Chapter 18

Machine Learning and Hybrid Methods for Metabolic Pathway Modeling

Miroslava Cuperlovic-Culf, Thao Nguyen-Tran, and Steffany A. L. Bennett

Abstract

Computational cell metabolism models seek to provide metabolic explanations of cell behavior under different conditions or following genetic alterations, help in the optimization of in vitro cell growth environments, or predict cellular behavior in vivo and in vitro. In the extremes, mechanistic models can include highly detailed descriptions of a small number of metabolic reactions or an approximate representation of an entire metabolic network. To date, all mechanistic models have required details of individual metabolic reactions, either kinetic parameters or metabolic flux, as well as information about extracellular and intracellular metabolite concentrations. Despite the extensive efforts and the increasing availability of high-quality data, required in vivo data are not available for the majority of known metabolic reactions; thus, mechanistic models are based primarily on ex vivo kinetic measurements and limited flux information. Machine learning approaches provide an alternative for derivation of functional dependencies from existing data. The increasing availability of metabolomic and lipidomic data, with growing feature coverage as well as sample set size, is expected to provide new data options needed for derivation of machine learning models of cell metabolic processes. Moreover, machine learning analysis of longitudinal data can lead to predictive models of cell behaviors over time. Conversely, machine learning models trained on steady-state data can provide descriptive models for the comparison of metabolic states in different environments or disease conditions. Additionally, inclusion of metabolic network knowledge in these analyses can further help in the development of models with limited data.

This chapter will explore the application of machine learning to the modeling of cell metabolism. We first provide a theoretical explanation of several machine learning and hybrid mechanistic machine learning methods currently being explored to model metabolism. Next, we introduce several avenues for improving these models with machine learning. Finally, we provide protocols for specific examples of the utilization of machine learning in the development of predictive cell metabolism models using metabolomic data. We describe data preprocessing, approaches for training of machine learning models for both descriptive and predictive models, and the utilization of these models in synthetic and systems biology. Detailed protocols provide a list of software tools and libraries used for these applications, step-by-step modeling protocols, troubleshooting, as well as an overview of existing limitations to these approaches.

Key words Metabolism modeling, Hybrid modeling, Metabolomics, Lipidomics, Flux analysis, Machine learning

Kumar Selvarajoo (ed.), *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology*, Methods in Molecular Biology, vol. 2553, https://doi.org/10.1007/978-1-0716-2617-7_18,
© The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2023

1 Introduction

Whether for production of biologics or bioremediation in metabolic engineering, understanding different metabolic states under physiological and disease conditions to identify new therapeutic targets or for predictive modeling of cell behavior in a changing environment, computer modeling of cell metabolism provides an in silico platform to test optimal culture conditions, intervention, or impact of target engagement. Such models have been used to advantage in multiple biopharmaceutical applications [1], drug target identifications [2], toxicogenomics including comparison of animal and human cell response [3], and, as detailed, kinetic models of simple cell systems, including red blood cells (erythrocytes) [4] and platelets [5]. These models can be further expanded into major biotechnology platforms designed to optimize the engineering of CHO cells for biologics [6] and HEK293 cells for vaccine particle production [7] and characterize the metabolic changes that influence pluripotency and stem cell fate [8].

Classical, mechanistic, cell metabolism models, generally, are either dynamic models that include detailed kinetic information for a limited number of reactions or steady-state, constrained models that simulate stationary behavior of a larger cellular, tissue, or organismal system [9]. These models are built based on biological knowledge and only for known metabolic reactions where subsets of reaction or flux parameters are optimized using data to fit specific conditions. Kinetic models allow dynamic simulation of the change in the system over time; constrained models assume the system is in steady-state, thus, only allowing simulation of the flux through reactions with the assumption of constant metabolite concentrations on the simulation timescale. When choosing between these extremes, the modeler is faced with a trade-off between the size of the model and the level of detail provided by the predicted solutions.

Different combinations of methods have been proposed to model metabolism including efforts to develop a genome-scale kinetic model combining large network coverage with detailed reaction and metabolite concentration analysis (reviewed in [10, 11]). Bringing together different types of mechanistic models, however, attempts to alleviate the deficits of constraint-based models given their lack of information about dynamic metabolite concentration and enzyme regulation while optimizing the kinetic framework to reduce shortcomings associated with nonlinearity, parameter identifiability, and uncertainty. Although these combined approaches can bring metabolism modeling closer to the optimal large scale, they fully depend on a priori biological knowledge. Moreover, the reality is that they will encompass multiple unknown parameters that require optimization or testing for

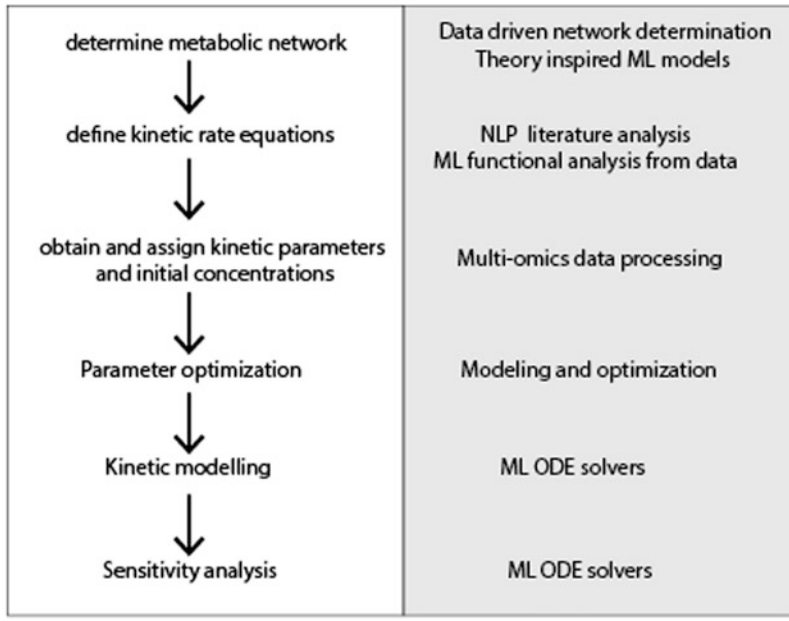

Fig. 1 Examples of possible contributions of ML in different steps of mechanistic model development. Although ML can be used for building of the complete model or for combination of model and experimental data, it can also help in determination of parameters and optimization of specific steps in development and utilization of mechanistic models. NLP natural language processing, ODE ordinary differential equations

outcome across many combinations and large ranges of values. Such hybrid models are, by their very nature, both computationally demanding and data-intensive. The application of machine learning (ML) methods to these models can address some of these issues. ML enables various types of data to be used simultaneously as well as provide more appropriate data-driven approaches that can provide more efficient parameter searches and more accurate, unbiased data-driven modeling. Thus, ML can both contribute to specific steps in the mechanistic model development, as outlined in Fig. 1, and present new global approaches for the expansion of hybrid methods combining both constrained and kinetic modeling described below.

