

Machine and deep learning meet genome-scale metabolic modelling

Guido Zampieri¹, Supreeta Vijayakumar¹, Elisabeth Yaneske¹, Claudio Angione^{1,2,*}

1 Department of Computer Science and Information Systems, Teesside University, Middlesbrough, TS1 3BX, United Kingdom

2 Healthcare Innovation Centre, Teesside University, Middlesbrough, TS1 3BX, United Kingdom

* Corresponding author: c.angione@tees.ac.uk

Abstract

Omic data analysis is steadily growing as a driver of basic and applied molecular biology research. Core to the interpretation of complex and heterogeneous biological phenotypes are computational approaches in the fields of statistics and machine learning. In parallel, constraint-based metabolic modelling has established itself as the main tool to investigate large-scale relationships between genotype, phenotype and environment. The development and application of these methodological frameworks have occurred independently for the most part, whereas the potential of their integration for biological, biomedical and biotechnological research is less known. Here we describe how machine learning and constraint-based modelling can be combined, reviewing recent works at the intersection of both domains and discussing the mathematical and practical aspects involved. We overlap systematic classifications from both frameworks, making them accessible to non-experts. Finally, we delineate potential future scenarios, propose new joint theoretical frameworks and suggest concrete points of investigation for this joint subfield. A multi-view approach merging experimental and knowledge-driven omic data through machine learning methods can incorporate key mechanistic information in an otherwise biology-agnostic learning process.

Introduction

Today, the search for biological mechanisms at molecular scale can leverage an unprecedented amount of information. With the recent development of high-throughput technologies, data collection has received an enormous impulse that has radically changed the perspective towards molecular biology. The main protagonist of this shift is *omic* data, namely experimental profiles with large coverage over multiple biological domains. Several levels of knowledge have become associated with emerging omic technologies [1–3]. The most widespread to date include DNA sequencing (genomics), microarrays and RNA sequencing (transcriptomics), DNA methylation and histone modifications (epigenomics), and protein or metabolite mass spectrometry (proteomics and metabolomics). As technology moves forward, its associated costs decrease and a growing wealth of data is being generated. Omic data therefore provide direct and convenient access to genetic variability and cellular activity. Undoubtedly, these datasets can be useful only if processed and deciphered through appropriate analytical tools.

A fundamental tool for the inspection, interpretation and exploitation of omic data is machine (and deep) learning, which has arguably fueled several leaps forward in

recent research and is expected to increasingly drive it in the near future [4,5]. Machine learning can be described as a set of algorithms that improve prediction accuracy through experience, given a certain processable input from which they are able to learn and generalize. Beyond their predictive power, their diffusion in bioinformatics and computational biology is also due to the limited assumptions they require compared to other statistical or computational approaches. This makes them essential in a number of tasks, ranging from the understanding of RNA folding to estimating the impact of mutations on splicing, and from the exploration of gene expression profiles to reconstructing phylogenetic trees [6–9].

In parallel, the increase in data and knowledge also favored the development of mathematical models for biomolecular systems. Contrary to data-driven approaches, hypothesis-driven analysis of large-scale omic domains typically remains prohibitive given the difficulty in pinpointing the underlying biological mechanisms. There are however some exceptions. Among the various approaches, constraint-based modelling (CBM) of metabolism is receiving a huge impulse thanks to its wide scope and flexibility, enabling mechanistic insights into the genotype-phenotype-environment relationship via integration with omic data [10]. With the recent advances in technology, we are now able to reconstruct large-scale metabolic reaction networks of prokaryotic and eukaryotic cells, and genome-scale metabolic models (GSMMs) are constantly increasing in number and variety across all life kingdoms [11–15].

These two computational frameworks have mostly been used in isolation, having distinct research communities associated with them. However, we believe that their complementary characteristics and common mathematical bases make them particularly suitable to be combined. Several works implemented this idea in various ways and were partially surveyed before [16,17]. Nevertheless, a comprehensive and systematic overview on this subject is lacking. In this work, we first review the existing approaches for integrating machine learning and CBM by compiling a thorough record of previous studies based on a combined classification of the two frameworks. Then, we suggest possible future research lines to develop new methodological approaches at the intersection of the two fields.

We therefore aim at providing a comprehensive and systematic catalog of existing interactions between CBM and machine learning, whilst distinguishing between the various methodological and applicative aspects concerned. In general, the central idea is that GSMMs can be used to generate an additional omic layer: the so-called fluxomic data. The multi-omic learning considered here then integrates this newly generated omic with the ones already available. For instance, concatenation of two datasets following normalization is a viable option. However, we will describe why in practice this may not be the best approach. Although it is outside of the scope of this work, we also remark that other computational techniques have successfully been used to build on CBM approaches and study the multi-omic nature of various organisms. These include Bayesian [18] and meta-heuristic optimization algorithms [19,20], as well as methods drawn from the theory of games [21], graphs [22], Markov chains [23] and information [24].

In the following sections, we first concisely summarize the rationale and scope of machine learning and CBM of metabolism. Next, we review and classify previous studies where these two frameworks were combined. Finally, we discuss similarities and differences among their mathematical bases, evaluate the advantages and limitations of computationally generating omic information, and outline aspects that have not been explored so far. To distinguish among the different types of mathematical models considered, throughout the text we will use the term *data-driven* to refer to machine and deep learning models, whereas *knowledge-driven* will refer to constraint-based models. If the meaning is intended to be more general, we will simply use the term

biological model. Overall, we show that mining and integrating experimental and GSMM-generated multi-omic data with machine learning techniques can unveil unknown mechanisms in a sample-specific manner, hence identifying relevant targets for biotechnology and biomedicine. Compared to approaches applying machine learning to omic data directly, we believe that a multi-view approach merging experimentally- and GSMM-generated omic data can include key mechanistic information in an otherwise biology-agnostic learning process.

Data-driven exploration of biomolecular systems

The key problem in an increasingly omic-based biology is the difficulty in extracting knowledge from large and complex datasets. This task can be conveniently tackled through machine learning algorithms, many of which can be adapted to specific settings and omic types. A number of recent developments in the application of machine learning to problems in molecular biology and biomedicine have been critically analyzed in previous surveys, along with their limitations and challenges [4–9, 25–27]. Here we concentrate on recalling the main characteristics of basic methods, with a focus on those suited for the simultaneous analysis of heterogeneous data.

Types of machine learning approaches

A fundamental distinction in machine learning is between *supervised* and *unsupervised* learning approaches. In *supervised learning*, the goal is to predict one or more targets associated with a given sample. For instance, pathogenicity resulting from mutations can be predicted starting from the sequence as a continuous risk score or a discrete risk class. Broadly speaking, supervised learning methods can be sub-divided into two main categories: classifiers, which aim to predict sample classes (e.g. pathogenic versus non-pathogenic variants), and regressors, whose task is to estimate numerical quantities (such as pathogenicity risk level). Several methods such as support vector machines (SVM) or artificial neural networks (ANN) can be used to solve both classification and regression problems.

In contrast, *unsupervised learning* allows the exploration of data collections by deconstructing variation or correlations among samples. Unsupervised learning approaches are largely classified as either association algorithms, which uncover latent rules or trends in data, or clustering algorithms, which partition samples based on their inherent and often hidden characteristics. Owing to the large volume of omic data, its condensation or simplification can prove to be useful in order to facilitate its interpretation. The most popular approaches for data dimensionality reduction are: (i) principal components analysis (PCA), which reduces data into low dimensional representations summarizing maximum variance among variables; (ii) factor analysis, which decomposes data based on latent relationships describing the correlation between variables; and (iii) matrix factorization, which breaks down data matrices into denoised constituents. For instance, non-negative matrix factorization (NMF) has been used to infer the ecological interaction networks of different gut microbial communities starting from high-dimensional metagenomic samples [28]. Finally, as regards clustering approaches, the most widespread ones fall within the k -means and hierarchical clustering families, but many other algorithms are available with several applications [29].

Machine learning for multi-omic data

A single type of data usually offers a partial view on biological complexity and limits our understanding of it. Data integration methods can facilitate the combined analysis

of multiple omic datasets, which may be heterogeneous, in order to more closely represent genotype-phenotype relationships [1, 2, 30–34]. Data may be generated starting from the same samples through different omic measurements, or even with different omic measurements across different samples measured in the same system. As omic domains are inherently interconnected, signals missing from a single dataset can be compensated for in a multi-omic data-driven model, therefore decreasing the likelihood of false negatives. At the same time, the mutual reinforcement of heterogeneous omic signals can limit false positives.

Most successful large-scale data integration approaches are meta-dimensional methods, which simultaneously span multiple data sources and can cope with variable inputs [2]. They are broadly categorized into *concatenation-based*, *transformation-based* and *model-based* integration, whose general characteristics are displayed in Figure 1. Alternatively, they are also called early-, intermediate- and late-stage integration methods respectively. In the machine learning context, algorithms dealing with data from multiple heterogeneous sources are referred to as *multi-view* or *multi-modal learning algorithms* [35, 36].

Fig 1. Omic data integration methods in machine learning. Multi-view omic data integration methods can be classified into three main domains. (a) Concatenation-based (early-stage) integration involves combining all omic data into one large matrix before applying machine learning (ML) methods to obtain a data-driven model. (b) Transformation-based (intermediate-stage) integration involves applying data transformations to obtain a uniform format, which can then permit the combination into one fused dataset. (c) Model-based (late-stage) integration involves obtaining individual machine learning models separately for each dataset before combining the outcomes, rather than combining data prior to the learning phase.

