1

From Genome to Actionable Insights in Biotechnology

James Morrissey, Benjamin Strain, and Cleo Kontoravdi

Imperial College London, Department of Chemical Engineering, London SW7 2AZ, UK

1.1 Introduction

Systems biology facilitates the understanding of biological processes [1, 2]. It provides a framework for contextualizing and understanding high-throughput data, allowing structured and meaningful insights to be gained. With the increasing generation of high-throughput omics data and usage of machine learning (ML) data analysis tools, it becomes ever more important to develop robust systems biology tools.

One of the key tools in a systems biologist's repertoire is a network. Networks in biology represent the flow and interactions between components such as genes, proteins, reactions, and metabolites, where network nodes correspond to these biological entities and edges define their interactions, such as biochemical conversion, regulation, and physical binding [3]. Networks improve interpretation and understanding of complex biological phenomena, but, most importantly, in the context of big data, they provide structure. This structure can help turn highly complex data from a "black box" to multiscale models with predictive and interpretable capabilities. Networks allow underlying biological mechanisms to be understood during big data approaches. This structural framework is particularly valuable in biotechnology, where datasets like genomics, transcriptomics, proteomics, and metabolomics must be carefully integrated and analyzed. Even if they are "smaller" in scale, both in network size and connectivity, than those in fields like image recognition or natural language processing, their biological complexity requires equal, if not greater, interpretative attention [4, 5].

By leveraging the structure provided by biological networks, researchers can generate predictions that help elucidate how cellular systems function under various conditions. These predictions, such as flux distributions, regulatory responses, or growth outcomes, offer mechanistic insight into the underlying biology [6]. With this understanding, targeted interventions can be designed, such as modifying gene expression, adjusting nutrient feeds, or engineering metabolic

Machine Learning and Big Data-enabled Biotechnology, First Edition. Edited by Hal S. Alper.

pathways. Crucially, these interventions often lead to positive outcomes, including improved productivity, robustness, or efficiency in biotechnological applications. Thus, networks serve not only as interpretative tools but also as platforms for driving practical, data-informed decisions.

The most widely utilized type of biological network is the metabolic network [7]. These networks describe the flow (or flux) of metabolites through biochemical reactions within a cellular system. Depending on the application, metabolic networks can range from small, pathway-specific subsystems to genome-scale reconstructions that aim to represent the entirety of an organism's metabolic capabilities [8]. These large-scale reconstructions are referred to as genome-scale metabolic models (GEMs). GEMs provide a computational framework for probing metabolism by predicting intracellular fluxes under defined conditions. By integrating experimental data into a metabolic network, GEMs enable simulation of cellular behavior, guiding hypothesis generation, strain engineering, and process optimization in biotechnology [9, 10]. GEMs also serve as a central scaffold for incorporating other biological networks, enabling the integration of multiomics data and supporting comprehensive, data-driven approaches [11, 12].

In this chapter, we explore how biological networks, particularly metabolic networks, can be constructed from genomic data, validated and refined, transformed into predictive models, and ultimately used to generate actionable insights in biotechnology. We highlight how high-throughput omics and ML tools can be used to enhance interpretability, constrain solution spaces, and improve predictive power to enable biotechnology applications.

1.2 From Genome to Network

Genomic data provides the foundational blueprint for all cellular processes. However, to extract actionable insight, this static information must be translated into dynamic representations of biological function. Networks offer a powerful way to achieve this, linking genes to their roles in metabolism, regulation, signaling, and molecular interactions [7].

From a genome sequence, various biological networks can be reconstructed, such as metabolic networks describing biochemical reactions, gene regulatory networks (GRNs) that capture transcriptional control, signaling networks that map cellular communication, and protein–protein interaction (PPI) networks that reveal the physical interactions within the proteome [3]. Each of these networks offers a different layer of understanding, and together they form a comprehensive systems-level view of the cell.

This section outlines how these networks are constructed from genomic data, with a focus on metabolic networks, either using bottom-up approaches that start from gene annotations and known biological functions or using top-down approaches that integrate omics data to refine or contextualize existing network structures. These reconstructions are the foundation for building models that can interpret data, generate predictions, and guide interventions in biotechnology.

Chapter 2 provides detailed information on how network models can be constructed from genomic information using computational algorithms and manual approaches.

1.2.1 Metabolic Networks

Among the various biological networks that can be reconstructed from genomic data, metabolic networks are the most widely applied. Metabolic networks represent the biochemical reactions linking metabolites, genes, enzymes, and reactions in the cell system [13]. This can either be a subset of biochemical reactions (e.g. just focusing on core functions) or can be genome-scale. In recent years, GEMs have been created for a multitude of organisms [14] due to the availability of high-throughput omics data and increased applications of these models in systems biology.