ML combines sets of algorithms that develop predictive models through experience, i.e., through learning and functional generalization from data. ML models can be developed only from the data and do not require any prior knowledge; however, they also benefit from inclusion of domain knowledge that can optimize ML methodology for specific applications. In this way, prior knowledge can reduce training data needs. ML methods can also contribute to individual steps in the development of a mechanistic model (Fig. 1 shows some possible applications). In this context, ML is not used in modeling but helps to gather information, optimize parameters, or provide better solvers for differential equations [12]. Alternatively, ML can be further integrated into mechanistic models to

provide analysis of the results or include theoretical information for development of knowledge-inspired metabolic ML models. Therefore, combined ML/mechanistic methods into "hybrid" cell metabolism models can augment the mechanistic knowledge about metabolism and the kinetics of metabolic reactions with data-driven methods for describing unknown parts of the system or for describing, more effectively, the underlying complexity of the system. These hybrid models are a very recent innovation, with great potential to provide new insight in metabolism and its influence on organism and cellular fate.

1.1 From Mechanistic to ML Models, There, and Back Again

The first hybrid model in systems biology was presented in 2010 [13], yet this potentially transformative approach remains in its infancy due to the complexity of the problem and the lack of appropriate data for most applications. In the most general case, mathematical modeling attempts to combine both internal and external metabolic reactions and interactions with ultimate goal to provide simulation of the complete metabolic network in all its detail, including complete metabolic pathways and individual reactions as well as activation and inhibition with formal, numerical representation providing as high level of accuracy and detail as achievable within our current level of information. For well-described systems and reactions, it is possible to develop highly accurate, mechanistic models, presenting detailed dynamic reaction information and providing the change in metabolism over time via differential equations allowing inclusion of the effects of inhibitors or activators of enzyme functions. The increasing availability of longitudinal omics data will allow optimization of kinetic parameters in these models. However, for the majority of reactions, this level of information is not available, and modeling is only possible using approximations of kinetic process simulation (e.g., Michaelis-Menten equation) or by reducing studies by assuming steady state and constraining potential responses, thereby making it possible to model a larger number of reactions. Hybrid models employing ML have been fueled by the increasing availability of large amounts of biomolecular data. ML models increase calculation speed, but, even more importantly, ML can assist in creating models for systems for which there is limited knowledge via a data-driven approach. ML methods can furthermore be used to combine data from different sources including multiomic data, enhance mechanistic models by providing additional in silico data, and optimize methods for parameter determination. ML methods can help in building and executing simulations to test outcome.

Kinetic models of metabolism integrate enzyme regulation and multiomic data with reaction network information to provide dynamic analyses and predictions of metabolite concentrations. These models present mechanistic representation of the processes in cells defined as a series of ordinary differential equations (ODEs)

and include details of rate expression and kinetic parameters to estimate dynamic behavior of each reaction in the model. The mathematical form of the model is shown in Eq. 1:

$$\frac{dc_i}{dt} = S \cdot V(E, C, k); i = 1, 2 \ldots n \quad (1)$$

where c_i is the concentration of metabolite i, S is the stoichiometric matrix, and V is the vector representation of reaction flux that depends on E (the enzyme abundance), C (metabolite concentrations), and k (kinetic parameters for the reaction). Equations are written for each metabolite in the system, requiring knowledge of appropriate parameters for each and every reaction. The sets of ODEs are then solved often using various approximation methods, including as two examples Michaelis-Menten or Hill kinetic equations [14]. As a result, the majority of kinetic models focus on a small subset of reactions within specific pathways. Kinetic models have been developed for a number of metabolic pathways in different organisms and are made available through dedicated repositories (listed in Table 1). While useful and effective, the possibility to develop large, genome-scale kinetic models remains challenging given issues of kinetic model nonlinearity, computational tractability, parameter identifiability, estimability, and uncertainty [10].

While kinetic information is available for a number of enzymes in several detailed databases [15, 16] (reviewed in Table 1), the majority of kinetic constants have been measured ex vivo. Without empirical validation, it is possible that they inadequately represent the in vivo situation. More accurate determination of kinetic parameters requires optimization from data; however, models generally have problems in identifying and optimizing large numbers of parameters given nonlinear mechanistic rate equations. Simplified kinetic models have been explored for different applications by either reducing the size of the pathway space or simplifying kinetic equations. Such approaches require optimization of these approximate parameters for each case. Improvements in the optimization and fitting of models to data have been proposed with methods such as approximate Bayesian computation (ABC) [17] presented as a way to improve fitting strategy by sampling values from an approximation of the posterior distribution while not calculating explicitly the likelihood function.

The alternative to kinetic models, constraint-based modeling, lacks the representation of metabolite concentration and enzyme regulation afforded by kinetic models. Instead, these so-called genome-scale metabolic models (GEMs) combine gene sequence information with omics data to provide a map of intracellular metabolism for an organism through calculation of the stoichiometric matrix. GEMs have been used for a number of different applications, for example, flux balance analysis (FBA) [18] or metabolic balance analysis (MBA) [19] as well as testing of synthetic

Table 1
Examples of resources available for model development, ML examples, as well as metabolic models

Metabolism model development	Software application	Examples of applications in cell culture metabolomics
Bayesian modeling	GRASP [17]	Methionine cycle modeling using approximate Bayesian computation [17]
Logical modeling	CellNetOptimizer (http://www.cellnopt.org) GINsim (http://ginsim.org)	Combination of cell line proteomics and metabolomics data logic mechanistic l modeling to explain heterogeneous drug response in cellular cholesterol regulation [62]
Dynamic modeling through ordinary differential equations	COPASI [63] CellDesigner [64] VCell [65]	Many examples of COPASI's use in biotechnology cell modeling are reviewed in [66] recent example of hybrid cybernetic modeling that combines dynamic modeling between different metabolic states for CHO cells [67]
Stochastic modeling	COPASI [63] StochKit [68] MaBoSS (http://maboss.curie.fr)	Theoretical foundation to study metabolism in conjunction with stochastic enzyme expression has been presented showing metabolic heterogeneity resulting from enzyme-level stochasticity [69]
Stoichiometric modeling	COBRA [57] CobraPy [57, 70] Raven 2.0 [58] Merlin [71]	Genome-scale stoichiometric reconstructions and computational models of mammalian metabolism particularly for CHO cells coupled to protein secretion [72]
Agent-based modeling	ARCADE [73]	Extensive review of agent-based methods for cancer cell modeling [37]
ML tools	**Software application**	**Examples of some application in cell culture metabolomics**
Longitudinal GPR (LonGP)	https://github.com/chengl7/LonGP [40]	Additive GPR method for non-parametric analysis of longitudinal data
LSTM used in metabolism modeling	https://github.com/youlab/pattern_prediction_NN_Shangying [37]	LSTM for improvement of parameter modeling based on mechanistic models
Metabolism model database	**Software application**	**Type of resource**
BioModels	https://www.ebi.ac.uk/biomodels [74]	Model repository
SABIO-RK	http://sabio.h-its.org/ [75]	Kinetic information
BRENDA	https://www.brenda-enzymes.org/ [16]	Kinetic information
eQuilibrator	https://equilibrator.weizmann.ac.il/ [76]	Database of biochemical equilibrium constants and Gibbs free energies

lethality of genes [20] and determination of off-target drug effects [21]. GEMs are built on a network connection of all metabolic reactions that are known to occur in an organism combining metabolites, genes, and protein information to inform observed changes in metabolite concentrations across conditions.