Concatenation-based integration (Figure 1a) fuses multiple data types together by concatenating data matrices into a single comprehensive matrix. Next, a learning algorithm is applied to this combined matrix. An advantage of this approach is the relative ease to apply statistical methods to any final data matrix. However, combining multiple matrices together can be challenging, due to differences in scaling or inherent biases of each data type. Normalization techniques can be used to ensure that data of different orders of magnitude converge on the same scale, but differences in noise and variance can still affect the results [37] and as such, this kind of approach can lack reliability. Moreover, a data reduction step may be necessary if too many variables make the analysis infeasible.

Transformation-based integration (Figure 1b) converts each dataset into an intermediate form such as a *graph* or a *kernel matrix* (i.e. a matrix describing a precise mathematically-defined similarity among observations) [38]. The integration of the two datasets is then performed at the level of transformed data, hence resulting in an integrative graph or kernel matrix, which is used in the learning phase. This approach has the advantage of preserving the original properties of the data and the capability to combine virtually any data structure or format by applying the appropriate transformations. The main disadvantage is the difficulty of detecting interactions among different sources, missing cross-omic correlations and therefore resulting in hard interpretation.

Model-based integration (Figure 1c) generates machine learning models from each dataset and subsequently combines them to produce a final data-driven model. This kind of integration can have even larger flexibility compared to the transformation-based approach. For instance, in patient-centered studies, it is possible to combine models coming from various groups of patients for which different data sources have been analyzed. However, this strategy can miss interaction among different

data types as well. Furthermore, it is particularly sensitive to over-fitting, so it is recommended when the data pool is extremely heterogeneous.

All these strategies are commonly applied to heterogeneous datasets obtained from different experimental sources. However, there are also computational methods for generating data on the omic levels where empirical means are inadequate. In particular, we concentrate on CBM of metabolism, as described in the following section.

Constraint-based analysis of metabolic networks

Metabolism is one of the major biological components that co-participates with the genotype in composing the phenotype. Metabolites can generate signals that are received at other omic levels, while metabolic feedbacks can compensate or modify genetic and environmental signals through complex non-intuitive routes [39,40]. Unfortunately, omic-scale metabolite probing is still immature and suffers from major limitations. The main obstacles are high biochemical heterogeneity and concentration variations that can occur within sub-second time scales and span several orders of magnitude [41,42]. In turn, metabolic reaction fluxes cannot be directly measured at large scale, and their estimation from indirect measurements presents even more challenges [43].

Genome-scale metabolic models (GSMMs)

Despite these experimental difficulties, metabolism remains the domain in molecular biology with the vastest knowledge, accumulated over the past century. Reconstructions of entire metabolic reaction networks have immediately followed after completing the first genomes in the late 1990s [44,45]. GSMMs are mathematical representations of such networks and their relationships with associated enzymes and encoding genes, comprising the metabolic functionality of a cell [46]. A vast range of computational methods have been developed upon the framework of GSMMs to investigate interactions between genotype, environment and phenotype [17,47,48]. Acting as integrative platforms for multi-omic data, they can also help identify non-intuitive phenomena in metabolism [49]. Importantly, they also enable evaluation of the complete metabolic state of cell populations even when metabolome profiling is infeasible.

The mathematical framework of GSMMs is grounded on two physical assumptions. First, mass and charge conservation, which guarantees that the total mass of produced substrates equals the total mass of those consumed. Second, the system must be at steady state, meaning that internal metabolite concentrations do not change over time. The steady-state assumption differentiates CBM from the modelling based on ordinary differential equations. The latter allows the study of metabolic systems in dynamical conditions, but it is computationally expensive and requires detailed knowledge of initial metabolic conditions and kinetic reaction parameters. For these reasons, it is only feasible for small systems, and therefore cannot capture long-range phenomena or general metabolic reprogramming. Conversely, GSMMs are restricted to steady-state conditions, but they can span the entire cellular metabolism or even multi-cellular communities [50].

Modelling fluxes can be crucial for gaining a better understanding of both metabolic activity and wider biological phenomena [10]. At a reaction and pathway level, flux balance analysis (FBA) is currently the most widely used tool to estimate the flow of metabolites in metabolic networks [46]. FBA allows determination of the flux configuration that yields maximal or minimal rate through one or more target reactions. In its basic form, it is mathematically defined as a linear optimization problem targeting a subset of reaction fluxes (Figure 2). Usually, when no other obvious cellular objective

is involved, the maximization of biomass is considered as a reasonable goal for bacteria under evolutionary pressure, but also for cancer cells under a proliferative regime [51]. For other types of cells, identifying the true objective is still a challenge, therefore biomass is commonly taken as a reasonable proxy. Various FBA variants take into account further biological constraints or regularizations, and are defined as quadratic, mixed-integer or multi-level programs [52].

Fig 2. Constraint-based data integration and fluxome generation. (a) Constraint-based metabolic modelling begins with the construction of a manually-curated GSMM recording all reactions taking place in the network. (b) Coded within the structure of a GSMM is the stoichiometric matrix \mathbf{S} , denoting the involvement of metabolites in each reaction. Constraints are applied to the model to identify a given metabolic goal, represented as the objective function \mathbf{c} , and using linear or quadratic optimization to maximize or minimize this objective. The steady state assumption ($\mathbf{S}\mathbf{v} = 0$) sets the product of the stoichiometric matrix \mathbf{S} and flux vector \mathbf{v} as invariant. (c) To compute a unique flux distribution, the objective function can be regularized by subtracting a concave function from it. In addition to \mathbf{v} being restricted between default lower and upper limits (\mathbf{v}_{\min} and \mathbf{v}_{\max}), external multi-omic data θ can be used to further constrain fluxes using the mapping function $\varphi(\theta)$, hence driving the output towards condition-dependent solutions.

Condition-specific constraint-based models

In a typical constraint-based metabolic model, fluxes are the variables whose values have to be determined. Since there are usually a greater number of reactions than metabolites in a GSMM, the problem is under-determined - meaning that multiple solutions can satisfy it. In order to determine biologically meaningful solutions, it is often necessary to further refine the model by applying additional biological, physical or chemical constraints. For example, these may account for enzyme capacity and promiscuity, spatial occupation, metabolite sequestration and multiple levels of gene, transcript and protein regulation [53]. Constraints derived from experimental data are particularly useful as they are employed to build GSMMs that directly reflect observed biological conditions, e.g. those in particular tissues or pathological states.

The development of condition-specific or context-specific GSMMs constitutes a further data integration framework, as shown in Figure 2. In this case, the process starts from raw data and knowledge on cellular physiology that are aggregated and converted into a GSMM. Although in the early phases of this field global reconstructions were built by long manual efforts to aggregate and make sense of scattered information, methods to partially automate this process are now available, and in principle they can be used to construct hundreds of knowledge-driven models [54–56]. General purpose GSMMs can then be used as scaffolds for the successive integration where omic data are mapped onto it, obtaining newly refined models with additional constraints. Mapped data can be transcriptomic, proteomic and metabolomic profiles, or information on splice isoforms or codon usage, as implemented in a number of works and software packages [57, 58]. Depending on the external data introduced, it is possible to generate GSMMs that reflect specific properties or states of particular tissues, cell types, microbial strains, or even individual cells.

Transcriptional profiles are the most popular omic to build context-specific GSMMs, via an array of methods utilizing different contextualizing criteria [17, 59]. Switch-based methods utilize a gene expression threshold to turn off reactions associated with lowly expressed genes, thereby pruning the metabolic network. Conversely, valve-based methods map the transcriptional information on the constraint-based model in a

continuous fashion. There are instead fewer approaches focused on the integration of proteomic and metabolomic data [60]. Commonly, algorithms such as iMAT [61], INIT [62] and METRADE [20] provide a framework for integrating both gene and protein expression data, with IOMA providing the opportunity to integrate proteomic and metabolomic data [63]. However, more specific approaches for protein data are being developed. For instance, a method known as GECKO constructs a GSMM with enzymatic constraints using kinetic and omic data [64]. This is achieved by expanding the stoichiometric matrix of the GSMM to include rows representing enzymes and columns representing enzyme usage in reactions, while enzyme kinetics (k_{cat} values) is modelled by pseudo-stoichiometric coefficients in this matrix. Constraining protein abundance in this way has the effect of significantly reducing flux variability and improving the accuracy of the predictions. Methodology and applications for condition-specific GSMMs have been reviewed in detail elsewhere [53, 59, 65, 66].

Combining constraint-based analysis and machine learning

The integration of CBM of metabolism with machine learning is based on two key ideas. The first is that genetic and environmental perturbations propagate in a non-linear fashion through metabolic networks and assume patterns on a reaction flux level that may be used to gain mechanistic insights into several research questions. The second is that GSMMs can act as both an analytical framework to represent biological systems and generators of information to be mined. In other words, flux solutions obtained by a GSMM can be treated like additional numerical data (another omic layer) and analyzed via learning algorithms. With the knowledge-driven metabolic model being set, the information extracted from it may depend on the task of interest and on the variables deemed relevant. As a result, it is possible to leverage the whole array of techniques defined on CBM [47] (see section Constraint-based analysis of metabolic networks). Additionally, constraints at the metabolic level can be used to enhance the learning in multi-omic settings, as explained in the following.

