published studies have shown that applying ML to the demanding manual curation process can increase the efficiency of COBRA. For example, an ML-based method, automated metabolic model ensemble-driven uncertainty elimination using statistical learning (AMMEDEUS), was developed for prioritizing metabolic reactions to be manually curated (Figure 1b) [15]. AMMEDEUS aims to identify metabolic reactions that have a significant impact on the simulation performance. First, an ensemble of GEMs is generated from a draft GEM by iterative gap-filling. K-means clustering is performed to cluster the ensemble members into two clusters based on the similarity of GEM simulation profiles. Then, a random forest classifier is trained to predict the cluster to which a GEM belongs, using the presence or absence of metabolic reactions in the GEM. Feature importance indicating how much a metabolic reaction contributes to the classifier performance and cluster ratio indicating how much a metabolic reaction is enriched in a single cluster are used to prioritize metabolic reactions to be manually curated. This study suggested that the tedious, yet important, manual curation steps can be incorporated into the automated GEM reconstruction pipeline through the use of ML, thereby improving the quality of the GEMs generated by automated reconstruction [4].

Reconstructed GEM is often subject to constraint-based simulations to predict metabolic phenotypes, and the predictive power of constraint-based simulations can be enhanced by incorporating high-throughput omics data into the GEM [16]. In the process of data integration, ML application can be a powerful method to expand the data coverage and improve the predictive capabilities of GEMs [17,18]. In a recent study, a GEM was parameterized using *in vivo* enzyme turnover numbers (k_{cat}s) with improved coverage through ML, allowing better prediction of quantitative proteome data (Figure 1c) [17]. More specifically, *in vivo* k_{cat}s values were estimated using proteomic and fluxomic data of 21 *Escherichia coli* strains, and were found to be robust against genetic perturbations. The *in vivo* k_{cat}s values were extrapolated to the genome-scale covering 1870 metabolic reactions in an *E. coli* GEM by an ensemble ML model of random forest, elastic net, and neural network using the input features of metabolic network properties, enzyme structural properties, and information on enzyme biochemistry. Two types of *E. coli* GEMs parameterized by the predicted *in vivo* k_{cat}s showed a higher predictive performance for the quantitative proteome compared to the GEMs parameterized by *in vitro* k_{cat}s [17].

ML applications using information-derived from GEMs

In addition to ML applications in COBRA, ML has been applied to advance and diversify the utilization of fluxomic information derived from GEM simulation. Recent ML applications using GEMs include frequent use of multiomics data, improved interpretability, and expansion of research scope to the comparative genomics level. Integration of different omics data types has the potential to better understand complex biological phenomena that cannot be understood by a single data type [1]. In addition, public databases for the omics data are constantly being updated with an ever-increasing amount of experimental data, and also the platforms that integrate different types of omics data have emerged (Table 1). Such data richness has made ML approaches using integrated multiomics more promising in performing studies in fundamental biology and biotechnology [5,6]. Fluxomic data obtained from GEM simulation or experimentally determined by ^{13}C metabolic flux analysis have been proven to be important for such ML applications. Also, using fluxome with other omics data has been shown to further improve the predictive power in ML applications. In a recent study, fluxomic

Figure 1



Recent ML applications in the COBRA of GEMs. **(a)** ML application in the annotation of metabolic functions. A deep learning-based method has been applied to predict EC numbers, which can be used for the annotation of enzyme functions in COBRA [13]. **(b)** ML application in GEM curation. An ML method was developed to prioritize reactions to be manually curated, demonstrating that ML application can help simplify the demanding manual curation of GEMs [15]. **(c)** ML application in data integration of GEMs. ML methods can expand the coverage of omics data to be integrated into GEMs, which consequently improves the predictive capabilities of GEMs. In a recent study, a GEM parameterized by *in vivo* k_{cat}s, which were extrapolated through ML, showed better predictive performance for quantitative proteomic data [17].

Table 1

The latest updates of the representative databases available for ML applications in GEM studies since 2018.