The potential to determine and work from the entire metabolic reaction network derived directly from genome information opens an opportunity for building complete metabolic maps for any organism as well as subsets of metabolic networks for different biological systems. GEMs can simulate flux for all known metabolites. Additionally, they can provide a platform for multiomic analysis as well as a system for an evaluation of the complete metabolome space with sparse metabolomic profiling data. However, their reaction maps are often underdetermined, with more reactions than metabolites; thus, they generate many possible solutions often too complex for the majority of applications [1]. A number of approaches to address this issue and simplify these models for specific applications include the utilization of transcriptomic, proteomic, and metabolomic data to remove unlikely reactions as well as the addition of biological, physical, or chemical constraints [22–24]. Gene expression data is commonly used to extract the subset of reactions that are active in a specific situation and silence reactions catalyzed by enzymes that are not expressed. Although this approach is efficient, it makes a very serious assumption that gene expression activity measured at a given time point in a mixture of cells is linked to gene-protein-reaction network at steady state. This assumption is an oversimplification of the highly complex relationship between proteins, metabolite fluxes, and gene expression. As an example, the most complete GEM for metabolism of human cells – Recon3D – provides a network of 10,600 reactions linking 5835 metabolites and 2248 genes [25]. Recon3D provides a very good coverage of hydrophilic metabolites; however, while it includes a number of lipid pathways, its coverage of the lipidome is essentially incomplete, making it difficult to extend beyond metabolomics.

The lack of network solutions for lipidomic data makes lipidomics highly amendable to data-driven modeling. Development of mechanistic lipid metabolism kinetic models or a complete representation of lipid processes via GEMs remains highly challenging due to the diversity of lipid functions and their enzymes. As classified by the LIPID MAPS consortium [26], lipids are divided into eight categories and further subdivided into multiple classes, subclasses, divisions, and molecular species each with specific roles and synthesized or remodeled by overlapping enzymatic pathways. Current estimate of the number of lipid species in biological life ranges from 9000 to 100,000 [27]. This diversity in lipid structures and functions makes the mapping of all interconnections of lipids impossible as of today. In addition, the enzymes which regulate

lipids are promiscuous, catalyzing several different reactions with different specificities for the hydrocarbon chains that define lipid identities [28]. Without detailed substrate affinities, it is difficult to predict which lipids at the molecular level will be impacted by a change in condition or state. As a further challenge to all metabolomic modeling, cellular reactions are compartmentalized, with enzymes localizing to specific organelles within cells and to specific tissues within an organism. Thus, modeling must consider not only lipid abundances and enzymatic function but also their transport and, ideally, their subcellular concentrations. As an example, acid ceramidase encoded by *ASAH1* localizes to the lysosome and catalyzes the hydrolysis of ceramides to their constituent sphingoid base and free fatty acid at pH = 4.5. If the enzyme is mislocalized or lysosomal pH is alkalinized, then acid ceramidase catalyzes the reverse reaction, increasing the abundance of ceramides from a sphingoid base and a free fatty acid [29, 30]. Under physiological conditions, acid ceramidase displays substrate preference for ceramides and free fatty acids with unsaturated N-acyl hydrocarbon chains of 6–16 carbons [29].

1.2 Improving Cell Metabolism Modeling with ML

ML methods can be viewed as a combination of algorithms that learn and generalize functional dependencies from experiences, data, to identify high-order correlations and then generate predictions from data. At the most basic level, ML methods can be divided into two approaches: unsupervised and supervised. Unsupervised methods aim to determine variation, correlations, groups, or functional dependencies among samples without any input of sample labels from an external "supervisor" [31]. Supervised methods on the other hand rely on the inputted sample labels and try to develop models that predict targets and underlay the supervised group classification. Regression analysis is part of supervised ML, where algorithms are trained with input and output features to provide predictive modeling for continuous outcome (e.g., metabolite concentration over time) based on the value of one or more predictor, input value, system parameter, or condition characteristic.

Specific roles of ML in combination with mechanistic metabolism modeling are:

1. Integration of in silico mechanistic modeling results with other omics data.
2. Determination of parameters for mechanistic models from data- or theory-driven ML.

We review example methods that have been applied with success below and then provide specific methodology protocols.

1.2.1 Integration of in Silico Mechanistic Modeling Results with Other Omics Data

To achieve integration, the user must first develop and optimize a mechanistic model and then use the data obtained from this model for ML analysis of the system. ML system exploration can use the results of the simulation or combine simulation outputs with other relevant data about the system under investigation. As a proof of principle, a combination of ML and multiomics data were used to effectively predict pathway dynamics in [32, 33]. In this approach, metabolism models can be done at any scale from whole network GEM models to very small models including successful recapitulation of lipid metabolism (reviewed in [34]). Here, ML is subsequently used as a tool for data mining rather than modeling. A small number of examples, combining GEM and ML methods, have shown potential for utilization of both supervised and unsupervised ML for this type of application. As an example, when used for analysis of the effect of inhibitors on metabolism, GEMs can provide simulation of flux differences following disruption of a specific metabolic step. In this approach, ML can be used to determine major changes across the network between control and in silico "treated" cases. Shaked et al. [35] have used support vector machine (SVM) and random forest (RF) ML methods to determine major metabolic alterations from simulated flux data obtained using flux variability analysis (FVA) following inhibitory drug simulation through gene deletion analysis. In this way, ML was used to determine drug side effects on the metabolic network [35]. In another very significant application, GEM and ML models were combined during learning tasks by embedding stoichiometric constraints in the ML model training process [36]. In this approach, dynamic elementary mode regression discriminant analysis was developed to identify the most discriminant pathway activation patterns between different conditions [36].

1.2.2 Determination of Parameters for Mechanistic Models from Data- or Theory-Driven ML

Mechanistic models require optimization of parameters from data where, in the majority of cases, models cannot be solved analytically; thus, parameter optimization requires numerical methods. These methods are often slow and, for a large number of parameters with exponentially increasing number of combinations, unable to perform large-scale explorations of the complete parameter space. Yet the complete parameter space must be interrogated in order to determine global, optimal parameter or input choices. Long short-term memory (LSTM) deep learning-based network analysis method has shown promising results for the acceleration of this parameter optimization with high accuracy [37]. LSTM was introduced as a way to resolve problems of exploding/vanishing gradients that recurrent or very deep neural networks face when trying to learn long-term dependencies [38]. LSTM has been developed for processing continuous series of data [39] including time course sequences (as is usually the case in mechanistic models) or series of outcomes for combinations of input parameters

(as needed for optimization of model routine). The strength of these deep learning methods lies in the capacity to establish a map of outcomes from the training data. In the LSTM application, a small subset of data generated using mechanistic models is used to train neural network that then provides faster coverage of the parameter space to determine optimal combination for a given system.