Despite these potential advantages, such integrated methodologies have remained confined to a few studies so far. In this section - to the best of our knowledge - we outline the existing examples of integration between machine learning and CBM grouped based on the task type as shown in Table 1 and Figure 3.

Fig 3. Multi-omic data analysis by combination of constraint-based modelling with machine learning. (a) Fluxomic analysis involves FBA or related techniques performed on a general-purpose GSMM, from which the flux data obtained can be used as input for unsupervised or supervised machine learning. (b) To improve the accuracy of machine learning predictions, multi-omic datasets are obtained using high-throughput analytics, e.g. transcriptomics (DNA microarrays, RNA sequencing), proteomics (2D gel electrophoresis, stable isotope labelling, mass spectrometry) or metabolomics (NMR spectroscopy, isotopic labeling, LC-MS, GC-MS). As these datasets are obtained from different sources, they must undergo several preprocessing stages such as filtration and normalization to maintain synchronicity, account for variance and reduce noise. Condition-specific knowledge-based models are generated by introducing these multiple datasets into GSMMs to obtain more precise flux estimations, from which machine learning techniques can be applied to infer biologically-relevant patterns in the data. (c) Alternatively, machine learning can be directly applied to single- or multi-omic datasets, to produce or improve GSMMs or fluxomic data.

Supervised fluxomic analysis

The baseline case is when biological targets are predicted based solely on metabolic fluxes obtained from general purpose GSMMs. The output of FBA or related techniques can then be fed to algorithms for supervised analysis, without data integration being involved (see Table 1).

For instance, Sridhara *et al.* investigated whether bacterial growth conditions could be inferred from intracellular flux configurations [67]. Multinomial logistic regression was used in conjunction with least absolute shrinkage and selection operator (LASSO) regularization to relate growth conditions to simulated metabolic fluxes. The regression enabled prediction of growth conditions for a particular FBA solution by using internal metabolic fluxes as input, with regularization serving to select the most relevant fluxes and prevent over-fitting.

In the context of human metabolism, integration of constraint-based models and machine learning has been shown to correctly identify side-effects of inhibitory drugs with higher accuracy than baseline methods [68]. Drug-specific actions were simulated by *in silico* gene deletions, and the associated metabolic perturbations were estimated through flux variability analysis (FVA), whose results were fed to an ensemble SVM. Artificially-reproduced metabolic alterations improved the results, compared to a predictor used on drug biochemical structures. In a similar fashion but for a different objective, a deep neural network and a differential search algorithm were applied to design gene deletion interventions in *E. coli* for the production of xylitol [69]. Also in this case, FBA coupled with artificial gene knockout served as a generator of genome-scale fluxomic data. Another recent study tested a flux-based data-driven approach for the prediction of titre, production rate and yield across different bioprocessing settings [70]. Based on an ensemble of state-of-the-art machine learning techniques, flux features were shown to boost predictive accuracy in this scenario, typically characterised by sparse data.

Importantly, CBM and machine learning can be formulated as a joint problem by embedding stoichiometric constraints in a learning task. As an example of supervised method, a discriminant analysis technique based on metabolic network constraints - called dynEMR-DA - was defined to identify pathway activation patterns that best discriminate between experimental conditions [71]. The methodology expands the concept of elementary flux modes (EFMs) - which are the simplest paths in a GSMM that characterize the associated flux space - to dynamic conditions. The algorithm seeks to determine the EFMs that differ the most in terms of time evolution.

Expanding the analysis of fluxes to an ecological scale, DiMucci *et al.* developed an approach to predict interactions among bacterial species starting from temporal simulations of co-cultures through dynamic flux balance analysis (dFBA) [72]. A random forest classifier was trained on binary vectors representing the exchange reactions in each GSMM, using dFBA relative yield predictions of co-cultures with respect to independent cultures. This data-driven model allowed better generalization than the simple distance-based criterion commonly employed in microbial community studies, and also allowed inferring the metabolic exchanges underlying the predicted interactions. In another ecological context, Chien and Larsen proposed that supervised classification of niches of bacterial species can benefit from the information generated by metabolic models [73]. They reconstructed GSMMs for 21 *Pseudomonas* species living in the endosphere and rhizosphere, and simulated 12 media formulations in order to generate predictive features. A cross-comparison of SVM, ANN and NMF suggested that metabolic flux features may be more predictive than purely genomic features.

Unsupervised fluxomic analysis

The exploration and statistical characterization of fluxomic profiles extrapolated from a GSMM can be of interest to shed light on the underlying physiology. In the absence of a well-defined biological target, unsupervised machine learning approaches can generally characterize correlation or variation across multiple samples. This allows clustering metabolic states or describing them in terms of sparser sets of variables.

This was first realized by Segre *et al.*, who exploited a GSMM to explore epistasis in yeast metabolism [74]. The task was accomplished by performing agglomerative clustering on the fitness landscape of single and double deleterious mutants for all genes involved in metabolism, where the fitness was defined on FBA growth rate ratios. The analysis identified a widespread modular organization of genes into groups linked exclusively by buffering or aggravating epistatic interactions, leading the authors to extend the concepts of modularity and epistasis based on the observed inter-module connections rather than on intra-module properties. An analogous approach was employed in the context of gut microbiome ecology, where Magnúsdóttir *et al.* performed a large-scale study on the ecological interactions among community members across a combination of western or high-fibre diets and aerobic or anaerobic conditions [75]. Similarly, these interactions were evaluated in terms of hierarchical clustering of the relative growth between interacting and non-interacting pairs predicted through FBA. The microbes were then profiled based on their interactions, identifying three major subgroups enriched in species with different carbohydrate fermentation capabilities. Positive interactions were observed mainly among metabolically-distant organisms, confirming independent studies.

Furthermore, dimensionality reduction techniques can be employed to deconstruct the entire flux space associated with constraint-based models, as done for *E. coli* [76]. In this case, PCA served to filter and synthesise the variation in biochemical reaction fluxes achievable by the metabolic network. Non-trivial cross-correlations among pathway activities can be captured, and associated metabolic capabilities can be comprehensively evaluated in terms of imposed constraints.

Finally, like in the supervised scenario, the analysis of multiple flux profiles can benefit from constraining a learning objective with stoichiometric knowledge. Alternative hybrids of PCA and stoichiometric flux analysis, termed as principal elementary mode analysis (PEMA) and principal metabolic flux mode analysis (PMFA), extract flux modes generated by metabolic models which contribute the most significantly towards variance, whilst penalizing deviations from the steady state [77, 78]. These methods are able to overcome some of the shortcomings of using general PCA for the statistical interrogation of flux distributions, e.g. the overlooking of reaction stoichiometry and the need for a predefined set of pathways. PEMA was also extended to analyze non steady-state EFMs [71].

Supervised multi-omic analysis

When experimental data is available, it can be aggregated with CBM-generated fluxomes to build multi-omic sets of features and predict targets of interest. Thanks to the peculiar advantages of each individual data-integration approach, there are multiple ways to combine them depending on the questions addressed and on the available resources. One-stage integration by machine learning methods is a possibility, as described in Section Machine learning for multi-omic data.

This strategy was first investigated to predict metabolic reaction essentiality in *E. coli*. FBA-like approaches coupled with artificial gene deletions can efficiently estimate essential reactions, although this often requires precise knowledge of nutrient availability in a given condition. The essentiality is usually evaluated merely based on the biomass

accumulation rate, which may be an imprecise estimator in some cases. Plaimas and co-workers [79] investigated whether it could be possible to improve FBA predictions by combining the estimated growth rate with additional topological, genomic and transcriptomic data. By using a SVM as classifier, they successfully verified an improvement in accuracy. An analogous approach was used by Szappanos *et al.* to predict positive and negative genetic interactions in *S. cerevisiae* [80]. A random forest was trained with FBA-based fitness and genetic interaction scores in addition to a large array of gene-pair characteristics such as paralogy, protein annotations, protein interaction network topology, single deletant fitness, mRNA expression, quantitative phenotypic correlation and compartment localization. Traditional features were shown to give low precision for the majority of gene interactions, while FBA-based features brought significant improvements in predictive precision and recall, indicating that genome-scale CBM captures relevant information that is missed by gene-level traditional features. The approach was tested again in the context of gene essentiality prediction by Nandi *et al.* [81], who instead employed flux coupling analysis (FCA) as feature generator to take gene adaptability into account in varying environmental conditions [82].

However, the metabolic capabilities of a cell population vary according to environmental and genetic conditions. For the sake of prediction, it is therefore important that metabolic information extracted by GSMMs reflects the differences between these conditions. This can be achieved through the creation of condition-specific metabolic models (Section Condition-specific constraint-based models). This constraint-based integration was used for the first time by Li *et al.* to predict novel drug-reaction interactions in cancer [83]. They employed a linear programming model to enforce the agreement between gene expression and metabolic fluxes in order to determine fluxomic profiles relative to 59 cell lines, which were used for binary classification by a kernel k -nearest neighbor (kNN) model. A similar procedure was used to explore the molecular biology of aging [84]. Using the transcriptomic data from the CD4 T-cells of 499 healthy participants, personalized CD4 T-cell metabolic models and their fluxomes were obtained with a continuous gene expression map [20]. Applying elastic net regression to these individual metabolic fluxes and the chronological ages of the individuals allowed establishing metabolic age predictors and their effect sizes. Using these poly-omic predictors, the metabolic age of an individual could be defined and calculated, providing a basis for improved prediction of individual aging and life expectancy. A similar strategy was employed to metabolically and mechanistically evaluate the impact of synthetic mutations in *Pseudomonas putida* starting from corresponding gene expression measurements [85].

