Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

# Recent advances on constraint-based models by integrating machine learning

Pratip Rana[1,3], Carter Berry[2,3], Preetam Ghosh[1] and Stephen S Fong[2]

Research that meaningfully integrates constraint-based modeling with machine learning is at its infancy but holds much promise. Here, we consider where machine learning has been implemented within the constraint-based modeling reconstruction framework and highlight the need to develop approaches that can identify meaningful features from large-scale data and connect them to biological mechanisms to establish causality to connect genotype to phenotype. We motivate the construction of iterative integrative schemes where machine learning can fine-tune the input constraints in a constraint-based model or contrarily, constraint-based model simulation results are analyzed by machine learning and reconciled with experimental data. This can iteratively refine a constraint-based model until there is consistency between experimental data, machine learning results, and constraint-based model simulations.

**Addresses**
[1] Computer Science, Virginia Commonwealth University, 401 West Main Street, Richmond, 23284, VA, USA
[2] Chemical and Life Science Engineering, Virginia Commonwealth University, 601 West Main Street, Richmond, 23284, VA, USA

Corresponding author: Fong, Stephen S (ssfong@vcu.edu)
[3] These authors contributed equally.

**Current Opinion in Biotechnology** 2020, **64**:85–91

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Yinjie J Tang** and **Ludmilla Aristilde**

For a complete overview see the Issue and the Editorial

Available online 5th December 2019

**https://doi.org/10.1016/j.copbio.2019.11.007**



## Introduction

With the development and improvements in DNA sequencing technology, the high-level goal of biological research has shifted towards understanding the genotype-phenotype relationship. Whole-genome sequencing enabled this pursuit, but it was quickly realized that genomic data on their own are not sufficient to extrapolate or predict function largely due to the multiple, interconnected layers of biological functional units. Subsequent advancements in methods and technology sought to fill the information gap between genotype and the functional phenotype by enabling systemic measurement of mRNA (using RNASeq), proteins (using mass spectrometry (MS), PCR etc.), metabolites (using gas chromatography (GC), liquid chromatography (LC), or capillary electrophoresis (CE) coupled with subsequent MS), pathways fluxes (using Nuclear magnetic resonance (NMR), gas chromatography-mass spectrometry (GC–MS)), and interactions between signal transduction, regulatory and metabolic network modules. Computational approaches have also been applied for integrating and analyzing large-scale biological data to gain better insight into biological function. However, there remains a need to develop approaches that can identify meaningful features/patterns in large-scale data and connect them to biological mechanisms to establish causality, bridging the gap between genotype and phenotype.

Two computational methods that have shown promise in addressing current large-scale biological analyses research are constraint-based modeling and machine learning. Both are generalized approaches that can be implemented for any biological system and can scale the levels of single cells, organisms, or multi-organism consortia. Constraint-based models were developed shortly after the first microbial genomes were sequenced as a method of directly utilizing genomic information to predict integrated metabolic function; thus, it has the potential to connect genotype to phenotype through gene-protein-reaction mechanisms. Machine learning (ML), on the other hand, encompass the algorithms or statistical models that can identify patterns and make hypotheses or inferences based on learning from the observed datasets. ML has grown and evolved as the scale of information has increased and has been used to identify significant features from large datasets while considering the presence of noise and interconnectedness of components. Given that both approaches can likely be implemented to study the same biological system and data and that the methods and results are largely complementary, a potentially fruitful computational approach to studying biological systems would be to combine constraint-based modeling and machine learning.

In this review, we provide a brief overview of the various elements that comprise the constraint-based modeling reconstruction pipeline highlighting instances where machine learning has successfully been used in conjunction with constraint-based modeling. Finally, we will comment on areas where opportunities for growth by developing or

implementing a combined constraint-based modeling and machine learning approach.