A very detailed outline of LSTM methodology with examples of LSTM architecture used for metabolism modeling is provided in [37, 38]. In this arrangement, the cell remembers, i.e., holds, values over some time or point intervals, and the gates control and regulate the flow of information into the cell. LSTM is ultimately built from a set of recurrently connected subnetworks where each block maintains its state and regulates information flow through its nonlinear gating units. In the applications reviewed in [37, 38], LSTM is used to determine mechanistic model input parameters as it was able to search through a larger space of parameter options with a relatively small training set of random parameters and mechanistic model predicted molecular outputs. LSTM networks were shown to provide reliable and, most importantly, novel patterns of parameters suggesting that they are not limited to passive repetition of the training information but provide real mapping between input and output parameters. In this approach, neural network model building focuses on an empirical mapping of combinations of input parameters to system outputs of interest and provides a much faster way to search input parameter space while, at the same time, providing very accurate models for output parameters. For exploration, Vanilla LSTM is readily available in Python or MATLAB applications.

An alternative approach to training ML models with data and mechanistic models is to use biological knowledge to develop more appropriate ML models that can then be trained with smaller datasets providing knowledge-constrained modeling. Gaussian process regression (GPR) is a method of great interest in this type of application. In GPR, analysis and modeling of time-series data and the determination of parameters and models can be viewed as a regression problem where the goal of inference is to determine the putative form of the time-dependent function and to obtain the probability distribution of the dependent value on the variable. In the sense of metabolism modeling, regression problems would take the form of $c_1(t) = f(c_2(t)) + \varepsilon$. This functional dependence determination can be viewed as a curve fitting that assumes that c_1 is ordered by c_2, where c_2 is a function of time. GPR models can provide nonlinear system modeling, can be trained with smaller datasets, and can automatically output values that include the variance and confidence interval of the model. In addition, prior knowledge can be incorporated into the GPR model before training through optimization of covariance and kernel function. Here,

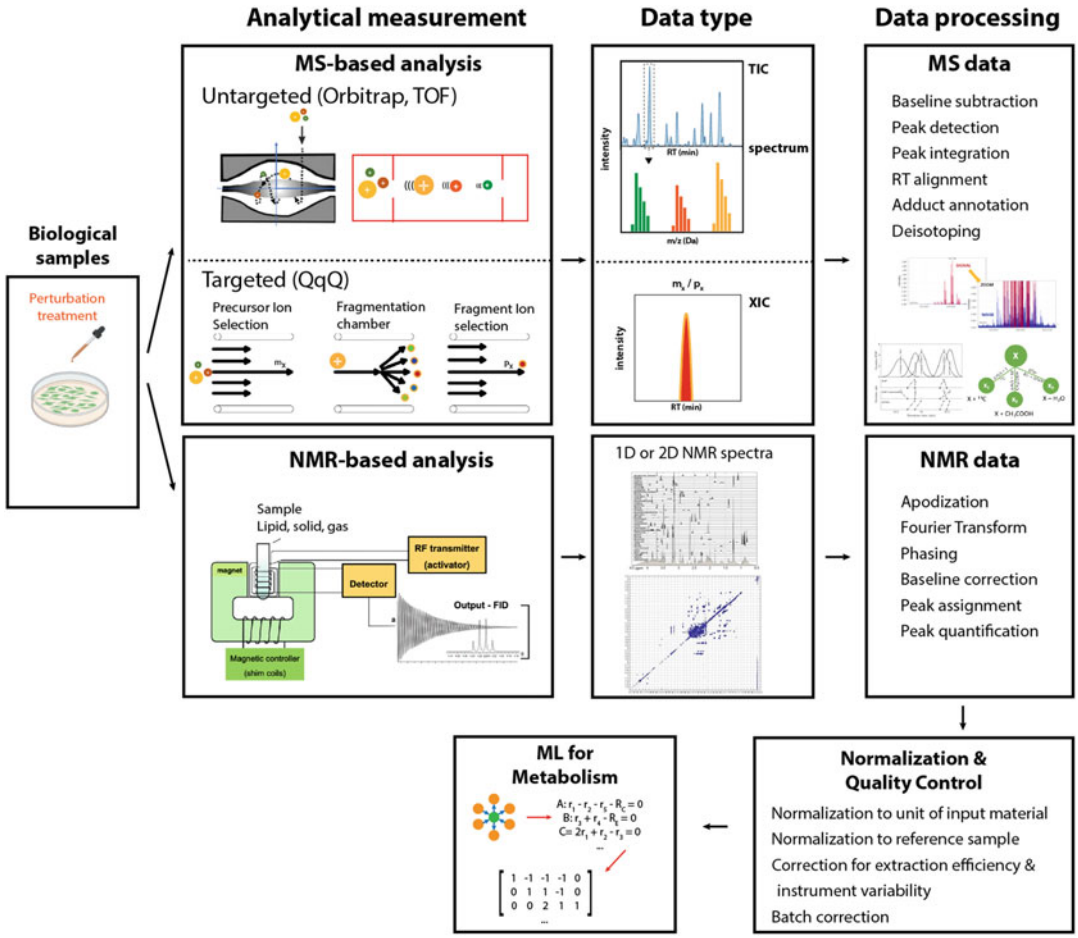

Fig. 2 Brief outline of two approaches linking mechanistic and machine learning (ML) models for (**a**) using ML for combined analysis of simulation results and omics data and (**b**) using ML for increased parameter space search coverage in order to increase

kernels can be viewed as flexible nonlinear functions that can be optimized and developed to define how quickly the regression function will vary. A related example of utilization of GPR in modeling of longitudinal processes was recently presented in [40].

Although many different ML approaches can be combined with mechanistic modeling in a variety of ways and for a range of applications, a number of similar procedural steps are required for application of any ML method in either analysis of model-derived data or augmentation of mechanistic models. Method section lists procedures for utilization of LSTM and GPR in modeling with similar protocols required for other ML model utilizations. The Materials section below provides some software tools and links to major metabolism modeling databases. The Methods section below provides detailed protocols with Fig. 2 giving a schematic presentation of these procedures.

2 Materials

Information about Web resources providing data, information, and software for metabolic modeling that can support ML and hybrid model development is presented in Table 1.

3 Methods

3.1 Using Mechanistic Models to Produce Data for Incorporation into ML Classifiers

Development of a high-quality model relies upon (1) the intimate knowledge of the system in question, (2) the articulation of appropriate hypotheses to test the models using experimental data, and (3) a feedback workflow to inform the model for rebuilding and validation. The experimental data used for modeling should be obtained using robust, high-throughput, analytical techniques that allow for rapid identification and reliable quantification of metabolites. In this context, metabolomic and lipidomic datasets are predominantly generated by mass spectrometry (MS)-based and nuclear magnetic resonance (NMR) approaches. Brief outline of methods is shown in Fig. 3.