Data type	Database name	Major latest updates	Latest reference
Genome	DDBJ	• Computational system has been upgraded with 30 petabytes of DNA data archiving storage. • An automatic file transfer system has been newly installed.	[34] (2020)
	ENA	• The accumulated data comprises 1.5 × 109 sequences and 8 × 1015 base pairs across 1.5 × 106 taxa. • Additional services, such as data coordination and presubmission validation, have been developed.	[35] (2019)
	GenBank	• Data increases include 57 synthetic chromosomal constructs and 60 chromosome-scale eukaryotic sequences. • Technical updates have been reported including new submission wizards for viral genomes and a new Genome Workbench version.	[36] (2019)
Transcriptome	ArrayExpress Archive	• Capacity for single-cell sequencing experiments has been increased. • ArrayExpress will be superseded by BioStudies.	[37] (2019)
	Expression Atlas	• The latest data includes expression data from 3126 studies across 33 species. • New features have been presented, including options to analyze gene set overlaps and to view genes with similar expression patterns.	[38] (2018)
	GTEx	• In-depth characterization of version 8 data (i.e., 15,201 RNA-sequencing samples from 49 tissues of 838 individuals) has been presented.	[39] (2020)
Multiomics	KEGG		• A predictive method tool, KEGG Mapper [40], of reconstructing molecular network systems and a network database, KEGG NETWORK [41], has been developed. [40] (2020)
	BioStudies	• BioStudies is a newly established database to deal with biological multimodal data, which include epigenetics, RNA, and protein expression data.	[42] (2020)
TCGA	• The data have accumulated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. • A standardized dataset, namely TCGA Pan-Cancer Clinical Data Resource, has been additionally featured.	[43] (2018)	
Proteome	BRENDA	• Major updates include a new search option for similarity and isomer, supplementation of BKMS-react with further reactions, and development of plant word maps.	[44] (2019)
	HPA		[45] (2018)

(*continued on next page*)

Table 1 (*continued*)

Data type	Database name	Major latest updates	Latest reference
	PRIDE	• The coverage of protein data has been expanded to 86% of all human protein-coding genes. • Three new Atlases (i.e., Blood Atlas, Brain Atlas, Metabolic Atlas) have been introduced. • To date, 10,110 datasets have been archived. Technical improvements have been made on the storage backend, API, and web interface. • The support for quantitative proteomics data has been improved.	[46] (2019)
	ProteomicsDB	• The data coverage has been expanded to include nearly 300 human tissues and cell lines and to include additional support for other organisms. • The platform has begun to further support data from other organisms and protein turnover data.	[47] (2020)
Metabolome	MetaCyc	• To date, 2749 pathways from more than 60,000 publications have been archived, of which 184 base pathways and 5 superpathways have been newly added. • GlycanBuilder and PathoLogic components have been specifically improved.	[48] (2020)

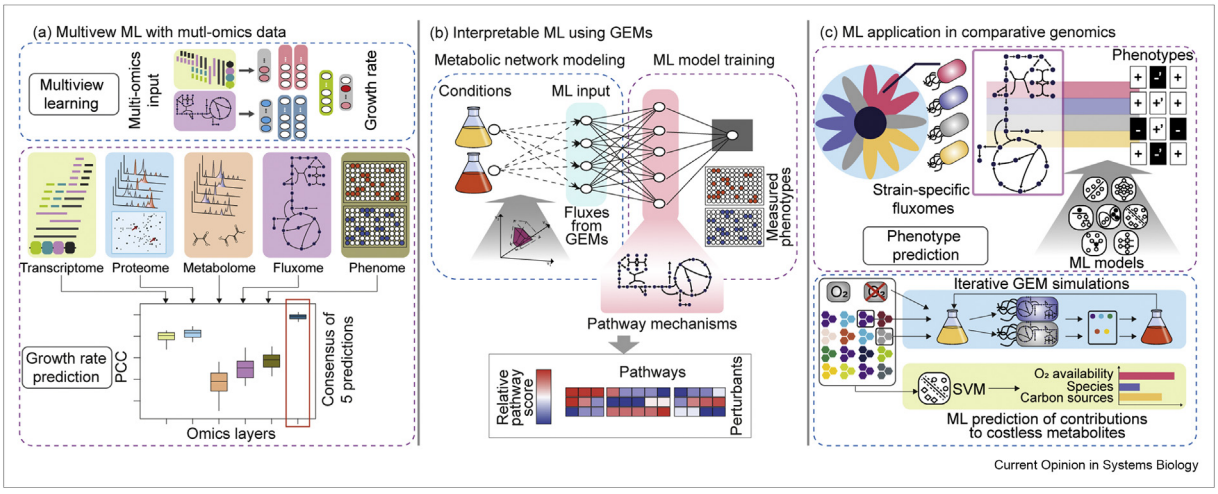
data obtained by GEM simulation and experimental transcriptomic data were used in combination to predict cell growth of numerous *Saccharomyces cerevisiae* strains [19] (Figure 2a). Gene expression data were used for preparing strain-specific GEMs, which were consequently simulated to obtain fluxomic data. By testing 27 different ML methods, a multiview neural network showed the highest performance for predicting cell growth. Importantly, the multiview neural network using both omics data outperformed the single-view learning with transcriptomic data only. This elegant study emphasized the importance of analyzing the fluxome as key omics data to predict metabolic characteristics under given genetic and environmental conditions, together with other omics data, including transcriptome.