Effectively learning from empirical omic profiles and associated GSMM-based metabolic states necessitates fully exploiting all the variety of multi-omic analysis methods. In this case, a two-stage integration can be achieved through the creation of condition-specific GSMMs and the subsequent machine learning-based data integration. This idea was used to predict the metabolic capabilities of Chinese hamster ovary (CHO) cells for diverse growth conditions [86]. In the study, it was shown that combining fluxomic and transcriptomic data in mammalian cells can provide a better estimation of secondary metabolite production such as lactate. The pipeline includes building bioreactor-specific GSMMs and bi-level FBA optimization [20], which provided information on the metabolism associated with each growth condition. Later, both fluxomic and transcriptomic data was used to predict lactate accumulation with improved accuracy. Considering a wider omic array, Kim *et al.* developed a general framework for multi-omic inference based on various machine learning methods [87]. Their platform can be used to perform cross-omic predictions among five biological layers: transcriptomic, proteomic, metabolomic, fluxomic and phenomic. All of them

are composed of experimental data aggregated from a number of studies, except the fluxomic layer which is the result of condition-specific FBA following the integration of transcriptomic and proteomic data.

More sophisticated data-integration pipelines have also been developed. A study used a method similar to sparse group LASSO to identify phenotypic extreme currents (ECs) based on a combination of metabolic network features and gene expression data [88]. Extreme pathways are sub-pathways (i.e. a subset of largely invariant pathways in the metabolic network that consistently yield steady-state flux), which are decomposed by linking them with a given phenotype. In other words, all ECs were associated with a gene set; based on gene expression data, those displaying a statistically-significant association to a given clinical phenotype were identified. Uncertainty in the kinetic properties of enzymes is one of the main challenges in developing kinetic models of metabolism. Andreozzi *et al.* designed a strategy called iSCHRUNK to minimize such uncertainty [89], where fluxomic and metabolomic data are integrated with a GSMM to create a thermodynamically-consistent GSMM. Subsequently, decision trees are used to evaluate kinetic parameters. Finally, a recent work has used CBM to support an artificial neural network (ANN). DeepMetabolism is an ANN method that integrates unsupervised pre-training with supervised training to build a deep learning model with the ability to predict phenotypic outcomes [90]. In its five-layer autoencoder, the first input gene layer was followed by two encoder layers (protein layer, phenotype layer) and the last two layers were decoders (reconstructed protein layer, reconstructed gene layer). Connections between the layers were regulated by biological priors, with FBA used to set the connectivity between the proteomic and the phenomic layer and therefore embed metabolic knowledge in the ANN architecture.

Unsupervised multi-omic analysis

Like in the supervised case, unsupervised algorithms can be applied on heterogeneous sets of experimental and GSMM-generated omic profiles. For instance, environmental condition-specific metabolic modelling was combined with statistical modelling by Angione *et al.* to estimate the metabolic pathway activation cascade triggered by different environmental stimuli [91]. The methodology was shown to better characterize the relationships among different pathways compared to static analysis, especially those occasionally interacting depending on the environmental conditions. In the same fashion, varying genetic conditions can be characterized in terms of associated changes on the metabolic level, and potentially exploited in synthetic biology studies. For instance, decomposition of mutant-specific fluxomic profiles through PCA lead to identifying novel biomarkers for rhamnolipids production [85]. Analogously, Yaneske and Angione utilized both agglomerative hierarchical clustering (AHC) and *k*-means clustering on transcriptomic data and fluxomic profiles in order to characterize the aging process in human [84]. Subsequent comparison of the clustering between transcriptomic and fluxomic data revealed that fluxomic profiles were better predictors of chronological age and age-associated metabolic biomarkers.

Moreover, metabolism and GSMMs can be used as a basis to understand underlying genomic variation. The Gene Expression Latent Space Encoder (GEESE) is a recently proposed approach [92] where transcriptomic information is fed into a deep generative model (specifically, a variational autoencoder) combined with a GSMM. Initially, gene expression data is provided as an input to the autoencoder, returning reconstructed gene expression vectors which are then used to train an FBA approximator. The deep generative model is trained to minimize the loss between the fluxes obtained by passing the reconstructed gene expression through the approximated FBA and the fluxes generated by the real FBA, whilst keeping the weights of this approximator constant. Based on this approach, latent patterns in gene regulation could therefore be identified

while mechanistically accounting for downstream metabolic perturbations. 475

Generation of constraint-based models and fluxomic data 476

Besides analyzing fluxomes generated via CBM, machine learning can be combined with 477
CBM itself to acquire novel fluxomic information. For instance, a suite of different 478
machine learning algorithms (SVM, kNN and decision trees) was used to directly 479
predict fluxomic configurations starting from genetic and environmental factors [93]. 480
The training was performed by aggregating ¹³C metabolic flux analysis estimations with 481
associated genetic and environmental information from a cohort of studies. In a second 482
stage, the predicted flux outputs were adjusted to satisfy stoichiometric constraints 483
using quadratic optimization, in order to account for the flux balance and boost their 484
accuracy. 485

While in the previous paragraphs we have presented examples of machine learning 486
applications on FBA outputs, data mining can even be used as a preliminary step to 487
gain additional constraints for CBM. For example, Brunk *et al.* [94] applied a series of 488
multivariate analysis methods (including PCA) on metabolomic data to better 489
understand inner correlations and identify key metabolites influencing inter-strain 490
variation. Consequently, this enabled fixing sets of flux constraints inside the *E. coli* 491
GSMM and achieving a better characterization of each culture phase. This strategy also 492
allows estimating metabolic fluxes in conditions that are not directly accessible to FBA, 493
such as in unsteady-state flux balance analysis (uFBA), where multiple flux profiles 494
associated with dynamic conditions can be predicted [95]. The underlying idea is to use 495
PCA and linear regression to define constraints for a FBA model starting from 496
metabolomics data. Since whole-metabolome measurements are generally difficult to 497
achieve, uFBA also includes an algorithm to estimate unmeasured metabolite 498
concentration differences on the basis of those that are measured. The obtained 499
constraint-based model can be used for traditional FBA, FVA or related analyses in 500
dynamic conditions. Finally, an ensemble of methods was used to estimate enzyme 501
catalytic turnover bounds for a whole *E. coli* GSMM, improving its predictions on 502
proteome allocation compared to the integration of turnover rates measured *in* 503
vitro [96]. It is interesting to note that, in this case, FBA solutions associated with 504
random environmental conditions were also included in the supervised learning phase, 505
corresponding to a fluxomic analysis as described in Section Supervised fluxomic 506
analysis. 507

Table 1. Overview of previous studies that integrated constraint-based modelling and machine learning, grouped by task type.

Study	Data integration approach	Machine learning component	CBM component	Task
Supervised fluxomic analysis				
[67]	-	Regularized multinomial logistic regression	FBA	Prediction of growth conditions
[68]	-	Bagging SVM, random forest	FVA, gene deletion	Inhibitory drug side effect prediction
[69]	-	ANN	FBA, gene deletion	Prediction of xylitol production
[73]	-	SVM, ANN, NMF	FBA	Prediction of bacterial ecological niches
[72]	-	Random forest	dFBA	Prediction of ecological interactions
[71]	-	Discriminant analysis	Elementary flux modes	Identification of distinguishing metabolic patterns between conditions
[70]	-	PCA, SVM, elastic net, random forest, XGBoost, kNN, ANN, ensemble learning	FBA	Estimation of titre, production rate and yield of microbial factories
Unsupervised fluxomic analysis				