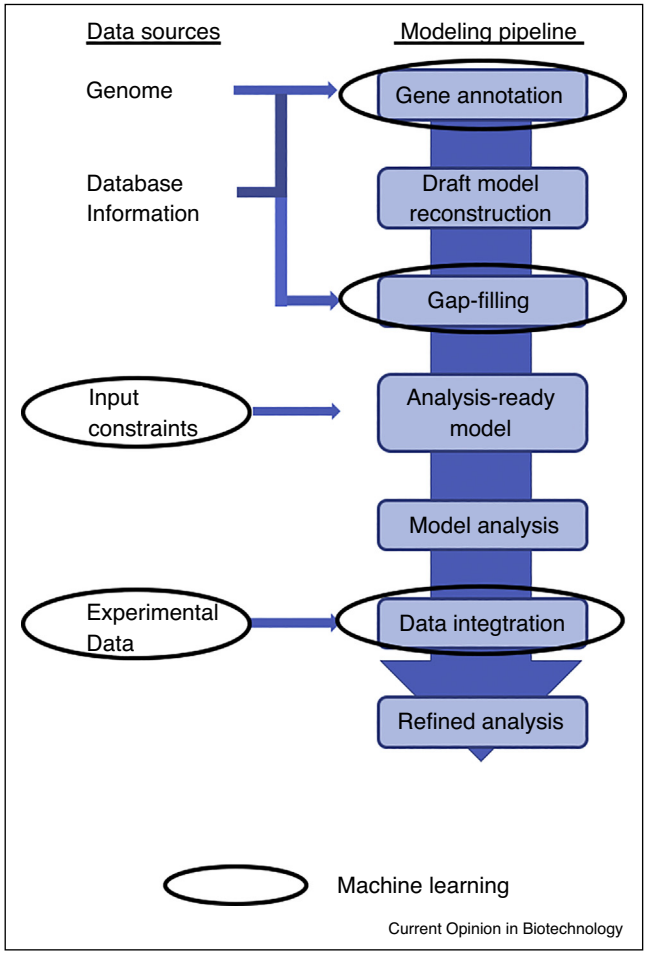
## Constraint-based modeling pipeline

Numerous review papers discuss various aspects of constraint-based modeling [1–4] and its applications. The typical constraint-based model building pipeline for model reconstruction and analysis will be used as the underlying framework for discussing work using machine learning approaches (Figure 1) and for proposing areas for potential future work. An overview of studies relevant to this integration of approaches is shown in Table 1.

### Annotation

With the ability to rapidly generate genomic data, the starting point for constructing constraint-based models is most often a DNA sequence file (FASTA file). One of the critical components of constraint-based models is the annotation of genes that constitute the core model contents. Currently there are multiple methods that can be used for gene annotation, such as, Rapid Annotations using Subsystems Technology (RAST) [5], Prokka [6], and a variety of other annotation approaches [7] such as Autograph [8] or GEM system [9] that generate gene-protein-reaction (GPR) associations using orthology. Despite the abundance of options for generating annotation and biochemical information using different tools, they can result in different gene annotations [10] which will have a direct effect on the stoichiometric matrix of constraint-based models and subsequent model-based predictions.

**Figure 1**




Overview of the constraint-based modeling pipeline highlighting areas where machine learning has been applied using the oval shapes. It additionally depicts four categories of data sources and seven broad steps in the CBM modeling pipeline.

To improve the annotation process, a machine learning based multiclass classification method applying seven different machine learning algorithms using three reaction fingerprints was developed to predict enzymatic reactions [11•]. The training data consisted of 1055 hydrolysis and 2510 redox reactions from KEGG and further validated on 213 hydrolysis and 512 redox reactions from Rhea database. Neural network and logistic regression-based models delivered the best performance and achieved around 0.9 F1 score for main class, subclass and superclass classification. Here F1 score is a measure of test accuracy, and is expressed as the harmonic mean of precision and recall. Continued, systematic use of machine learning to improve gene annotations could potentially significantly impact and improve the metabolic content of constraint-based models as errors in gene annotation can directly lead to failure modes [12].

### Gap filling

After an initial metabolic reconstruction is built from the genomic and biochemical information, it is necessary to analyze the metabolic network and potentially fill metabolic gaps [7] (e.g. missing pathways where experimental evidence indicates the organism has that functionality). As with genome annotation, a variety of gap-filling approaches have been developed [13,14].