1.2.1.1 Bottom-Up Approaches for Network Reconstruction

The bottom-up approach is a lengthy and manual approach to GEM creation, but it is recommended to create high-quality GEMs from scratch. The reader is pointed to a protocol [8], which contains an in-depth protocol for GEM reconstruction. In this section, we summarize the key steps as illustrated in Figure 1.1. A GEM reconstruction begins with a genome annotation for the organism of interest. To create a draft metabolic network, metabolic reactions can be extracted from the genome annotation using gene ontology (GO) [16], enzyme commission (EC) numbers [17], and biochemical reaction databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] and BRaunschweig ENzyme DAtabase (BRENDA) [19]. Chapters 4 and 5 discuss how protein function can be extracted using these methods as well as other ML frameworks.

This draft annotation is then subject to manual curation to scrutinize every gene and reaction entry. The entries must be relevant to the organism of interest, for example, ensuring correct cofactor and substrate specificity for each enzyme, as well as the correct gene localization. It is preferable to have literature or experimental data to support the presence and function of genes and reactions. When data is lacking, phylogenetically close organisms can be used. During manual curation, a confidence score is useful for assessing the amount of information available for each entry. Gene-protein-reaction (GPR) associations indicate which genes are required for reaction to occur, which states the presence of isozymes and enzyme complexes. The GPRs must be manually refined using databases and literature searches.

The next step is to ensure correct reaction stoichiometry. Metabolites in databases (and hence the draft reconstruction) are represented with uncharged formulas, but their protonation state varies depending on the pH of the cellular environment. The charged formula is determined based on the pKa values of functional groups, which can be predicted using computational tools or literature for the correct pH in the subcellular location. Once the charged formulas are assigned, the correct reaction stoichiometry is established by ensuring mass and charge balance across reactions, incorporating protons and water where necessary. Correct balancing is crucial to avoid artificial energy generation. Correct reaction

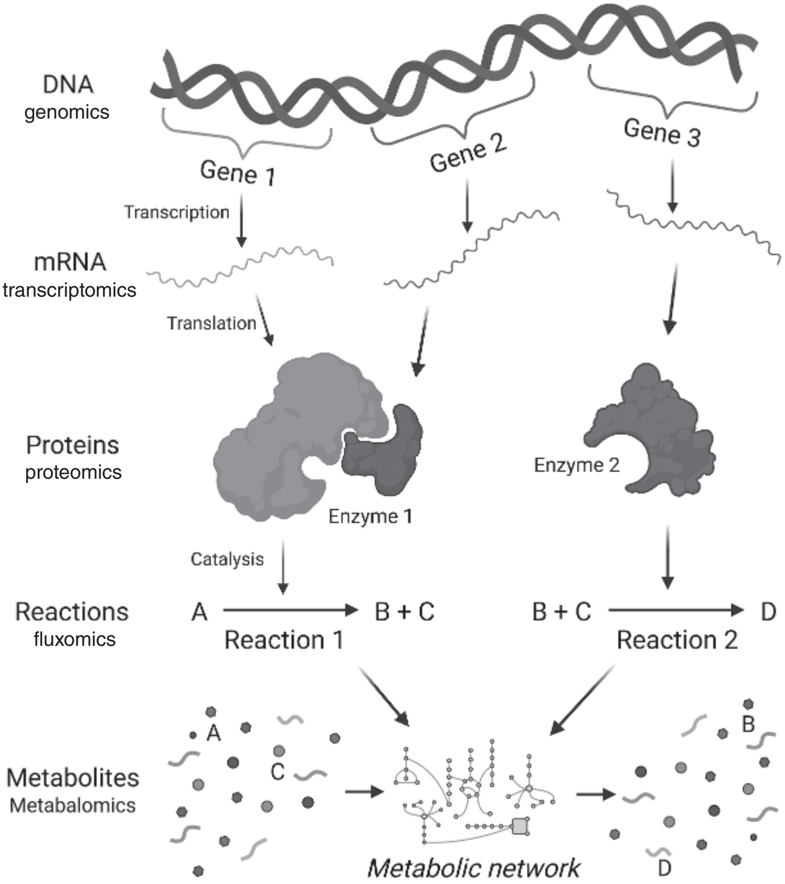

Figure 1.1 Key omics data and steps in the bottom-up construction of a metabolic network from an organism's genome. *Source*: Strain et al. [15] / Elsevier / CC BY 4.0.

directionality is also essential for preventing irreversible reactions from running backward, which would lead to incorrect predictions and thermodynamically infeasible loops (futile cycles), which are discussed in Section 1.3.2. Correct directionality is determined using existing biochemical data, but when this is unavailable, Gibbs free energy change can be obtained from databases or calculated using group contribution methods [20]. To complete and validate the network, further reactions must be added, which are discussed in Section 1.3.1.