Table 1 continued from previous page

[74]	-	Hierarchical clustering	FBA	Characterization of epistasis in yeast metabolism
[76]	-	PCA	Random sampling	Decomposition of metabolic flexibility
[77]	-	PCA	Elementary flux modes	Identification of metabolic patterns
[75]	-	Hierarchical clustering	FBA	Exploration of ecological interactions
[78]	-	PCA	Stoichiometric constraints	Identification of responsive pathways
[71]	-	PCA	Elementary flux modes	Identification of metabolic patterns in dynamic conditions
Supervised multi-omic analysis				
[79]	Concatenation-based	SVM	FBA, reaction deletion	Reaction essentiality prediction
[83]	Constraint-based	Kernel kNN	Maximization of consistency between reaction activity and gene expression	Drug target prediction
[80]	Concatenation-based	Random forest, logistic regression	FBA	Genetic interactions prediction in yeast
[87]	Constraint-based, concatenation-based, model-based	RNN, LASSO regression, ensemble learning	FBA	Cross-omic states prediction in <i>E. coli</i>
[89]	Constraint-based	Decision trees	TFBA	Estimation of kinetic parameter range and identification of key enzymes
[81]	Concatenation-based	SVM-RFE	FCA	Prediction of gene essentiality
[88]	Transformation-based	Sparse-group LASSO	Extreme currents	Identification of disease-deregulated pathways
[86]	Constraint-based, concatenation-based	Elastic net regression, PCA, GLM	Bi-level FBA	Prediction of lactate production in CHO cells
[90]	Model-based	ANN, autoencoder	FBA, gene deletion	Phenotypic predictions in <i>E. coli</i> based on multi-omic data
[84]	Constraint-based	Elastic net regression	Bi-level FBA	Identification of poly-omic predictors of aging
[85]	Constraint-based	Elastic net regression	Geometric FBA	Identification of disrupted pathways in <i>P. putida</i> mutants
Unsupervised multi-omic analysis				
[91]	Constraint-based	Bayesian factor modelling	Bi-level FBA	Prediction of temporal pathway activation in <i>E. coli</i>
[84]	Constraint-based	Hierarchical clustering, <i>k</i> -means clustering	Bi-level FBA	Poly-omic characterization of aging
[85]	Constraint-based	PCA	geometric FBA	Identification of biomarkers for rhamnolipids biosynthesis
[92]	Constraint-based, model-based	ANN	Stoichiometric constraints	Interpretation of gene expression data in <i>E. coli</i>
Generation of constraint-based models and fluxomic data				
[93]	-	kNN, decision trees, SVM	Stoichiometric constraints	Metabolic flux estimation based on general genetic and environmental conditions
[94]	Constraint-based	PCA	FBA, Monte Carlo sampling	Characterization of engineered <i>E. coli</i> strains variation
[95]	Constraint-based	PCA, linear regression	FBA	Metabolic flux estimation in dynamic conditions
[96]	Concatenation-based, constraint-based	Elastic net regression, random forest, neural networks, ensemble learning	FBA, pFBA, ME model	Prediction of proteomic data

Table 1 continued from previous page

The studies reviewed here and included in the table are grouped by task type: supervised or unsupervised fluxomic analysis, supervised or unsupervised multi-omic analysis, generation of constraint-based models and fluxomic data. Each study is annotated with the methodological building blocks related to the two computational frameworks (CBM and machine learning). List of abbreviations: support vector machine (SVM), artificial neural networks (ANN), non-negative matrix factorization (NMF), principal component analysis (PCA), extreme gradient boosted trees (XGBoost), k -nearest neighbors (kNN), recurrent neural networks (RNN), least absolute shrinkage and selection operator (LASSO), SVM based on recursive feature elimination (SVM-RFE), generalized linear models (GLM), flux balance analysis (FBA), flux variability analysis (FVA), dynamic flux balance analysis (dFBA), thermodynamics-based flux balance analysis (TFBA), flux coupling analysis (FCA), parsimonious flux balance analysis (pFBA), metabolism and gene expression GSMM (ME model).

Perspective

As detailed in the previous section, a number of data- and knowledge-driven workflows can be devised depending on the research goals and on the available resources. Ideally, multi-omic settings appear the most promising for effectively grasping meaningful biological patterns, not only because of the well-known advantages of data integration, but also considering the complementary characteristics of experimental and GSMM-based data. In the following we firstly articulate this point (Section Advantages and limitations of expanding the multi-omic array *in silico*), highlighting the strengths and limitations of both omic types. In Section Emerging applications, we outline important scenarios where we believe this multi-omic machine learning framework could be applied, and which are largely or entirely unexplored so far.

At the same time, many novel integrative methods could be developed given the variety of algorithms existing within the machine/deep learning and CBM frameworks. In particular, in the last section we discuss two related aspects that we believe could inspire the design of novel integrative methods: the importance of interpretability in biological data-driven models and the connections of both CBM and machine learning to mathematical programming (Section Building on common mathematical roots: towards predictive and interpretable biological models).

Advantages and limitations of expanding the multi-omic array *in silico*

Due to their generation process, fluxomic profiles obtained through a GSMM provide an alternative and mechanistic perspective on the underlying biology compared to traditional omics. Both possess complementary benefits and drawbacks in scientific and operational terms, as outlined in the following, which make them particularly suitable for integration.

As previously pointed out, important differences exist first of all in terms of genetic coverage and prior knowledge [97]. Experimentally-generated omic data can span vast portions of the genome, transcriptome or proteome, despite the limitations of some technologies to achieve full coverage [98]. CBM is instead normally limited to metabolic networks, although extensions to other domains have been advanced [99, 100]. Second, generation of traditional omics requires no prior information, whereas GSMM construction assumes extensive knowledge of the metabolic system under consideration, although a semi-automated knowledge-driven model creation partially alleviates this burden [54–56]. On one hand, experimental data generation can be therefore more readily translated to new systems. On the other hand, experimental data is also prone to contain false-positive cues and can sometimes be superficial or ambiguous in its

biological meaning. For instance, the high expression level of a gene does not necessarily lead to an increased enzyme activity if it is part of an enzymatic complex, as it would be limited by the expression of the other genes in the complex. Conversely, GSMMs are usually highly curated and provide a mechanistic description of biological processes, linking together genes, enzymes, metabolites and reactions. GSMM are therefore able to account for isozymes and enzymatic complexes through gene-protein-reaction rules. Compared to annotations with an abstract structure, they can also describe the functional role of genes more precisely as they provide a direct representation of biochemical processes. Despite their well-defined meaning, the scope and precision of fluxes generated *in silico* is however limited by the quality of the metabolic model used, and by the available knowledge and understanding of a system, which may often be partial.

If experimentally-generated omics are the first step towards a comprehensive understanding of living systems, the use of condition-specific GSMMs can therefore help contextualize and interpret them on a large scale. This fusion can also help identify gaps or inconsistencies in knowledge-driven models and maintain a comprehensive biological scope. Likewise, errors arising during experimental measurements might be mitigated through constraint-based integration, also controlling for biological soundness.

Furthermore, cost and time factors may motivate the integration of the two data types, though distinguishing between a GSMM, its condition-specific variants, and the calculation of flux data. The initial building and curation of general-purpose knowledge-driven models can in fact be time-consuming and require up to months or even years, despite aid from computational pipelines [101]. However, if a baseline GSMM is already available, the creation of context-specific counterparts and associated flux solutions is generally fast through dedicated software. Besides, the generation of experimental omic data is notoriously cheaper than ever, but remains a non-negligible cost especially when dealing with numerous samples. For cell systems with already validated GSMMs, FBA and related techniques can therefore quickly provide an additional omic layer to integrate with the others at extremely low cost. This consideration is especially important in the case of large sample numbers, which are essential for machine learning methods to identify robust and biologically meaningful patterns.

There are nevertheless unsolved issues involving, to some extent, all omic data types and posing major limitations to studies based on them. In particular, we mention the quality of estimated biological phenomena and related biases. As previously mentioned, experimental measurement is subject to intrinsic noise and uncertainty that has to be corrected through appropriate normalization, and small numbers of technical replicates may undermine the statistical significance of the observed signals. Additionally, traditional omics are affected by sampling or technology-specific systematic errors [102], in particular batch effects [103]. In some cases, technology-specific issues can even compromise the overall data quality, like in the sequencing of PCR-challenging regions [98]. Besides, *in silico* calculation of fluxomes has to deal with uncertainty and bias on different levels as well. The steady-state assumption poses a limit to the kind of fluxomes that can be reasonably estimated [104], and in several situations it may be unclear how to choose among multiple valid flux solutions. In addition to this, uncertainties arising in experimental settings may propagate to omic-based condition-specific GSMMs. As a result, external validation of FBA-predicted fluxes is generally required, at least on the level of cellular growth or most relevant pathways. Thorough GSMM evaluations are highly beneficial for the improvement of these platforms, but they have been conducted only in a limited number of systems, such as *E. coli* and *S. cerevisiae* [105–109]. However, as a consequence of the iterative refinement of GSMMs through the accumulation of new knowledge and data, their

coverage and quality are rapidly increasing [14,110].

All these points are very important to bear in mind for a correct and meaningful analysis and interpretation of the underlying biology. Overall, knowledge-driven fluxomic data relies on strong assumptions that require cautious evaluation to ensure biological soundness. At the same time, experimental data generation has to deal with issues that in some cases risk undermining any scientific conclusions. For these reasons, signals obtained from both experimental and GSMM-based omic studies should always be thoroughly evaluated through careful study design, appropriate statistical methods and independent data (when available), without omitting negative results in downstream reports [111,112].

Emerging applications

Despite the challenges highlighted in the previous section, omic data analysis will probably remain fundamental in numerous contexts and spread to new ones. Given the complementary advantages of GSMM-generated fluxomes and other omic data, their integration has therefore the potential for many novel applicative scenarios. As long as steady-state metabolism is deemed relevant to the task at hand, CBM can be employed to extend the omic or multi-omic data array, and machine learning techniques can be used to identify hidden patterns. For instance, metabolic engineering could enormously benefit from integrative biological models, which are more efficient and cost-effective than empirical trials [113], in terms of both pathway design [20,85] and bioprocess parameters [70,86,114]. Further, CBM extensions for modelling dynamic conditions can overcome the intrinsic limitations of FBA, and open the door to another range of applications. In spite of this, only a few studies have investigated this scenario so far, as visible in Table 1.