Several machine learning methods including logistic regression, decision trees, naive Bayes have been used to identify missing reactions and enzymes in a model [15]. This work used a collection of 123 pathway features from 5610 pathway instances for learning. These machine learning methods provide similar performance compared to other pathway prediction algorithms. These ML methods also allow greater explainability (i.e. feature values are causally related to model predictions, although they may additionally require expert level verification) and extensibility (e.g. adaptable to larger datasets or even other data sources) by providing the probability for each predicted pathway and ability to improve the predictions using sophisticated input features. The association rule mining method also performed efficiently when automating functional prediction of proteins [16]. By training on UniProtKB/Swiss-Prot entries, the rule mining method

**Table 1**

**Overview of machine learning algorithms applied/applicable to metabolic network model**

| CBM pipeline component | Reference | Machine learning method | Objective |
|---|---|---|---|
| Annotation and protein function prediction | [15] | Logistic regression, decision trees, naive Bayes | Pathway prediction |
| | [16] | Rule based | Function prediction of protein |
| | [17] | Deep learning | Pathway completion and functional discovery of genes |
| | [11•] | Neural network, logistic regression, decision tree (DT), KNN, naive Bayes, random forest, and SVM | Describe the type of reactions |
| Substrate constraints | [18] | Logistic regression, random forest, scalable tree boosting system, neural network, KNN, and SVM | Predict feed substrate |
| Metabolome section | [24] | SVM | Identifying metabolites from mass spectroscopy data |
| | [25•] | Random forest and ensemble prediction | Metabolite identification |
| | [26] | Multiple linear regression | Predict metabolome from enzyme expression proteome data |
| Fluxomics | [27] | SVM, KNN, decision tree | Accelerated the flux quantification |
| Interactomics | [46] | | Automate metabolic model refinement using gene Interaction data |
| Kinetics | [47••] | Different regression algorithms including random forest and deep neural network | Protein turnover number |
| | [38] | Decision tree (CART algorithm) | Reduce the range of kinetic parameters |
| | [39•] | Supervised Learning of Metabolic Dynamics | Automate the prediction of the model dynamics |
| Multi-omic data integration | [29] | Kernel based | Full fusion of multiview data |
| | [30] | Minimizing the disagreement between the kernel matrix with imposing constraints | Full fusion of multiview data |
| | [32] | Penalized matrix tri-factorization | Intermediate fusion of multiview data |
| gene essentiality | [48] | SVM | Identify the essential genes |
| | [49] | SVM | Identify the essential genes |
| Drug effects/ targeting | [50] | SVM | Predict the side effects of a drug |
| | [51] | Bayesian model | Drug prediction |
| other | [52] | Supervised machine learning | Optimize the production of specialty chemicals |

achieved very high accuracy (F-measure = 0.982) while predicting the pathway. One deep-learning-based model, Stacked Denoising Autoencoder Multi-Label Learning (SdaMLL), was also used for pathway completion and functional discovery of genes [17]. This multi-label classification model achieved a moderate 0.577 coverage precision after training on the feature matrix derived from the term frequency of genes from 18 930 articles from biomedical literature and gene annotation from the KEGG database.

One facet of constraint-based models that bridges the gap-filling step to running model-based simulations is the need to specify input constraints. For example, flux balance analysis would typically require input constraints associated with numerical uptake rates of incoming nutrients (i.e. carbon source, oxygen, etc.). This is also important for the gap-filling stage of model building as it may be difficult to know what inputs enter a cell for systems such as unculturable organisms, symbiotic organisms, or microbial communities, thus affecting the way a model is gap-filled. One approach that has used machine learning to address the issue of inputs focused on analyzing microbial communities that can be utilized for microbial fuel cells [18]. Six machine learning algorithms were trained on four different input variables from 69 samples to predict feed substrate from genomics dataset. The four input features were built using family taxonomic level and phylum level features, as well as dimensionally reduced features using Principal Coordinate Analysis and Non-metric Multidimensional Scaling of the dataset. The model based on a neural network algorithm provided the highest accuracy of $93 \pm 6\%$. Reapplication of this or similar approaches may help guide the decision-making process in gap-filling and input constraints for complex or poorly characterized systems. However, to ensure its use for practical biosensing applications, significantly more samples and input features need to be considered in model training and performance evaluation.