In another study, a curated multiomics compendium for *E. coli* was built and used to train an integrative ML model, consisting of multiple layers, each of which processes an individual type of omics data (i.e., transcriptome, proteome, metabolome, fluxome, and phenome) to predict genome-scale expression and growth states [20] (Figure 2a). The ML model receives experimental conditions, such as strains, genetic perturbations, media, and environmental stresses as inputs and predicts corresponding cellular states expressed by the concentrations of cellular components, metabolic fluxes, and growth rates. The integration of the multiple omics layers significantly increased the prediction performance, measured by Pearson's correlation coefficient, and coverage compared to individual data layers.

Although the application of ML has shown great performance in predicting input–output correlations from biological data, causal mechanisms underlying the predictions are often uninterpretable. Biological network models are mechanistically constructed using interaction data, which can serve as a mechanistic framework for ML to deduce causal genotype–phenotype relationships [21]. Accordingly, recent studies have shown that ML applications can become interpretable through the GEM simulation and provide new mechanistic insights [8,22]. In a recent study, biochemical screening, GEM simulation, and ML were performed to predict antibiotic lethality and elucidate the metabolic mechanisms of action for antibiotics (Figure 2b) [8]. The effects of various supplements on antibiotic lethality were experimentally screened in *E. coli*. The metabolic network states for the perturbations were simulated through a GEM with the results of the biochemical screening. A least-squares model was trained with the fluxes of reactions selected by a multitask elastic net and the lethality data measured in IC_{50} values for each antibiotic. Through this ML process, a novel metabolic mechanism for antibiotic lethality involving antibiotic-induced adenine limitation was discovered. This study demonstrated how mechanistic GEMs can help ML unveil underlying causal relationships between perturbations and phenotypes.

Another interesting interpretable ML method, namely Metabolic Allele Classifier (MAC), was developed using flux balance analysis (FBA), the most popular GEM simulation method, to predict allele-specific antimicrobial resistance (AMR) of a specific strain [22]. In detail, MAC has been formulated within the FBA framework using allele-specific flux capacity constraints and individual antibiotic-specific objective functions. MAC receives a genome sequence of a *Mycobacterium tuberculosis* strain and outputs the predicted AMR phenotypes (i.e., resistant or susceptible to a specific antibiotic) by optimizing the antibiotic-specific objective function.

Figure 2



Major trends in recent ML applications using information derived from GEMs. **(a)** Multiview ML with multiomics data. In a recent study, fluxomic data from GEM simulations and experimental transcriptomic data were used in combination to predict cell growth of *S. cerevisiae* strains. Among several ML methods, a multiview neural network showed the highest performance in predicting cell growth (upper panel) [19]. Similarly in another study, the integration of the multiple data types (i.e., transcriptome, proteome, metabolome, fluxome, and phenome) from a curated multiomics compendium for *E. coli* significantly increased the prediction performance presented by Pearson's correlation coefficient (PCC) (lower panel) [20]. **(b)** Interpretable ML using GEMs. An exemplary study trained a least-squares regression model with the fluxes calculated from GEM simulation and experimental phenotypes [8]. Through the ML process, known and novel metabolic mechanisms for the phenotype prediction were found. **(c)** ML application in comparative genomics. GEMs can serve as a platform for ML applications to analyze genotype–phenotype relationships from the datasets of pan-genome, GWAS (upper panel) [22], or genomes in a microbial community (lower panel) [24].