Our survey also shows that many previous studies dealt with bacterial systems, in particular *E. coli*, which is arguably the organism with the most highly curated GSMMs. However, for several eukaryotic organisms - including human - constraint-based models are now available, with increasing scope and precision, which constitute a promising platform for integrative biological models. Initial studies have demonstrated this in the context of human aging and disorders [84,88], as well as for drug development [68,83]. The accumulating wealth of data extracted from human tissues is a particularly valuable resource, which has however yet to be fully exploited through data- and knowledge-driven approaches. This union has the potential to unveil novel clinical biomarkers and drug targets if properly implemented in omic studies.

Moreover, there are emerging research areas that are likely to require strong analytical and automation skills in the near future. In particular, we refer to those applications that require GSMMs of growing size, such as for cell populations and microbial communities. At present, CBM can be used to describe multiple cell types, tissues [115] or even the heterogeneity within cell populations [116]. Given that no cell lives in isolation, all these models will be increasingly important to understand its interactions and behavior in larger systems. At the same time, focusing on individual cells is increasingly revealing essential in fields like cancer biology, where single-cell technologies are being improved and expanded to new omic layers. Single-cell RNA-seq alone will make the amount of data generated scale up to the millions of samples, or even higher numbers [34]. The spread of these technologies can further fuel the emergence of a larger omic-data era, with the associated challenges in terms of data analysis and interpretation. Even in this context, data- and knowledge-driven computational tools appear essential to cope with these challenges.

Finally, as mentioned above, CBM is extensible to biomolecular domains other than metabolism. Efforts have particularly focused on integrated constraint-based models of gene expression and metabolism [99,100]. The formulation and validation of novel

constraints could aid in developing further methods for multi-omic data mining, but at the same time it poses challenges associated with the increase in knowledge-driven model size and heterogeneity.

In all these contexts, we believe that effectively combining machine learning and CBM allows achieving a richer and more meaningful mechanistic comprehension of inherently multi-omic system. New integrative approaches are also expected to ultimately contribute to the progress of applicative fields such as biotechnology, bioengineering and biomedicine.

Building on common mathematical roots: towards predictive and interpretable biological models

One last aspect of pressing importance concerns the trustworthiness of integrative data- and knowledge-driven models, and their capacity to produce novel insights. Interpretability is a desirable property for any mathematical model, and constitutes a particularly delicate and widespread criticality in machine learning. Indeed, most automatically-generated models are complex and provide no direct explanation for their predictions. At the same time, interpreting results, generating hypotheses and testing them is imperative to maintain scientific rigor [117]. As discussed before, in this context, model interpretation relies on model transparency and post-hoc analyses [118].

Transparency refers to the human understanding of a whole model, a learning algorithm or their parts, variables and parameters. For example, in multiple kernel learning the weights assigned to input kernels can in some sense be regarded as their contribution to the given task. Alternatively, it is possible to infer relationships between input and output through additional algorithms or reasoning applied *a posteriori*. For instance, some methods can determine data samples whose predictions are similar, or they can compute local dependencies on input features.

Note that interpretability does not uniquely depend on the data-driven model form, but also on the input preprocessing. A neural network trained on intuitively meaningful features learns a data representation that can be visualized and reasoned upon more easily than a linear model trained on heavily processed features.

It can be therefore argued that CBM constitutes a vehicle for obtaining biological knowledge in the form of coherent information equipped with mechanistic relations at a single-reaction level. From this point of view, the generation of flux data from a condition-specific GSMM can be regarded as an elaborate but transparent feature engineering step, where a fluxome is the result of combining available omics with expert knowledge and mathematical optimization. Therefore, we believe that CBM could be the key to building more interpretable machine learning models, for instance by providing variables of clear meaning [68, 72].

Perhaps even more importantly, paths for building more interpretable and mechanistically meaningful biological models exist also on a methodological level. In particular, it may be useful to consider that both machine learning methods and FBA-based approaches are grounded in mathematical programming (also called mathematical optimization), even though from distinct points of view.

In machine learning, optimization tasks target any cost function that is assumed to minimize the true predictive error and allow the final model to generalize. Like in traditional mathematical optimization research, this discipline seeks to formulate investigative questions in terms of tractable and scalable problems. In addition, other qualities such as easy implementation and interpretation are important, disregarding high accuracy and robustness across wide classes of problems [119]. Many of these goals are shared also by metabolic CBM. First, tractability and scalability remain of primary interest, especially with the growing size of GSMMs and microbial community models.

Second, highly precise solutions are not indispensable, as long as they define the phenotypic state(s) associated with prior assumptions. In fact, the use of regularization is starting to be recognized as a standard in FBA with the goal of identifying more realistic solutions, especially when the full flux distribution is used for inference or post-processing [52]. Moreover, the ease in implementation and interpretation of FBA-related approaches is usually guaranteed by embedding physical, biochemical or evolutionary assumptions.

This underlying connection can potentially be exploited to develop novel hybrid methods and provide a second way towards more transparent biological models. For instance, the learning problem can be formulated by integrating constraints borrowed from a knowledge-driven model. This idea has been already implemented both in the context of unsupervised fluxomic analysis [77, 78] and supervised fluxomic analysis [71], as described in Section Combining constraint-based analysis and machine learning. However, the existing case studies are limited and more work is needed to understand how to effectively integrate mechanistic biological information in data-driven algorithms, especially given their variety and heterogeneity. Although challenging, the idea appears particularly promising if extended to multi-view learning methods, that would thereby fully leverage an expanded combination of data and knowledge.

Conclusion

The use of machine and deep learning in computational and systems biology will keep growing in parallel with the rapid advancement of high-throughput omic technologies. However, extensions of current methodologies are needed to adapt to the heterogeneous, multi-dimensional nature of omic data. Here we have explored the joint application of machine learning and genome-scale metabolic modelling in the context of multi-omic analysis, evaluating strengths and pitfalls in developing hybrid methods that draw from both fields. Machine learning is a valuable tool for deconstructing biological complexity, for the purposes of condensing high-volume multi-omic datasets and extracting relevant outputs from them. In turn, CBM makes it possible to analyse metabolic activities associated with distinct properties or states specific to each cell, tissue or community. This is achieved through multi-omic data integration and the estimation of an additional (flux)omic layer that is closer to cellular phenotype.

Constraint-based modelling can provide ways to inject mechanistic knowledge within novel multi-view methods, aiding in the achievement of data- and knowledge-driven analysis of biological systems. Given the increasing recognition of the importance of metabolism and mechanism-aware omic data analysis in a range of biomedical and biotechnological problems, we envisage that this multi-omic machine learning approach could be useful to researchers across computational biology.

References

1. Joyce AR, Palsson BØ. The model organism as a system: integrating 'omics' data sets. *Nature reviews Molecular cell biology*. 2006;7(3):198.
2. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*. 2015;16(2):85.
3. Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. *Trends in Genetics*. 2017;33(2):155–168.

4. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 2015;16(6):321.
5. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*. 2018;15(141):20170387.
6. Zhang Y, Rajapakse JC. *Machine learning in bioinformatics*. vol. 4. John Wiley & Sons; 2009.
7. Leung MK, DeLong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE*. 2016;104(1):176–197.
8. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Molecular systems biology*. 2016;12(7):878.
9. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Briefings in bioinformatics*. 2017;18(5):851–869.
10. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*. 2014;15(2):107.
11. Durot M, Bourguignon PY, Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS microbiology reviews*. 2008;33(1):164–190.
12. de Oliveira Dal’Molin CG, Nielsen LK. Plant genome-scale metabolic reconstruction and modelling. *Current opinion in biotechnology*. 2013;24(2):271–277.
13. Geng J, Nielsen J. In silico analysis of human metabolism: Reconstruction, contextualization and application of genome-scale models. *Current Opinion in Systems Biology*. 2017;2:29–38.
14. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. *Nature biotechnology*. 2014;32(5):447.
15. Yilmaz LS, Walhout AJ. Metabolic network modeling with model organisms. *Current opinion in chemical biology*. 2017;36:32–39.
16. Cuperlovic-Culf M. Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. *Metabolites*. 2018;8(1):4.
17. Vijayakumar S, Conway M, Lió P, Angione C. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in bioinformatics*. 2017; p. bbx053.
18. Heino J, Tunyan K, Calvetti D, Somersalo E. Bayesian flux balance analysis applied to a skeletal muscle metabolic model. *Journal of theoretical biology*. 2007;248(1):91–110.
19. Machado D, Herrgård MJ. Co-evolution of strain design methods based on flux balance and elementary mode analysis. *Metabolic Engineering Communications*. 2015;2:85–92.
20. Angione C, Lió P. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific reports*. 2015;5:15147.