3.1.1 MS-Based Lipidomic and Metabolomic Data

MS offers a sensitive, quantitative, technical solution and includes the possibility of devising and coupling experiments to produce structural information of countless metabolites in a single acquisition. Considerations of data processing are as follows:

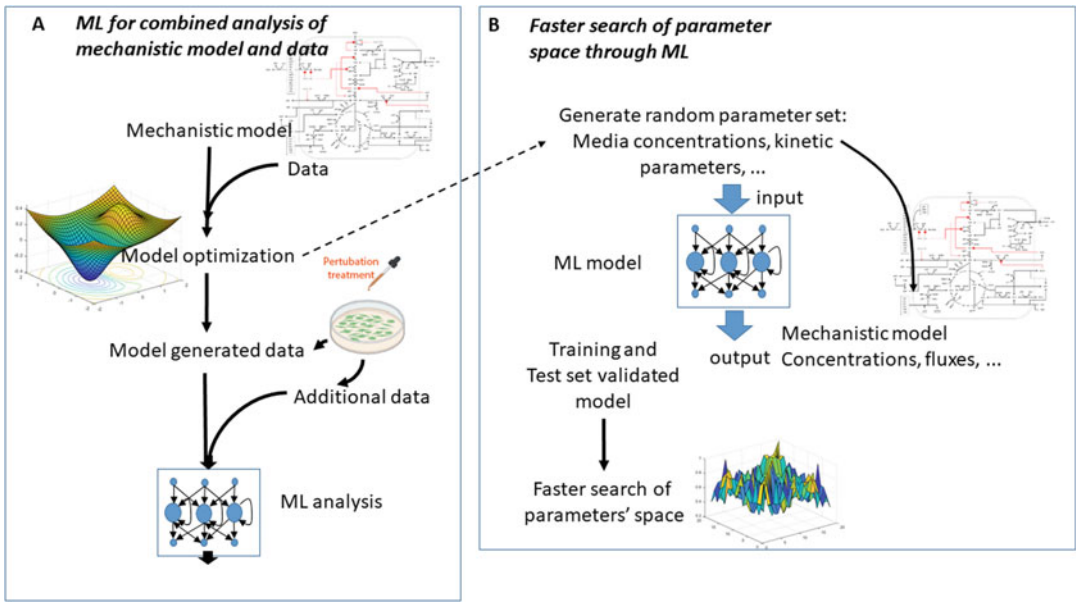

Fig. 3 Schematic representation of NMR- and MS-based metabolomics and lipidomics analysis providing data for model development. Included are major steps going from sample preparation, analytical methodologies, assignment, and data preprocessing

1. Untargeted MS analyses provide an unbiased approach to simultaneously measure a large number of metabolites or lipids within a sample without prior knowledge of lipid and metabolite categories. Strengths are the broad coverage afforded by the high-resolution mass analyzers used to discriminate lipids based on mass to charge (m/z). Weaknesses lie in the complexity of the matrices analyzed such that high abundance metabolites are favored over low abundance ones despite multiple front-end separation approaches (i.e., gas chromatography, liquid chromatography, ion mobility, etc.). Quantification is done in a semiquantitative manner. Without reducing matrix complexity, the large quantity of metabolites and lipids results in ion suppression due to co-elution, as well as in detector saturation. These limitations are offset by the high-resolution mass scanning of the precursor ion which enables identification based on m/z. A comprehensive review of the technologies is provided in [41, 42].
2. Targeted MS analyses focus on a predefined set of metabolites and lipids by parking on a diagnostic ion using triple quadrupole or QTRAP mass analyzers wherein the third quadrupole can be switched to trap fragmented ions for structural verification (reviewed in [41, 42]). By coupling chromatography to targeted MS methods, higher-resolution and more reliable quantification of metabolites can be achieved. In addition to derivatization by GC, a variety of LC methods such as normal phase, reversed phase, and hydrophilic interaction LC, ion pair chromatography is another strategy commonly employed in metabolomic analysis for the separation of ionic metabolites [41, 42]. The targeted metabolomic and lipidomic pipelines generally utilize tandem mass spectrometry to obtain high selectivity, enhanced sensitivity, and reliable quantification of metabolic targets by reducing noise from isobaric species. As such, targeted MS analyses aim to perform close to absolute quantification. This is achieved by performing tandem MS experiments such as multiple reaction monitoring (MRM, with or without schedule) to restrict analysis to a predefined set of metabolites or lipids. The data reduces complexity by quantifying a single lipid or metabolite subclass at a time (aka exploring 1000 in lieu of ~10,000 metabolites at a time). Limitations are the number of analyses required to explore the entire lipidome/metabolome. It is important to note that data from both untargeted and targeted approaches complement metabolomic modeling approaches.
3. Post-acquisition data processing in both MS approaches involves noise filtering and baseline correction, peak detection/selection, adduct annotation and deisotoping, peak alignment, and further deconvolution if necessary. Typically, in untargeted MS analyses, due to the broad coverage of

metabolites, the mass spectrum and chromatogram are saturated with noise signals. The removal of these noise signals involves establishing a set threshold and subtracting this threshold from the measurement. Similarly, this type of analysis likely will also contain detection of isotopic peaks of metabolites, which need to be removed to simplify the final dataset. For both untargeted and targeted MS analyses, specific parameters such as Gaussian smoothing, peak splitting, acceptable peak width, and retention time windows must be established for peak picking. This ensures consistency in data analysis and avoids false-positive signals. Finally, peak alignment is an important step in post-acquisition data processing to obtain correct identity assignment for each MS signal. Peak alignment and annotation are often performed by multiple peak features dependent on the separation methodology employed. Several alignment programs and algorithms have been developed for this purpose [43–47].

4. For post-acquisition normalization, the MS signal corresponding to each monitored metabolite or lipid, whether obtained in untargeted or targeted approaches, is normalized against an internal standard, critically of the same class as the analyte and either expressed as pmol equivalents of this standard or placed back onto standard curve of a known, normalized standard. Following this quantification from sample extract, the normalized MS signals need to be expressed according to the amount of starting biological material (e.g., liquid volume, cell number, tissue wet weight, etc.).

3.1.2 NMR-Based Data

NMR can be used for nondestructive, continual, or in vivo measurements in biofluids, tissues, and intact tissues and in solid, semisolids, and gas phases, with variety of different experiments and instrument profiles and measurements of multiple different nuclei (e.g., ^{1}H, ^{15}N, ^{13}C, ^{31}P), separately or simultaneously. In terms of metabolism modeling, NMR can provide longitudinal measurement for a system by either continual sampling or in vivo NMR measurement. Sample acquisition is limited with NMR experiments monitoring between 50 and 200 metabolites of high abundance (with concentrations greater than 1 μM). Briefly, steps in data derivation using NMR are as follows:

1. It is essential to select the appropriate experiment for the system of interest – for fast, high-throughput, or continual sample monitoring and quantification, preferred are 1D experiments with water suppression (e.g., 1D NOESY or 1D CPMG) that require minimal sample preprocessing (in the basic case only involving addition of NMR reference material and pH buffer), while 2D NMR provides possibility for analysis of complex systems with unknown metabolites. Sample preparation for different applications is reviewed in great detail elsewhere [48].