The formulation of MAC is beneficial in that a flux state corresponds to each strain-antibiotic classification, which can provide a biochemical interpretation of the genotype-phenotype map. Application of MAC recapitulated the known mechanisms of AMR by identifying genetic determinants and pathways responsible for discrimination of AMR phenotypes, which proved the interpretability of the GEM-based ML method. As comparative genomic studies are being actively performed with advanced genome sequencing technologies [23], the studies on ML applications using GEMs are also expanding to the level of comparative genomics. For example, the GEM-based ML framework, MAC, was applied to a genome-wide association study (GWAS) dataset comprising 1595 drug-tested *M. tuberculosis* strains and differentiated their AMR phenotypes (Figure 2c) [22]. Notably, the GEM was able to explain most of the known genetic determinants of AMR for multiple strains. This study showed that GEMs can serve as a platform for ML applications to extract insights on biological networks and genotype–phenotype relationships from pan-genome-level and GWAS datasets.

Another interesting study on community GEMs has investigated the contributions of costless secreted metabolites, which incur no fitness cost on the producer, to the intermicrobial interactions in a microbial community by performing over 2 million pairwise simulations of 24 species under various environmental conditions (Figure 2c) [24]. By doing so, a number of metabolites that can be costlessly secreted and consequently are useful for cross-feeding were identified. Along with the extensive GEM simulations, support vector machine (SVM) models were additionally constructed and trained to quantify the extent to which conditional variables (e.g., oxygen availability, species identity, and carbon sources) contribute to the secretion of costless metabolites. As a result of the ML analysis, oxygen availability was predicted to have the strongest association with the costless metabolites secretion. This study showed that ML-assisted GEM analysis can provide specific insights into the interspecies interactions in the microbial communities.

Conclusions

In this paper, we reviewed recent trends in ML applications for COBRA and for the advanced use of GEM-derived information. Recent studies showed that ML applications in GEM studies can improve predictive power, data coverage, and the interpretability of predictions. Despite the advances in ML applications in GEM studies, several challenges remain to be addressed. First, the ML applications in GEM studies have been performed in a limited number of systems, mostly microorganisms, possibly because reliable GEMs have been built mainly for microorganisms. Recently, ML was applied to the metabolic fluxes determined from a human GEM to predict genotype–phenotype

relationships associated with drug side effects [25]. This study, however, used the first generic human GEM released in 2007 [26], and thus, can be better performed using the human GEMs that have been updated and revised [27–31]. In addition to human GEMs, high-quality genomes and their GEMs for other higher organisms, including animals and plants, have become available [4,31–33]. Thus, it is expected that ML applications in GEM studies will increase for better biological understanding and biotechnological applications. Second, the suitability of biological data for ML applications is still restricted. Omics data, including fluxomic data from GEMs, are complicated to preprocess for ML due to their heterogeneity and have high data dimensions when considering integrative use of multiomics data, which might reduce the reliability of ML results [1]. Although the amount of omics data has been increasing rapidly, it is not yet always sufficient to apply state-of-the-art deep learning methods; however, it should be noted that omics data in the DBs are being rapidly updated and made consistent to facilitate their secondary uses such as ML applications (Table 1). With such advances, it is expected that ML and deep learning will be increasingly applied to GEMs to further improve the practical use and find new applications of GEMs.

Conflict of interest statement

Nothing declared.

Acknowledgements

This work was supported by the Technology Development Program to Solve Climate Changes on Systems Metabolic Engineering for Biorefineries (NRF-2012M1A2A2026556 and NRF-2012M1A2A2026557) from the Ministry of Science and ICT (MSIT) through the National Research Foundation (NRF) of Korea.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Xu C, Jackson SA: **Machine learning and complex biological data**. *Genome Biol* 2019, **20**:76.

2. Kim GB, Kim WJ, Kim HU, Lee SY: **Machine learning applications in systems metabolic engineering**. *Curr Opin Biotechnol* 2020, **64**:1–9.

3. Muzio G, O'Bray L, Borgwardt K: **Biological network analysis with deep learning**. *Briefings Bioinf* 2020, https://doi.org/10.1093/bib/bbaa257.

4. * Gu C, Kim GB, Kim WJ, Kim HU, Lee SY: **Current status and applications of genome-scale metabolic models**. *Genome Biol* 2019, **20**:1–18.
This paper comprehensively reviews the recent advances in GEM studies, which can be further analyzed by ML approaches. At the time of the paper, GEMs for 6239 organisms (i.e., 5897 bacteria, 127 archaea, and 215 eukaryotes) have been reconstructed.

5. Zampieri G, Vijayakumar S, Yaneske E, Angione C: **Machine and deep learning meet genome-scale metabolic modeling**. *PLoS Comput Biol* 2019, **15**, e1007084.