21. Ruppin E, Papin JA, De Figueiredo LF, Schuster S. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Current opinion in biotechnology*. 2010;21(4):502–510.
22. Angione C, Conway M, Lió P. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics*. 2016;17(4):83.
23. Spahn PN, Hansen AH, Hansen HG, Arnsdorf J, Kildegaard HF, Lewis NE. A Markov chain model for N-linked protein glycosylation—towards a low-parameter tool for model-driven glycoengineering. *Metabolic engineering*. 2016;33:52–66.
24. Pierobon M, Sakka Z, Catlett JL, Buan NR. Mutual information upper bound of molecular communication based on cell metabolism. In: *Signal Processing Advances in Wireless Communications (SPAWC), 2016 IEEE 17th International Workshop on*. IEEE; 2016. p. 1–6.
25. Bhaskar H, Hoyle DC, Singh S. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in biology and medicine*. 2006;36(10):1104–1125.
26. Tarca AL, Carey VJ, Chen Xw, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS computational biology*. 2007;3(6):e116.
27. Zeng ISL, Lumley T. Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science). *Bioinformatics and Biology Insights*. 2018;12:1177932218759292.
28. Cai Y, Gu H, Kenney T. Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization. *Microbiome*. 2017;5(1):110.
29. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering*. 2010;3:120–154.
30. Buescher JM, Driggers EM. Integration of omics: more than the sum of its parts. *Cancer & metabolism*. 2016;4(1):4.
31. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*. 2016;17(4):628–641.
32. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*. 2015;12(112):20150571.
33. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome biology*. 2017;18(1):83.
34. Colomé-Tatché M, Theis F. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*. 2018;7:54–59.
35. Sun S. A survey of multi-view machine learning. *Neural Computing and Applications*. 2013;23(7-8):2031–2038.
36. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*. 2016; p. bbw113.
37. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. *Briefings in bioinformatics*. 2015;17(5):891–901.

38. Wang X, Xing EP, Schaid DJ. Kernel methods for large-scale genomic data analysis. *Briefings in bioinformatics*. 2014;16(2):183–192.
39. Zampieri M, Sauer U. Metabolomics-driven understanding of genotype-phenotype relations in model organisms. *Current Opinion in Systems Biology*. 2017;6:28–36.
40. Yugi K, Kuroda S. Metabolism as a signal generator across trans-omic networks at distinct time scales. *Current Opinion in Systems Biology*. 2017;8:59–66.
41. Sriyudthsak K, Shiraishi F, Hirai MY. Mathematical modeling and dynamic simulation of metabolic reaction systems using metabolome time series data. *Frontiers in molecular biosciences*. 2016;3:15.
42. Aretz I, Meierhofer D. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *International journal of molecular sciences*. 2016;17(5):632.
43. Nidenführ S, Wiechert W, Nöh K. How to measure metabolic fluxes: a taxonomic guide for ¹³C fluxomics. *Current opinion in biotechnology*. 2015;34:82–90.
44. Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Journal of Biological Chemistry*. 1999;274(25):17410–17416.
45. Edwards J, Palsson B. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*. 2000;97(10):5528–5533.
46. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nature biotechnology*. 2010;28(3):245.
47. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*. 2012;10(4):291.
48. O’Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell*. 2015;161(5):971–987.
49. Ebrahim A, Brunk E, Tan J, O’Brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nature communications*. 2016;7:13091.
50. Gottstein W, Olivier BG, Bruggeman FJ, Teusink B. Constraint-based stoichiometric modelling from single organisms to microbial communities. *Journal of the Royal Society Interface*. 2016;13(124):20160627.
51. Lewis NE, Abdel-Haleem AM. The evolution of genome-scale models of cancer metabolism. *Frontiers in physiology*. 2013;4:237.
52. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nature protocols*. 2019; p. 1.
53. Jensen K, Gudmundsson S, Herrgård MJ. Enhancing Metabolic Models with Genome-Scale Experimental Data. In: *Systems Biology*. Springer; 2018. p. 337–350.

54. Thorleifsson SG, Thiele I. rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics*. 2011;27(14):2009–2010.
55. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*. 2018;46:7542–7553.
56. Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. *Biochemical Society Transactions*. 2018;46(4):931–936.
57. Yang L, Yurkovich JT, Lloyd CJ, Ebrahim A, Saunders MA, Palsson BO. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Scientific reports*. 2016;6:36734.
58. Angione C. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics*. 2018;34(3):494–501.
59. Vivek-Ananth R, Samal A. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems*. 2016;147:1–10.
60. Töpfer N, Kleessen S, Nikoloski Z. Integration of metabolomics data into metabolic networks. *Frontiers in plant science*. 2015;6:49.
61. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. *Bioinformatics*. 2010;26(24):3140–3142.
62. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology*. 2012;8(5):e1002518.
63. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*. 2010;26(12):i255–i260.
64. Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*. 2017;13(8):935.
65. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems*. 2013;9(2):167–174.
66. Fouladiha H, Marashi SA. Biomedical applications of cell-and tissue-specific metabolic network models. *Journal of biomedical informatics*. 2017;68:35–49.
67. Sridhara V, Meyer AG, Rai P, Barrick JE, Ravikumar P, Segrè D, et al. Predicting growth conditions from internal metabolic fluxes in an in-silico model of *E. coli*. *PloS one*. 2014;9(12):e114608.
68. Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E. Metabolic Network Prediction of Drug Side Effects. *Cell Systems*. 2016;2(3):209–213.
69. Yousoff SNM, Baharin A, Abdullah A. Differential Search Algorithm in Deep Neural Network for the Predictive Analysis of Xylitol Production in *Escherichia Coli*. In: *Asian Simulation Conference*. Springer; 2017. p. 53–67.

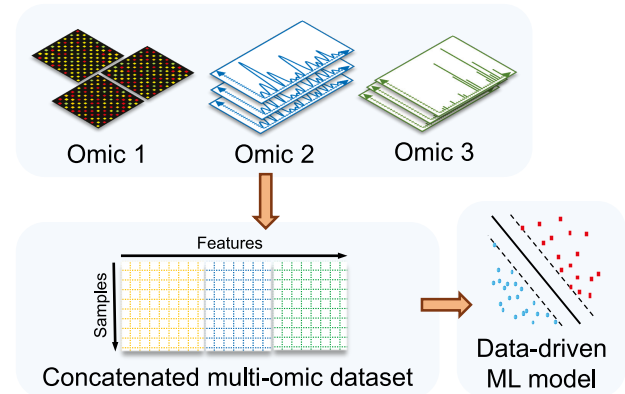
70. Oyetunde T, Liu D, Martin HG, Tang YJ. Machine learning framework for assessment of microbial factory performance. *PloS one*. 2019;14(1):e0210558.
71. Folch-Fortuny A, Teusink B, Hoefsloot HC, Smilde AK, Ferrer A. Dynamic elementary mode modelling of non-steady state flux data. *BMC systems biology*. 2018;12(1):71.
72. DiMucci D, Kon M, Segre D. Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. *mSystems*. 2018;3(5):e00181–18.
73. Chien J, Larsen P. Predicting the Plant Root-Associated Ecological Niche of 21 *Pseudomonas* Species Using Machine Learning and Metabolic Modeling. *arXiv preprint*. 2017; p. 1701.03220.
74. Segre D, DeLuna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. *Nature genetics*. 2005;37(1):77.
75. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature biotechnology*. 2017;35(1):81.
76. Barrett CL, Herrgard MJ, Palsson B. Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC systems biology*. 2009;3(1):30.
77. Folch-Fortuny A, Marques R, Isidro IA, Oliveira R, Ferrer A. Principal elementary mode analysis (PEMA). *Molecular BioSystems*. 2016;12(3):737–746.
78. Bhadra S, Blomberg P, Castillo S, Rousu J, Wren J. Principal metabolic flux mode analysis. *Bioinformatics*. 2018;1:9.
79. Plaimas K, Mallm JP, Oswald M, Svara F, Sourjik V, Eils R, et al. Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC systems biology*. 2008;2(1):67.
80. Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature genetics*. 2011;43(7):656.
81. 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. *Molecular BioSystems*. 2017;13(8):1584–1596.
82. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research*. 2004;14(2):301–312.
83. Li L, Zhou X, Ching WK, Wang P. Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines. *BMC bioinformatics*. 2010;11(1):501.
84. Yaneske E, Angione C. The poly-omics of ageing through individual-based metabolic modelling. *BMC bioinformatics*. 2018;19(14):415.
85. Occhipinti A, Eyassu F, Rahman TJ, Rahman PK, Angione C. In silico engineering of *Pseudomonas* metabolism reveals new biomarkers for increased biosurfactant production. *PeerJ*. 2018;6:e6046.

86. Zampieri G, Coggins M, Valle G, Angione C. A poly-omics machine-learning method to predict metabolite production in CHO cells. In: *Metabolomics, The 2nd International Electronic Conference on*. MDPI AG; 2017. p. 4993.
87. Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nature communications*. 2016;7:13090.
88. Samal SS, Radulescu O, Weber A, Fröhlich H. Linking metabolic network features to phenotypes using sparse group lasso. *Bioinformatics*. 2017;33(21):3445–3453.
89. 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. *Metabolic engineering*. 2016;33:158–168.
90. Guo W, Xu Y, Feng X. DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing. *arXiv preprint*. 2017; p. 1705.03094.
91. Angione C, Pratanwanich N, Lió P. A hybrid of metabolic flux analysis and bayesian factor modeling for multiomic temporal pathway activation. *ACS synthetic biology*. 2015;4(8):880–889.
92. Barsacchi M, Andres-Terre H, Lió P. GEESE: Metabolically driven latent space learning for gene expression data. *bioRxiv*. 2018; p. 365643.
93. Wu SG, Wang Y, Jiang W, Oyetunde T, Yao R, Zhang X, et al. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. *PLoS computational biology*. 2016;12(4):e1004838.
94. Brunk E, George KW, Alonso-Gutierrez J, Thompson M, Baidoo E, Wang G, et al. Characterizing strain variation in engineered *E. coli* using a multi-omics-based workflow. *Cell systems*. 2016;2(5):335–346.
95. Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson OE, Palsson BO. Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. *Nature Communications*. 2017;7:46249.
96. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*. 2018;9(1):5252.
97. Robinson JL, Nielsen J. Integrative analysis of human omics data using biomolecular networks. *Molecular BioSystems*. 2016;12(10):2953–2964.
98. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global-omics data. *Genome biology*. 2015;16(1):186.
99. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, et al. In silico method for modelling metabolism and gene product expression at genome scale. *Nature communications*. 2012;3:929.
100. Motamedian E, Mohammadi M, Shojaosadati SA, Heydari M. TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinformatics*. 2016;33(7):1057–1063.