## Data integration

After the gap-filling step, the contents of the stoichiometric matrix (S) are specified making it possible to run analyses such as flux balance analysis (FBA). Improvements to the predictive ability of constraint-based models have largely utilized high-throughput experimental data (i.e. omics data) integration. Successful approaches have included use of transcriptomics or proteomics with mixed-integer linear programming (MILP) [19], integration of proteome allocation theory [20], simulated

annealing [21], and parsimonious FBA (pFBA) [22]. A brief review on the integration of transcriptomics with CBM is provided in [23].

Utilization of machine learning approaches for analysis of high-throughput data is potentially fruitful and has been conducted for several data types. Metabolomic data can have problems with metabolite identification and classification as well as predicting common hidden features between metabolites. A machine learning approach named FIngerID learns the metabolite fingerprint for identifying metabolites from mass spectroscopy data using a kernel-based approach [24]. Next, it outputs a ranked list of candidate metabolites matched with molecular databases. Metabolite identification is a laborious and time-intensive step, and machine learning methods can shorten this time-frame and improve the accuracy of metabolites identification. Another knowledge-based machine learning tool for metabolite identification called BioTransformer [25•] consists of two components, a metabolic prediction tool (BMPT) and a metabolite identification tool (BMIT). BMPT uses a machine learning model based on random forest and ensemble prediction for prediction of substrate sensitivity and filtering for BMIT resulting in high precision and recall values. An application-driven analysis of metabolomic data and machine learning used multiple linear regression (MLR) to predict the metabolome from kinase knockout enzyme expression proteome data [26] where multifactorial relationships between enzyme expression and metabolite concentration were found.

Constraint-based models are often used to predict metabolic fluxes, so fluxomic data can be the most direct experimental data to use with CBMs. Machine learning on fluxomics data discovered the hidden relationships between genetic factors and reaction fluxes and accelerated the flux quantification of the model [27]. This study modeled flux prediction as a regression problem using five categorical features and sixteen continuous features; next, they applied various machine learning algorithms such as Support vector machine (SVM), K-Nearest Neighbors (KNN) and Decision tree on fluxomic data to formulate a quadratic optimization problem for flux correction. Recently, data-driven methods (e.g. data augmentation and ensemble learning that alleviates the challenges of sparse, non-standardized, and incomplete data sets) were also integrated with genome scale metabolic models to provide influential features and bioprocessing variables using multiple correspondence and principal component analysis for assessment of microbial bio-production in terms of fermentation yield, titer and rate [28].

## Multi-omic data integration

In addition to the application of machine learning to single data types, analysis of heterogeneous data (e.g. transcriptomics and omics) in different conditions can reduce the effects of noise and highlight significant features. Multiview or multimodal learning algorithms have become increasingly useful for multi-source data integration. Multiview machine learning algorithms reconstruct a comprehensive view of data by fusing different sources of data. Data fusion from multiple sources is not straightforward, due to the different inherent biases and noise of data sources, and often carry complementarity information. Wang *et al.* solved this problem by constructing a network of samples from different data sources individually and then efficiently fusing them using a nonlinear combination method [29]. This type of data fusion is known as kernel based data fusion or full fusion, where the similarity measure between samples is first mapped to a proper nonlinear kernel similarity function and the similarity kernel matrix is next iteratively fused to achieve a single comprehensive view of the data. Another similar fusion algorithm is MCGS (multiview Consensus graph clustering) where the consensus graph is constructed by minimizing the disagreement between the kernel matrices by imposing constraints on the rank of the Laplacian matrix [30]. Though these methods were not directly applied to metabolic reconstruction as yet, they are generic enough and hold immense potential towards improving metabolic models. An excellent review of this topic can be found in [31••].