6. Rana P, Berry C, Ghosh P, Fong SS: **Recent advances on constraint-based models by integrating machine learning**. *Curr Opin Biotechnol* 2020, **64**:85–91.

7. Zhang J, Petersen SD, Radivojevic T, Ramirez A, Pérez-Manríquez A, Abeliuk E, Sánchez BJ, Costello Z, Chen Y, Fero MJ, *et al.*: **Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism**. *Nat Commun* 2020, **11**:4880.

8. ** Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO, *et al.*: **A white-box machine learning approach for revealing antibiotic mechanisms of action**. *Cell* 2019, **177**:1649–1661. e9.
This study demonstrated how mechanistic GEMs can help ML discover underlying causal relationships between perturbations and phenotypes. By performing biochemical screening, GEM simulations, and ML, the antibiotic lethality of bacteria and the metabolic mechanisms of action for the bactericidal antibiotics could be predicted.

9. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93–121.

10. Machado D, Andrejev S, Tramontano M, Patil KR: **Fast automated reconstruction of genome-scale metabolic models for microbial species and communities**. *Nucleic Acids Res* 2018, **46**:7542–7553.

11. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ: **Raven 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor***. *PLoS Comput Biol* 2018, **14**, e1006541.

12. Karlsen E, Schulz C, Almaas E: **Automated generation of genome-scale metabolic draft reconstructions based on KEGG**. *BMC Bioinf* 2018, **19**:467.

13. Ryu JY, Kim HU, Lee SY: **Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers**. *Proc Natl Acad Sci* 2019, **116**:13996–14001.

14. Norsigian CJ, Fang X, Seif Y, Monk JM, Palsson BO: **A workflow for generating multi-strain genome-scale metabolic models of prokaryotes**. *Nat Protoc* 2020, **15**:1–14.

15. Medlock GL, Papin JA: **Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning**. *Cell Syst* 2020, **10**:109–119.e3.

16. Ramon C, Gollub MG, Stelling J: **Integrating omics data into genome-scale metabolic network models: principles and challenges**. *Essays Biochem* 2018, **62**:563–574.

17. * Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, Feist AM, Gonzalez DJ, Palsson BO: **Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers**. *Proc Natl Acad Sci* 2020, **117**:23182–23190.
This study parameterized GEMs using *in vivo* k_{cat}s, which were extrapolated to genome scale by ML. The GEMs parameterized by *in vivo* k_{cat}s showed a better prediction of quantitative proteome data than those parameterized by *in vitro* k_{cat}s. These results suggested that ML application can improve data coverage and predictive power of GEMs.

18. 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**. *Nat Commun* 2018, **9**:5252.

19. ** Culley C, Vijayakumar S, Zampieri G, Angione C: **A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth**. *Proc Natl Acad Sci* 2020, **117**:18869–18879.
This paper used experimental transcriptomic data and fluxomic data produced by GEM in an integrated manner to predict cell growth of a number of yeast strains. Among various ML methods tested, a multi-view learning method outperformed single-view learning methods. This result demonstrated that the use of multiomics data could improve the prediction power of ML.

20. Kim M, Rai N, Zorraquino V, Tagkopoulos I: **Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli***. *Nat Commun* 2016, **7**:13090.

21. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ: **Next-generation machine learning for biological networks**. *Cell* 2018, **173**:1581–1592.

22. * Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO: **A biochemically-interpretable machine learning classifier for microbial GWAS**. *Nat Commun* 2020, **11**:2580.

This study applied a GEM-based ML framework to a GWAS dataset of *M. tuberculosis* strains and differentiated their AMR phenotypes. This study demonstrated that GEMs can be a good platform for ML applications to extract information on genotype−phenotype relationships from pan-genome-level or GWAS datasets.

23. Kim Y, Gu C, Kim HU, Lee SY: **Current status of pan-genome analysis for pathogenic bacteria**. *Curr Opin Biotechnol* 2020, **63**:54−62.

24. * Pacheco AR, Moel M, Segrè D: **Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems**. *Nat Commun* 2019, **10**:103.

This study elucidated the contributions of costless secreted metabolites to the intermicrobial interactions by simulating a large space of species and metabolic conditions. ML was also applied to evaluate the contributions of environmental variables to the secretion of costless metabolites.

25. Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E: **Metabolic network prediction of drug side effects**. *Cell Syst* 2016, **2**: 209−213.

26. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proc Natl Acad Sci* 2007, **104**:1777−1782.

27. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, *et al.*: **A community-driven global reconstruction of human metabolism**. *Nat Biotechnol* 2013, **31**:419−425.

28. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, Zielinski DC, Ang KS, Gardiner NJ, Gutierrez JM, *et al.*: **Recon 2.2: from reconstruction to model of human metabolism**. *Metabolomics* 2016, **12**:109.

29. Ryu JY, Kim HU, Lee SY: **Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism**. *Proc Natl Acad Sci* 2017, **114**: E9740−E9749.

30. Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, Gatto F, Nilsson A, Gonzalez GAP, Aurich MK, *et al.*: **Recon3D enables a three-dimensional view of gene variation in human metabolism**. *Nat Biotechnol* 2018, **36**:272−281.

31. Robinson JL, Kocabaş P, Wang H, Cholley P-E, Cook D, Nilsson A, Anton M, Ferreira R, Domenzain I, Billa V, *et al.*: **An atlas of human metabolism**. *Sci Signal* 2020:13.

32. Sherman RM, Salzberg SL: **Pan-genomics in the human genome era**. *Nat Rev Genet* 2020, **21**:243−254.

33. Zahn LM: **A high-quality rhesus macaque genome**. *Science* 2020, **370**:1428−1430.

34. Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T: **DDBJ Database updates and computational infrastructure enhancement**. *Nucleic Acids Res* 2020, **48**:D45−D50.

35. Harrison PW, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, *et al.*: **The European nucleotide archive in 2018**. *Nucleic Acids Res* 2019, **47**:D84−D88.

36. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I: **GenBank**. *Nucleic Acids Res* 2019, **47**:D94−D99.

37. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, *et al.*: **ArrayExpress update − from bulk to single-cell expression data**. *Nucleic Acids Res* 2019, **47**:D711−D715.

38. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Füllgrabe A, Fuentes AM-P, George N, *et al.*: **Expression Atlas: gene and protein expression across multiple studies and organisms**. *Nucleic Acids Res* 2018, **46**: D246−D251.

39. Consortium TGte: **The GTEx Consortium atlas of genetic regulatory effects across human tissues**. *Science* 2020, **369**: 1318−1330.

40. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M: **KEGG: integrating viruses and cellular organisms**. *Nucleic Acids Res* 2020, https://doi.org/10.1093/nar/gkaa970.

41. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M: **New approach for understanding genome variations in KEGG**. *Nucleic Acids Res* 2019, **47**:D590−D595.

42. Sarkans U, Füllgrabe A, Ali A, Athar A, Behrangi E, Diaz N, Fexova S, George N, Iqbal H, Kurri S, *et al.*: **From ArrayExpress to BioStudies**. *Nucleic Acids Res* 2020, https://doi.org/10.1093/nar/gkaa1062.

43. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, *et al.*: **An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics**. *Cell* 2018, **173**: 400−416.e11.

44. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D: **BRENDA in 2019: a European ELIXIR core data resource**. *Nucleic Acids Res* 2019, **47**:D542−D549.

45. Thul PJ, Lindskog C: **The human protein atlas: a spatial map of the human proteome**. *Protein Sci* 2018, **27**:233−244.

46. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, *et al.*: **The PRIDE database and related tools and resources in 2019: improving support for quantification data**. *Nucleic Acids Res* 2019, **47**:D442−D450.

47. Samaras P, Schmidt T, Frejno M, Gessulat S, Reinecke M, Jarzab A, Zecha J, Mergner J, Giansanti P, Ehrlich H-C, *et al.*: **ProteomicsDB: a multi-omics and multi-organism resource for life science research**. *Nucleic Acids Res* 2020, **48**: D1153−D1163.

48. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes - a 2019 update**. *Nucleic Acids Res* 2020, **48**:D445−D453.

49. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, *et al.*: **Hmdb 4.0: the human metabolome database for 2018**. *Nucleic Acids Res* 2018, **46**:D608−D617.

50. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C: **MetaboLights: a resource evolving in response to the needs of its scientific community**. *Nucleic Acids Res* 2020, **48**:D440−D444.

51. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, *et al.*: **The reactome pathway knowledgebase**. *Nucleic Acids Res* 2020, **48**: D498−D503.

52. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, *et al.*: **MEMOTE for standardized genome-scale metabolic model testing**. *Nat Biotechnol* 2020, **38**:272−276.

53. Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, King Z: **BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree**. *Nucleic Acids Res* 2020, **48**:D402−D406.