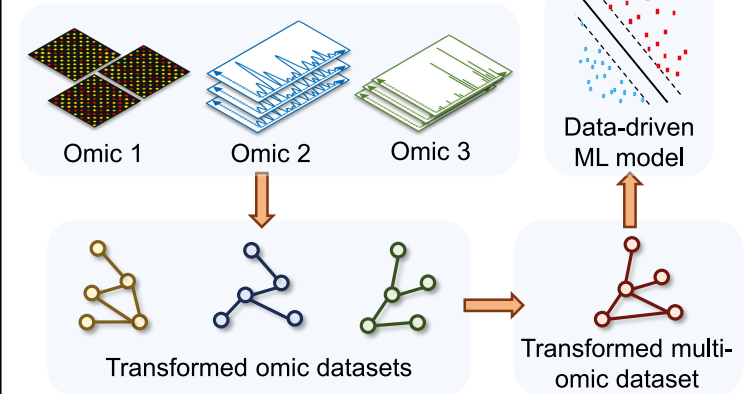
101. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*. 2010;5(1):93.
102. Semmes OJ. The “omics” haystack: defining sources of sample bias in expression profiling. *Clinical Chemistry*. 2005;51(9):1571–1572.
103. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*. 2017;35(6):498–507.
104. Tummler K, Klipp E. The discrepancy between data for and expectations on metabolic models: How to match experiments and computational efforts to arrive at quantitative predictions? *Current Opinion in Systems Biology*. 2017;8:1–6.
105. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature biotechnology*. 2001;19(2):125.
106. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, et al. Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *Journal of bacteriology*. 2006;188(23):8259–8271.
107. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*. 2010;6(1):390.
108. King ZA, O’Brien EJ, Feist AM, Palsson BO. Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion. *Metabolic engineering*. 2017;39:220–227.
109. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*. 2014;10(4):e1003580.
110. Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature biotechnology*. 2008;26(6):659.
111. Lay Jr JO, Liyanage R, Borgmann S, Wilkins CL. Problems with the “omics”. *TrAC Trends in Analytical Chemistry*. 2006;25(11):1046–1056.
112. Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND. Molecular signatures from omics data: from chaos to consensus. *Biotechnology journal*. 2012;7(8):946–957.
113. Long MR, Ong WK, Reed JL. Computational methods in metabolic engineering for strain design. *Current opinion in biotechnology*. 2015;34:135–141.
114. Campbell K, Xia J, Nielsen J. The Impact of Systems Biology on Bioprocessing. *Trends in Biotechnology*. 2017;35(12):1156–1168.
115. Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC systems biology*. 2011;5(1):180.
116. Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, et al. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS computational biology*. 2019;15(2):e1006733.

117. Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A. Disaggregating asthma: Big investigation versus big data. *Journal of Allergy and Clinical Immunology*. 2017;139(2):400–407.
118. Lipton ZC. The Mythos of Model Interpretability. *Queue*. 2018;16(3):30.
119. Bennett KP, Parrado-Hernández E. The interplay of optimization and machine learning research. *Journal of Machine Learning Research*. 2006;7(Jul):1265–1281.

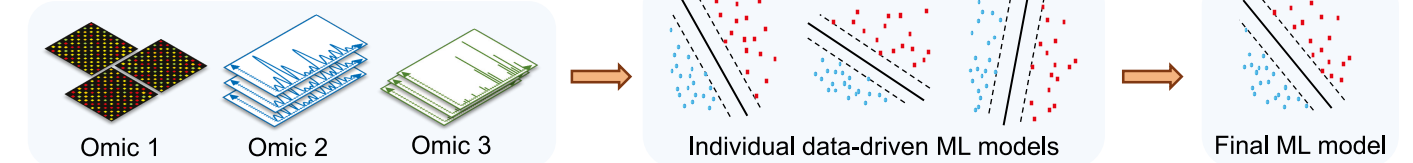
(a) Concatenation-based integration



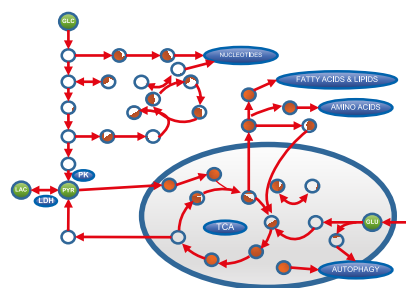
(b) Transformation-based integration



(c) Model-based integration



(a) Genome-scale metabolic reconstruction



(b) Mass balance and metabolic steady state

$$Sv = \frac{dx}{dt} = 0$$

Reactions →

	1	0	0	0	
Metabolites ↓	-1	0	-1	0	v_1
	0	-1	0	1	v_2
	0	1	1	0	\dots
					v_n
	S				v

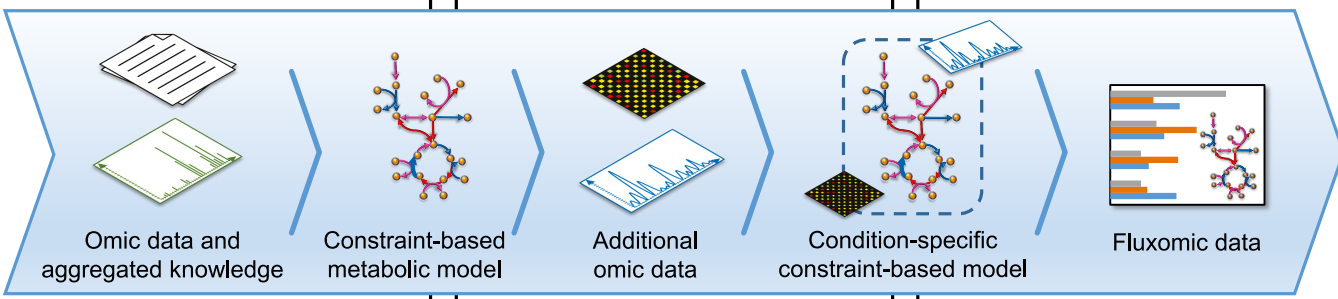
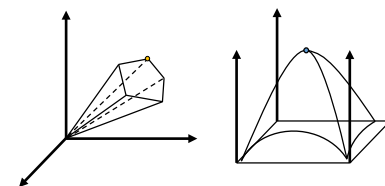
$\cdot = 0$

(c) Regularised linear or quadratic programming

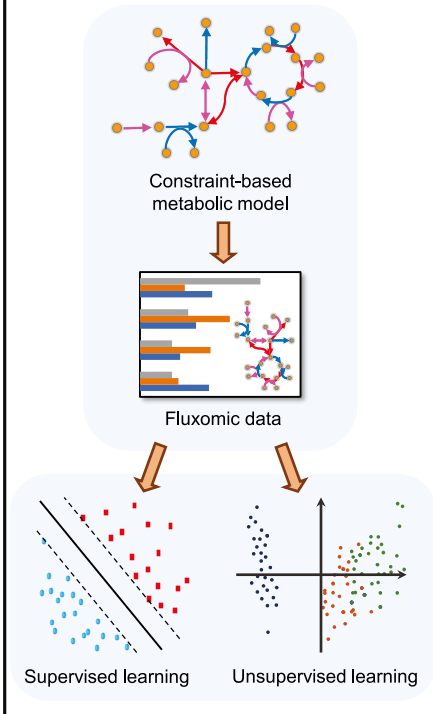
$$\max_v c^T v - \frac{\sigma}{2} v^T v$$

such that $Sv = 0$,

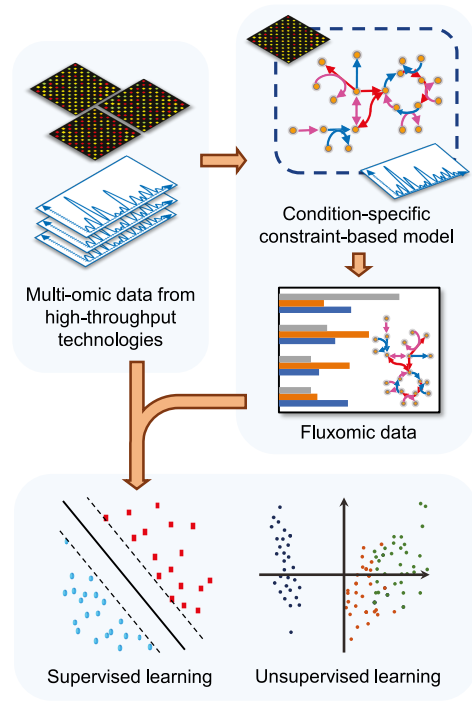
$$v_{\min} \varphi(\theta) \leq v \leq v_{\max} \varphi(\theta)$$



(a) Fluxomic analysis



(b) Multi-omic analysis



(c) Generation of constraint-based models and fluxomic data