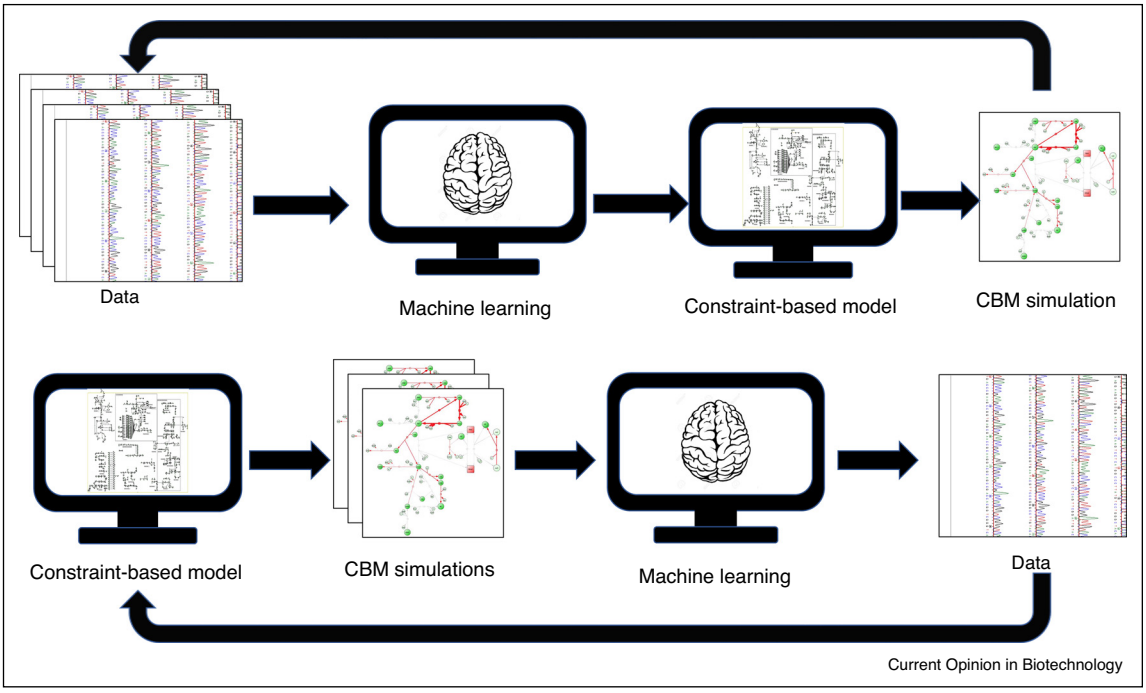
One problem with these types of kernel-based data fusion methods is that all data sources must represent a kernel matrix of similar types. This may result in loss of information and undermine cross-domain relations. In this context, intermediate fusion or partial integration becomes relevant which uses a single joint model for multisource data. Intermediate fusion methods like DFMF (Data Fusion by Matrix Factorization) have been successfully applied to predict disease-disease associations [32]. DFMF considers a single object type ($r_i$) for each data source and pairs of object type ($r_i$,rj) to relate two data sources. Next, it performs data fusion on the pair-wise relation matrix using a three-factor penalized matrix factorization. Moreover, this method requires a little transformation of the input data and can also handle missing objects or relations.

## Kinetics

Another significant area where efforts were made to improve constraint-based models is the ability to simulate cellular dynamics. These include methods such as dynamic FBA (dFBA) [33], extensions of dFBA [34] and [35], unsteady-state FBA (uFBA) [36], and DynamicME that incorporates dynamic prediction of protein expression [37•].

Application of machine learning on top of FBA has shown some promise to estimate kinetic parameters for the genome-scale model. Andreozzi *et al.* used machine

**Figure 2**




Proposed high-level iterative schemes integrating experimental data sets with machine learning, constraint-based models, and constraint-based model simulation results. Top: Machine learning is applied to the input data to identify the important features for constructing reduced order constraint-based models; the CBM simulations can be iteratively matched with input data for convergence until the proper set of features are identified. Bottom: Machine learning is iteratively applied to CBM simulations to reconcile with experimental data. Interplay between the Top and Bottom parts can iteratively lead to convergence between CBM simulations, experimental data and machine learning based predictions.

learning, named CART, with a kinetic modeling method on steady-state flux profiles of metabolites to reduce the range of kinetic parameters [38]. Here, the machine learning algorithm was able to identify the reduced subspace of kinetic parameters where the kinetic model is feasible. Another approach used machine learning on time-series multi-omics data to automate the prediction of the model dynamics [39•]. This method first performs a derivative to extract the relationship between metabolomics and proteomics data and next feed derivative pairs to the training phase of the ML algorithm to learn the dynamics. This supervised learning method significantly outperformed a handcrafted dynamic model for the limonene pathway.