2. Data processing from any NMR experiment involves signal processing (apodization, Fourier transform, phasing) and normalization (relative to NMR reference). Resulting spectrum provides both peak positions (in ppm) that can be used for assignment and peak intensities that are directly related to the analytes' concentrations. With addition of internal reference, NMR can be used for absolute quantification of metabolites in the sample and comparison between different samples or time points.
3. Metabolite assignment is performed in reference standards as described in [49–51] with a number of methods available for different sample types. Important considerations are that peak position shifts due to sample properties (i.e., pH, osmolality) and that line widths change with change in the magnetic field strength, sample viscosity, and composition possibly leading to changes in peak overlaps that can lead to errors in assignments. Thus, assignment and quantification should be done using information for comparable systems with specific assignment and quantification methods available, for example, for human blood or cerebrospinal fluid [52]. Several general methods are available, but prior to their utilization, the user should adjust parameters for specific sample set (reviewed recently in [53]).

3.2 Prepare Omics Data for Further Model Development

A number of preprocessing steps are universally required for the development of mechanistic models regardless of the modeling approach and omics data collected. These include:

1. Data assignment and quantification.
2. Using either novel data or information available in published databases, high quality, and relevant longitudinal data is required to build the model and optimize parameters. For metabolism modeling, it is essential to have assigned and quantified features measured for the specific biological system under conditions of interest. Genomics, transcriptomics, and/or proteomics should be used for contextualization of genome-scale models, and metabolomics/lipidomics or flux data are used for parameter determination in kinetic models or network optimization in GENs. Kinetic parameters are available for many enzymatic reactions from ex vivo measurements (Table 1).
3. *Missing data imputation:* Due to biological or technical reasons, some features will remain unidentified or unquantified. Depending on the cause for missing data, analysts should follow different strategies. Features with a large number of missing values across conditions (of the order of 20–30% missing values) should be excluded from further analysis. Features with low abundance or undetected in specific samples where values fall below levels of detection can be imputed with a value that is

a ratio of the lowest measurable value for the species (using 1/3 or 1/5 of the lowest measured value for that feature) or set to 0. Values missing due to experimental or technical errors can be imputed using computational methods, calculating missing values based on comparison with measured values in other samples determined to be similar. Extensive benchmarking of imputation methods has been presented recently [54] showing that in the majority of tests, random forest-based imputation provides an excellent approach for missing data estimates.

4. *Data scaling from different experimental platforms*: As a variety of data sources can be used in the development of a metabolic model, it is crucial to perform appropriate normalization for each data type using either standard or internal references or relative feature levels before combining data for model building. The analyst must also decide if low and high abundance analytes are placed on the same scale to ensure equal representation. Methods have been discussed in great details previously [55, 56].

3.3 Develop a Mechanistic Model of Metabolic Processes of Interest

For the network of interest, first develop a set of ODEs or PDEs describing all reactions of interest in the model with appropriate dependencies and sink points in the format of Eq. 1. For large systems, an exact solution is not possible, and generally two approaches are applied. (1) Generate a quasi-steady-state assumption and resolve to the genome-scale model (2.b), or (2) use mathematical functions to describe $V(E, c, k)$ function applying available, measured, or estimated values for parameters (2.c):

1. For genome-scale model development, omics data provided for the system of interest (e.g., genomics, transcriptomics, proteomics, metabolomics, lipidomics) are used for the development of the personalized genome-scale FBA model. In particular, gene transcription and gene mutation information are integrated to develop contextualized genome-scale models where information about lack of function (through either mutation or gene knockdown) can be used directly to delete unrelated reactions. Methods for optimization of models are available in COBRA [57] or RAVEN [58]. Both tools operate in MATLAB or Python and provide a variety of different optimization routines for the development of contextualized models and optimization of metabolic flux. Recon3D provides a complete known metabolic network [25, 57]. The COBRA platform allows for the addition of new reactions and features.
2. For dynamic network reactions, thermodynamic information can be obtained from existing databases (Table 1) ensuring that the kinetic information is curated and is up-to-date and for the appropriate species under investigation. The functional form of $V(E, c, k)$ can be approximated using Michaelis-Menten

equation or other, more detailed formalisms and can possibly include inhibition and activation interactions. It is critical to ensure that used kinetic constants match the model type and units of metabolomic data.

3. FBA must be optimized for desired properties. This can be achieved by maximizing, for example, biomass production or cell growth using COBRA [57]. For dynamic models, kinetic parameters can be optimized from available data for the system. Optimization can be done using numerical methods or ML methods (e.g., LSMA; *see* Method B).
4. Experimentally validate FBA model by comparing predicted individual metabolite levels with matched pairs of metabolites measured in the metabolomic screen.
5. If stochastic aspects are significant for simulation, include randomness, for example, by using chemical Langevin formulation or Poisson mixture model (PMM) as recently presented [59].

3.4 Integrate Mechanistic Model of Metabolic Processes with ML

1. *Integrate* in silico *fluxomic and other omics data*: Data integration can be performed in three ways – (a) early integration, concatenation of data into a unique dataset, (b) intermediate integration wherein the ML model is built using a combined transformation of the separate input sets, and (c) late integration, where a separate model is built for each dataset and models are fused. Following integration, all data should be scaled, for example, by z-score scaling (*see* 2c). In the cross-validation process, training data should be normalized, and the same normalization parameters should be used for the test set. In the case of z-score normalization, the training set is normalized, and the mean and standard deviation values of the training set are used to normalize the test set in order to prevent information leakage.
2. *Develop ML architecture that allows analysis of integrated data*: A variety of methods are available and can be explored with method proposed below resulting from [60]. Approaches for fusing experimental results with knowledge-based in silico models through interpretable ML are reviewed here [33].

3.5 Examples of Methods

1. *Combination of data:* Data-independent ensemble ML can be used to combine all data (using the late integration approach; *see* above) including omics as well as the predicted metabolic data run by individual base learners. Subsequently, prediction and probabilities of prediction are combined for each base learner under meta-learner output with weights for each predictor. The final probability of result is $p = \sum p_i w_i$ where i is base learner with probability of prediction p_i and weight w_i. Alternatively, fluxomic data can be combined with other omics data and analyzed together using ML (with early or

intermediate data integration). Multimodal artificial neural network (MANN) method has shown the best performance for combined analysis of fluxomic and transcriptomic data [33]; however, different combinations and sizes of data require optimization of ML methods for any given application.

2. *Optimization of hyperparameters for the model*: Gradient boosting machine (GBM) algorithm can be used with Bayesian optimization for determining optimal hyperparameter values. Bayesian optimization is run in multiple iterations with fivefold cross validation used to determine the performance of selected hyperparameters. The weighted log loss must be calculated to determine performance metric for GBM and also to determine model performance on validation sets. The formula for weighted log loss is:

$$\frac{1}{N_S}\sum_{i=1}^{N_S}\left[-\left(w_R y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right)\right] \tag{2}$$

with y_i the true class label of sample i, p_i *the* predicted probability of sample i having predicted label, w_R the weight for given label, and N_s the total number of samples. Overfitting can be prevented by early stopping of the optimization process. Mean-weighted log loss with one standard error over all five folds of cross validation is used to determine the best hyperparameter set performance.

3. *Test quality of ML model using cross validation*: Data are split into training and testing and validation datasets. The training set, usually randomly selected 80% of the complete dataset, is used for training the model with a user-defined set of hyperparameters. The validation part of the data (usually the remaining 20%) is used to assess model performance according to the set of hyperparameters optimized using the training set.