## Conclusion

As our ability to study biology in greater depth has increased, so has the need for computational tools to aid in testing knowledge, analyzing data, and predicting function. While all computational approaches have limitations, the respective strengths of constraint-based modeling and machine learning make them natural complementary approaches where identification of significant features/patterns from machine learning can be evaluated through the mechanistic framework of constraint-based modeling. To date, there has been limited work to directly integrate machine learning and constraint-based modeling approaches for cellular systems. One could envision iterative integrative schemes (Figure 2) where machine learning could be used to analyze data that could be used as input constraints in a constraint-based model with a reconciliation step between CBM simulation results and experimental phenotype. Additionally, constraint-based model simulation results could also be analyzed by machine learning and reconciled with experimental data. Feedback from both parallel paths could iteratively refine a constraint-based model until there is consistency between experimental data, machine learning results, and constraint-based model simulations.

The first steps to the realization of this systems-level approach have started through studies described above focused on specific components or data types. Incremental improvements to constraint-based models and our collective biological knowledge should naturally occur as machine learning approaches are applied to analyze more data sets. As machine learning analyses cover a broader range of a single data type (e.g. genomics) more confidence will be gained in results that are consistent (i.e. gene annotation). Significant challenges remain in applying machine learning to analyze heterogeneous data but there is potential for significant discoveries through multi-data approaches.

In addition to the need for further use of machine learning for data analysis/integration, other facets of constraint-based modeling can be addressed. Use of machine learning approaches can help prune or extrapolate time course data to help improve kinetic constraint-based model predictions. Additional features including thermodynamic feasibility [40] as well as protein cost and kinetic variability [41] can identify the most likely pathways from a less well-characterized set of reactions and result in better feedback to the ML based annotation or gap-filling steps.

Ultimately, the model building process and improvements made to the process are intended to produce the best possible constraint-based model that can be used with confidence to analyze and predict biological function. It is difficult to quickly externally gauge the inherent 'quality' of a newly built constraint-based model. If model predictions do not reasonably predict biological function, is the source of the discrepancy found in the model contents, the computational algorithm used to make the prediction, or in some biological variation? Some discussion has occurred to address the potential for computational error [42] and the potential for identifying interesting biological hypotheses is one of the main goals of building and analyzing these types of models. As for evaluating the model contents, there have been attempts to implement standards such as the development of SBML level 3 [43] and quality scoring metrics, such as Memote [44] and [45]. Eventually, the true gauge for the quality of a model will need to reflect its functional utility and machine learning has the potential to aid in improving numerous facets of a constraint-based model.

## Conflict of interest statement

Nothing declared.

## Acknowledgement

This work was partially supported by NSF grant #1802588 to Dr Preetam Ghosh.

## References and recommended reading

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

- • of special interest
- •• of outstanding interest


1. Stalidzans E, Seiman A, Peebo K, Komasilovs V, Pentjuss A: **Model-based metabolism design: constraints for kinetic and stoichiometric models**. *Biochem Soc Trans* 2018, **46**:261-267.
2. Ramon C, Gollub MG, Stelling J: **Integrating –omics data into genome-scale metabolic network models: principles and challenges**. *Essays Biochem* 2018, **62**:563-574.
3. Sen P, Orešič M: **Metabolic modeling of human gut microbiota on a genome scale: an overview**. *Metabolites* 2019, **9**.
4. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93-121.
5. Aziz RK, Bartels D, Best A, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M *et al.*: **The RAST server: rapid annotations using subsystems technology**. *BMC Genomics* 2008, **9**:1-15.
6. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinf* 2010, **11**.
7. Faria JP, Rocha M, Rocha I, Henry CS: **Methods for automated genome-scale metabolic model reconstruction**. *Biochem Soc Trans* 2018, **46**:931-936.
8. Notebaart RA, van Enckevort FHJ, Francke C, Siezen RJ, Teusink B: **Accelerating the reconstruction of genome-scale metabolic networks**. *BMC Bioinf* 2006, **7**:1-10.
9. Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M: **GEM system: automatic prototyping of cell-wide metabolic pathway models from genomes**. *BMC Bioinf* 2006, **7**:1-11.
10. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**:2068-2069.
11. • Cai Y, Yang H, Li W, Liu G, Lee PW, Tang Y: **Multiclassification prediction of enzymatic reactions for oxidoreductases and hydrolases using reaction fingerprints and machine learning methods**. *J Chem Inf Model* 2018, **58**:1169-1181.

Reaction fingerprints and machine learning algorithms were successfully employed to assign Enzyme Commission (EC) numbers to enzyme-catalyzed reactions. This paper demonstrates that machine learning has the potential to improve the crucial annotation step of constraint-based modeling while highlighting advantages over current prediction methods utilizing reaction fingerprints. Limitations and areas requiring refinement were also addressed.

12. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: **Systems approach to refining genome annotation**. *Proc Natl Acad Sci U S A* 2006, **103**:17480-17484.
13. Pan S, Reed JL: **Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries**. *Curr Opin Biotechnol* 2018, **51**:103-108.
14. Orth JD, Palsson B: **Systematizing the generation of missing metabolic knowledge**. *Biotechnol Bioeng* 2010, **107**:403-412.
15. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction**. *BMC Bioinf* 2010, **11**:15.
16. Boudellioua I, Saidi R, Hoehndorf R, Martin MJ, Solovyev V: **Prediction of metabolic pathway involvement in prokaryotic uniprotkb data by association rule mining**. *PLoS One* 2016, **11**: e0158896.
17. Guan R, Wang X, Yang MQ, Zhang Y, Zhou F, Yang C, Liang Y: **Multi-label deep learning for gene function annotation in cancer pathways**. *Sci Rep* 2018, **8**:267.
18. Cai W, Lesnik KL, Wade MJ, Heidrich ES, Wang Y, Liu H: **Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells**. *Biosens Bioelectron* 2019, **133**:64-71.
19. Shlomi T, Cabili MN, Herrgård MJ, Palsson B, Ruppin E: **Network-based prediction of human tissue-specific metabolism**. *Nat Biotechnol* 2008, **26**:1003-1010.
20. Zeng H, Yang A: **Flux balance analysis incorporating a coarse-grained proteome constraint for predicting overflow metabolism in *Escherichia coli***. *Comput Aided Chem Eng* 2019:865-870.
21. Gonzalez OR, Küper C, Jung K, Naval PC, Mendoza E: **Parameter estimation using simulated annealing for S-system models of biochemical networks**. *Bioinformatics* 2007, **23**:480-486.
22. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D *et al.*: **Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models**. *Mol Syst Biol* 2010, **6**.
23. Machado D, Herrgård M: **Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism**. *PLoS Comput Biol* 2014, **10**:e1003580.

24. Shen H, Zamboni N, Heinonen M, Rousu J: **Metabolite identification through machine learning— tackling CASMI challenge using FingerID**. *Metabolites* 2013, **3**:484-505.

25. • Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS: **BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification**. *J Cheminform* 2019, **11**:1-25.

BioTransformer is a open-source command line tool for rapid, accurate and comprehensive metabolite prediction.

26. Zelezniak A, Vowinckel J, Capuano F, Messner CB, Demichev V, Polowsky N, Mülleder M, Kamrad S, Klaus B, Keller MA *et al.*: **Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts**. *Cell Syst* 2018, **7**:269-283.e6.

27. Wu SG, Wang Y, Jiang W, Oyetunde T, Yao R, Zhang X, Shimizu K, Tang YJ, Bao FS: **Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming**. *PLoS Comput Biol* 2016, **12**:e1004838.

28. Oyetunde T, Liu D, Martin HG, Tang YJ: **Machine learning framework for assessment of microbial factory performance**. *PLoS One* 2019, **14**:e0210558.

29. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale**. *Nat Methods* 2014, **11**:333-337.

30. Zhan K, Nie F, Wang J, Yang Y: **Multiview consensus graph clustering**. *IEEE Trans Image Process* 2019, **28**:1261-1270.

31. •• Li Y, Wu FX, Ngom A: **A review on machine learning principles for multi-view biological data integration**. *Brief Bioinform* 2018, **19**:325-340.

This is an excellent review showcasing multiview machine learning algorithms which integrates multiple sources of data with great potential for biological modeling applications.

32. itnik M, Zupan B: **Data fusion by matrix factorization**. *IEEE Trans Pattern Anal Mach Intell* 2015, **37**:41-53.

33. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in *Escherichia coli***. *Biophys J* 2002, **83**:1331-1340.

34. Vargas FA, Pizarro F, Pérez-Correa JR, Agosin E: **Expanding a dynamic flux balance model of yeast fermentation to genome-scale**. *BMC Syst Biol* 2011, **5**:17-19.

35. Feng X, Xu Y, Chen Y, Tang YJ: **Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of shewanella oneidensis MR-1**. *PLoS Comput Biol* 2012, **8**.

36. Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson ÓE, Palsson BO: **Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics**. *Sci Rep* 2017, **7**:1-12.

37. • Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO: **DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression**. *BMC Syst Biol* 2019, **13**:1-16.

DynamicME is an algorithm for time scale simulation of metabolism and macromolecular expression. DynamicME has the ability to correctly predict time-course proteome allocation.

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

39. • Costello Z, Martin HG: **A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data**. *NPJ Syst Biol Appl* 2018, **4**:1-15.

This paper demonstrated that a machine learning approach yielded superior predictive power over a classical Michaelis–Menten model when applied to pathway dynamics. Kinetic modeling is a desired additional feature yet, often difficult to implement. Shown here is promise that machine learning methods may enhance aspects of metabolic modeling previously stymied by classical approaches.

40. Hädicke O, von Kamp A, Aydogan T, Klamt S: **OptMDFpathway: identification of metabolic pathways with maximal thermodynamic driving force and its application for analyzing the endogenous CO2 fixation potential of *Escherichia coli***. *PLoS Comput Biol* 2018, **14**:1-24.

41. Dinh HV, King ZA, Palsson BO, Feist AM: **Identification of growth-coupled production strains considering protein costs and kinetic variability**. *Metab Eng Commun* 2018, **7**:1-11.

42. Ebrahim A, Almaas E, Bauer E, Bordbar A, Burgard AP, Chang RL, Dräger A, Famili I, Feist AM, Fleming RMT *et al.*: **Do genome-scale models need exact solvers or clearer standards?** *Mol Syst Biol* 2015, **11**:831.

43. Hucka M, Bergmann FT, Hoops S, Keating SM, Sahle S, Schaff JC, Smith LP, Wilkinson DJ: **The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core**. *J Integr Bioinform* 2015, **12**:266.

44. Lieven C, Beber ME, Olivier BG, Bergmann FT, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, Diener C *et al.*: **Memote: a community driven effort towards a standardized genome-scale metabolic model test suite**. *bioRxiv* 2018 http://dx.doi.org/10.1101/350991.

45. Carey MA, Dräger A, Papin JA, Yurkovich JT: **Community standards to facilitate development and address challenges in metabolic modeling**. *bioRxiv* 2019 http://dx.doi.org/10.1101/700112.

46. Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL *et al.*: **An integrated approach to characterize genetic interaction networks in yeast metabolism**. *Nat Genet* 2011, **43**:656-662.

47. •• 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.

This paper demonstrates how accurate prediction of enzyme turnover number improves the metabolic model parameterization. Authors used machine learning on integrated data based on protein structure, enzyme biochemistry, assay condition and network context to predict the catalytic turnover number of enzymes.

48. Plaimas K, Mallm J-P, Oswald M, Svara F, Sourjik V, Eils R, Konig R: **Machine learning based analyses on metabolic networks supports high-throughput knockout screens**. *BMC Syst Biol* 2008, **2**:67.

49. Nandi S, Subramanian A, Sarkar RR: **An integrative machine learning strategy for improved prediction of essential genes in *Escherichia coli* metabolism using flux-coupled features**. *Mol Biosyst* 2017, **13**:1584-1596.

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

51. Ekins S, de Siqueira-Neto JL, Mccall LI, Sarker M, Yadav M, Ponder EL, Kallel EA, Kellar D, Chen S, Arkin M *et al.*: **Machine learning models and pathway genome data base for trypanosoma cruzi drug discovery**. *PLoS Negl Trop Dis* 2015, **9**:1-18.

52. Jervis AJ, Carbonell P, Vinaixa M, Dunstan MS, Hollywood KA, Robinson CJ, Rattray NJW, Yan C, Swainston N, Currin A *et al.*: **Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli***. *ACS Synth Biol* 2019, **8**:127-136.