4. *Test classifier performance for multiple iterations of randomized training/validation and testing data split*: Preferred performance metrices are weighted log loss (Eq. 2), area under the receiver operator curve (AUROC), as well as measures comparing true positive (TP), false positive (FP), true negative (TN), and false negative (FN) including:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FN}} \tag{4}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{5}$$

5. *Determine the importance of features in predictive models or classification:* Feature selection for both individual groups of samples and across combined samples can be done by calculating SHapley Additive exPlanations (SHAP) values for each classifier [61].

3.6 Determination of Parameters for Mechanistic Models from Data- or Theory-Driven ML

1. Develop kinetic or constrained metabolic model as listed in 3.3.
2. *Generate combinations of input parameters randomly*, if information is available, constrain parameter values within allowed range. Parameters can include, for example, kinetic constants, cell growth rate, cell motility, and media metabolite concentrations. Model output values can include metabolite concentration change over time, biomass information, and cell density as calculated by metabolic model.
3. *Develop LSTM architecture* with input layer, a fully connected layer, LSTM arrays, and two output layers, one for predicting peak values of distributions and one for predicting the normalized distributions. Vanilla LSTM is available in MATLAB and Python (TensorFlow or PyTorch). In the application of GPR, with prior information, architecture development requires selection or generation of appropriate kernel functions with possibility for additive kernel functions.
4. *Perform input and output data preprocessing*, including data scaling with, for example, min-max scaling to get all data to the 0–1 range or z-score normalization.
5. Use the calculated molecular value distribution obtained in 3.6.2. with a random combination of parameters to train ML models.
6. In the application of LSTM, parameters are used as input and molecular values as output of the neural network model. Randomly divide the data into training and test sets for cross-validation assessment of model accuracy, or use leave-one-out cross validation.
7. In LSTM, model input parameters are connected first to all neurons in the fully connected layer. Select the activation function (e.g., exponential linear unit), and initialize connection weights randomly.
8. Optimize the network using, for example, cross entropy, and calculate the cost function of the neural network using mean squared error.
9. Evaluate the model using the test set with, for example, calculation of root mean square error (RMSE) to determine the difference between LSTM and mechanistic model results.

10. For prediction of new values, use developed LSTM with new parameter inputs, and for enhanced accuracy, use the ensemble approach, for example, with Wisdom of the Crowd analysis. In this approach, calculations are rerun with the same input, and similarity scores are calculated between different predictions using RMSE, R2, or some other similarity assessment function. Each prediction is evaluated with an assessment score relative to the average prediction and the result with the minimal score, i.e., minimal deviation from the average score is selected as the final prediction result.

Acknowledgments

Work was supported in part by operating grants AI-4D-102-3 to SALB and MCC from the National Research Council AI for Design Challenge Program, RGPIN-2019-06796 to SALB from the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as an NSERC CREATE Matrix Metabolomics Training grant to SALB. TTN received an NSERC CREATE Matrix Metabolomics Graduate Scholarship.

References

1. Richelle A, David B, Demaegd D et al (2020) Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry: a process systems biology engineering perspective. NPJ Syst Biol Appl 6(1):6
2. Puniya BL, Amin R, Lichter B et al (2021) Integrative computational approach identifies drug targets in CD4(+) T-cell-mediated immune disorders. NPJ Syst Biol Appl 7(1):4
3. Blais EM, Rawls KD, Dougherty BV et al (2017) Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nat Commun 8:14250
4. Bordbar A, Jamshidi N, Palsson BO (2011) iAB-RBC-283: a proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst Biol 5:110
5. Thomas A, Rahmanian S, Bordbar A et al (2014) Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance. Sci Rep 4:3925
6. Rico J, Nantel A, Pham PL et al (2018) Kinetic model of metabolism of monoclonal antibody producing CHO cells. Current Metabolomics 6
7. Nguyen TNT, Sha S, Hong MS et al (2021) Mechanistic model for production of recombinant adeno-associated virus via triple transfection of HEK293 cells. Mol Ther Methods Clin Dev 21:642–655
8. Chandrasekaran S, Zhang J, Sun Z et al (2017) Comprehensive mapping of pluripotent stem cell metabolism using dynamic genome-scale network modeling. Cell Rep 21(10): 2965–2977
9. Cuperlovic-Culf M (2018) Machine learning methods for analysis of metabolic data and metabolic pathway modeling. Meta 8(1)
10. Srinivasan S, Cluett WR, Mahadevan R (2015) Constructing kinetic models of metabolism at genome-scales: a review. Biotechnol J 10(9): 1345–1359
11. Helmy M, Smith D, Selvarajoo K (2020) Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering. Metab Eng Commun 11:e00149
12. Borzì A (2020) Modelling with ordinary differential equations: a comprehensive approach, 1st edn. Chapman and Hall/CRC
13. von Stosch M, Peres J, de Azevedo SF et al (2010) Modelling biochemical networks with intrinsic time delays: a hybrid semi-parametric approach. BMC Syst Biol 4:131
14. Srinivasan B (2021) A guide to the Michaelis-Menten equation: steady state and beyond. FEBS J

15. Wittig U, Kania R, Golebiewski M et al (2012) SABIO-RK--database for biochemical reaction kinetics. Nucleic Acids Res 40(Database issue): D790–D796
16. Chang A, Jeske L, Ulbrich S et al (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49(D1):D498–D508
17. Saa PA, Nielsen LK (2016) Construction of feasible and accurate kinetic models of metabolism: a Bayesian approach. Sci Rep 6:29635
18. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28(3): 245–248
19. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol 6:401
20. Zhang C, Bidkhori G, Benfeitas R et al (2018) ESS: a tool for genome-scale quantification of essentiality score for reaction/genes in constraint-based modeling. Front Physiol 9: 1355
21. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4): 291–305
22. Richelle A, Joshi C, Lewis NE (2019) Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS Comput Biol 15(7):e1007185
23. Opdam S, Richelle A, Kellman B et al (2017) A systematic evaluation of methods for tailoring genome-scale metabolic models. Cell Syst 4(3): 318–329. e316
24. Aurich MK, Fleming RM, Thiele I (2016) MetaboTools: a comprehensive toolbox for analysis of genome-scale metabolic models. Front Physiol 7:327
25. Brunk E, Sahoo S, Zielinski DC et al (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol 36(3):272–281
26. Fahy E, Subramaniam S, Murphy RC et al (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50(Suppl):S9–S14
27. Shevchenko A, Simons K (2010) Lipidomics: coming to grips with lipid diversity. Nat Rev Mol Cell Biol 11(8):593–598
28. Bennett SAL, Valenzuela N, Xu H et al (2013) Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's disease. Front Physiol 4:168
29. Mao C, Obeid LM (2008) Ceramidases: regulators of cellular responses mediated by ceramide, sphingosine, and sphingosine-1-phosphate. Biochim Biophys Acta 1781(9): 424–434
30. Teichgraber V, Ulrich M, Endlich N et al (2008) Ceramide accumulation mediates inflammation, cell death and infection susceptibility in cystic fibrosis. Nat Med 14(4): 382–391
31. Bastanlar Y, Ozuysal M (2014) Introduction to machine learning. Methods Mol Biol 1107: 105–128
32. Costello Z, Martin HG (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. NPJ Syst Biol Appl 4:19
33. Culley C, Vijayakumar S, Zampieri G et al (2020) A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proc Natl Acad Sci U S A 117(31):18869–18879
34. Mc Auley MT, Mooney KM (2015) Computationally modeling lipid metabolism and aging: a mini-review. Comput Struct Biotechnol J 13: 38–46
35. Shaked I, Oberhardt MA, Atias N et al (2016) Metabolic network prediction of drug side effects. Cell Syst 2(3):209–213
36. Folch-Fortuny A, Teusink B, Hoefsloot HCJ et al (2018) Dynamic elementary mode modelling of non-steady state flux data. BMC Syst Biol 12(1):71
37. Metzcar J, Wang Y, Heiland R et al (2019) A review of cell-based computational modeling in cancer biology. JCO Clin Cancer Inform 3: 1–13
38. Van Houdt G, Mosquera C, Nápoles G (2020) A review on the long short-term memory model. Artif Intell Rev 53:5929–5955
39. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8): 1735–1780
40. Cheng L, Ramchandran S, Vatanen T et al (2019) An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat Commun 10(1): 1798
41. Mass spectrometry-based lipidomics approaches (2016) In: Hsu F-F (ed) Lipidomics. pp 53–88
42. Lipidomics (2017) Springer Protocols
43. Chitpin JG, Surendra A, Nguyen TT et al (2021) BATL: Bayesian annotations for targeted lipidomics. Bioinformatics. in press
44. Tsugawa H, Arita M, Kanazawa M et al (2013) MRMPROBS: a data assessment and metabolite identification tool for large-scale multiple

reaction monitoring based widely targeted metabolomics. Anal Chem 85(10):5191–5199

45. Domingo-Almenara X, Montenegro-Burke JR, Ivanisevic J et al (2018) XCMS-MRM and METLIN-MRM: a cloud library and public resource for targeted analysis of small molecules. Nat Methods 15(9):681–684
46. Niu W, Knight E, Xia Q et al (2014) Comparative evaluation of eight software programs for alignment of gas chromatography-mass spectrometry chromatograms in metabolomics experiments. J Chromatogr A 1374:199–206
47. Wang Y, Ma L, Zhang M et al (2019) A simple method for peak alignment using relative retention time related to an inherent peak in liquid chromatography-mass spectrometry-based metabolomics. J Chromatogr Sci 57(1): 9–16
48. Lin CY, Wu H, Tjeerdema RS et al (2007) Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. Metabolomics 3(1):55–67
49. Wishart DS, Jewison T, Guo AC et al (2013) HMDB 3.0--the human metabolome database in 2013. Nucleic Acids Res 41(Database issue): D801–D807
50. Velankar S, Burley SK, Kurisu G et al (2021) The protein data bank archive. Methods Mol Biol 2305:3–21
51. Romero PR, Kobayashi N, Wedell JR et al (2020) BioMagResBank (BMRB) as a resource for structural biology. Methods Mol Biol 2112: 187–218
52. Ravanbakhsh S, Liu P, Bjorndahl TC et al (2015) Accurate, fully-automated NMR spectral profiling for metabolomics. PLoS One 10(5):e0124219
53. Wang RCC, Campbell DA, Green JR et al (2021) Automatic 1D (1)H NMR metabolite quantification for bioreactor monitoring. Meta 11(3)
54. Jager S, Allhorn A, Biessmann F (2021) A benchmark for data imputation methods. Front Big Data 4:693674
55. Jauhiainen A, Madhu B, Narita M et al (2014) Normalization of metabolomics data with applications to correlation maps. Bioinformatics 30(15):2155–2161
56. Walach J, Filzmoser P, Hron K (2018) Data normalization and scaling: consequences for the analysis in omics sciences. Compr Anal Chem 82:165–196
57. Heirendt L, Arreckx S, Pfau T et al (2019) Creation and analysis of biochemical constraint-based models using the COBRA toolbox v.3.0. Nat Protoc 14(3):639–702
58. Wang H, Marcisauskas S, Sanchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ (2018) RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS Comput Biol 14(10): e1006541
59. Cornish-Bowden A (2014) Fundamentals of enzyme kinetics. Elsevier
60. Lewis JE, Kemp ML (2021) Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. Nat Commun 12(1):2700
61. Guyon I (2017) Advances in neural information processing system 30 pre-proceedings. NeurIPS 2017
62. Blattmann P, Henriques D, Zimmermann M et al (2017) Systems pharmacology dissection of cholesterol regulation reveals determinants of large pharmacodynamic variability between cell lines. Cell Syst 5(6):604–619.e607
63. Sahle S, Gauges R, Pahle J, et al. Simulation of Biochemical Networks Using Copasi – A Complex Pathway Simulator. In: Proceedings of the 2006 Winter Simulation Conference, 2006
64. Matsuoka Y, Funahashi A, Ghosh S et al (2014) Modeling and simulation using CellDesigner. Methods Mol Biol 1164:121–145
65. Resasco DC, Gao F, Morgan F et al (2012) Virtual cell: computational tools for modeling in cell biology. Wiley Interdiscip Rev Syst Biol Med 4(2):129–140
66. Bergmann FT, Hoops S, Klahn B et al (2017) COPASI and its applications in biotechnology. J Biotechnol 261:215–220
67. Martinez JA, Bulte DB, Contreras MA et al (2020) Dynamic modeling of CHO cell metabolism using the hybrid cybernetic approach with a novel elementary mode analysis strategy. Front Bioeng Biotechnol 8:279
68. Sanft KR, Wu S, Roh M et al (2011) StochKit2: software for discrete stochastic simulation of biochemical systems with events. Bioinformatics 27(17):2457–2458
69. Tonn MK, Thomas P, Barahona M et al (2019) Stochastic modelling reveals mechanisms of metabolic heterogeneity. Commun Biol 2:108
70. Ebrahim A, Lerman JA, Palsson BO et al (2013) COBRApy: COnstraints-based reconstruction and analysis for python. BMC Syst Biol 7:74
71. Dias O, Rocha M, Ferreira EC et al (2015) Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res 43(8): 3899–3910
72. Gutierrez JM, Feizi A, Li S et al (2020) Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. Nat Commun 11(1):68

73. Yu JS, Bagheri N (2020) Agent-based models predict emergent behavior of heterogeneous cell populations in dynamic microenvironments. Front Bioeng Biotechnol 8:249
74. Malik-Sheriff RS, Glont M, Nguyen TVN et al (2020) BioModels-15 years of sharing computational models in life science. Nucleic Acids Res 48(D1):D407–D415
75. Wittig U, Rey M, Weidemann A et al (2018) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res 46(D1):D656–D660
76. Flamholz A, Noor E, Bar-Even A et al (2012) eQuilibrator--the biochemical thermodynamics calculator. Nucleic Acids Res 40(Database issue):D770–D775