1.2.1.2 Top-Down Approaches for Network Reconstruction

In contrast to building draft networks from scratch, top-down approaches to GEM reconstruction rely on pre-existing networks, knowledge, and omics data to infer metabolic networks. The top-down approach to GEM reconstruction is typically faster and more automated than the bottom-up method and is well-suited for generating draft models, particularly when high-throughput data is available. These approaches begin with existing genome annotations and rely on automated pipelines that map annotated genes to known reactions in curated metabolic databases such as KEGG, MetaCyc [21], or ModelSEED [22]. Automated draft reconstructions can be rapidly generated using tools such as CarveMe [23] or Pathway Tools [24].

Top-down approaches can also be used to build context-specific models by integrating omics data into existing generic GEMs [25, 26]. Algorithms like GIMME [27], iMAT [28], and CORDA [29] filter and adjust reaction content based on expression thresholds or activity likelihood, producing models tailored to particular tissues, conditions, or phenotypes.

1.2.2 Networks Beyond Metabolism

GEM reconstructions focus on metabolic functions, but systems biology encompasses a broader range of network reconstructions that capture other aspects of cellular function. These include gene regulation, signaling, protein interaction, and transcription–translation networks, each offering complementary insights into cellular behavior [3].

These network types can be integrated into multilayer models, as discussed in Section 1.5.4, to improve predictive accuracy and capture cellular complexity. While GEMs are well-established, the ongoing development of GRNs, signaling, and transcription-translation models is crucial for building more comprehensive whole-cell models.

GRNs describe the interactions between transcription factors (TFs), regulatory elements, and their target genes [30, 31]. They are often inferred from gene expression data, using statistical methods (e.g. mutual information, Bayesian inference) or curated from literature. The matrix formalism has been used to describe the functional states of such systems, providing a structured way to analyze regulatory dynamics. A key example is the genome-scale reconstruction of the *Escherichia coli* transcriptional regulatory system, which enables simulation of transcriptional responses to environmental changes [32].

Signaling networks map how extracellular signals lead to cellular responses through cascades of molecular interactions. These can be constructed from curated pathway databases or derived from proteomics data but are less commonplace than other networks [33]. PPI networks represent the physical interactions between proteins within a cell [34]. Unlike metabolic or regulatory networks, these are typically undirected and do not reflect stoichiometry or directionality but instead capture the structural and functional connectivity of the proteome. PPIs are essential for nearly all cellular processes, including enzyme complexes, signaling cascades, structural assemblies, and transport mechanisms.

To model the impact of gene expression beyond regulation, genome-scale reconstructions of transcriptional and translational machinery have been developed. For instance, multiscale models of metabolism and expression (ME models) integrate metabolism with macromolecular expression systems, accounting for RNA polymerase, ribosomes, and enzyme synthesis [35, 36]. These ME models provide a platform to simulate how cellular resources are allocated between growth and gene expression and are discussed further in Section 1.5.4.

1.3 From Draft to Functional Network

In order to turn the draft network into a functional model, several refinements have to occur. These include adding additional reactions to the network, which are not represented by particular gene(s), such as sink reactions, spontaneous reactions, and the biomass reaction, but are required to create a metabolic model. These reactions are key; they provide a termination point of reaction pathways to prevent infinite accumulation of reaction flux, as well as provide routes for experimental data to be integrated into GEMs.

In addition, some basic network validation functionality checks must be implemented to verify the integrity of the network. These are elaborated in the following section.

1.3.1 Additional Reactions

1.3.1.1 Exchange Reactions

Exchange reactions represent the transfer of metabolites between the cell and its extracellular environment. The exchange reactions include sugars, amino acids, water, ammonia, sulfur, and many other potential compounds known to be taken up or secreted by the cell. The exchange reaction of metabolite A is modeled in the form:

$$EX_A\ :\ A_e \leftrightarrow$$

The format of the exchange reactions is such that a negative flux represents uptake from the extracellular environment and a positive flux represents secretion. Metabolites that are secretion-only are modeled as irreversible reactions to prevent a negative flux.

1.3.1.2 Demand Reactions

Demand reactions are additional sinks added to model terminal metabolites without secretion. This can represent metabolites required by the cell for normal function, desired products, e.g. antibodies or generic demand reactions to replace specific biosynthetic pathways, which may not be understood. The demand reaction of metabolite A in the cytosol is modeled in the form:

$$DM_A\ :\ A_c \rightarrow$$

1.3.1.3 Transport Reactions

GEMs also contain location information for metabolites, allowing the inclusion of transport reactions between compartments. For example, A_c refers to metabolite A in the cytosol (c) and A_e refers to metabolite A in the extracellular space (e). Other locations modeled include [m] for mitochondria, [x] for peroxisome, [r] for endoplasmic reticulum, and [g] for the Golgi apparatus. Thus, transport reactions are modeled in the form:

$$A_e \leftrightarrow A_c$$

1.3.1.4 Spontaneous Reactions

Spontaneous reactions are those that occur without enzymatic catalysis and therefore are not associated with specific genes; hence, these are not added to the network during the initial reconstruction phase. They ensure network completeness by allowing certain transformations to proceed even when no known enzyme is responsible. Spontaneous reactions are typically irreversible and should be added when they have at least one metabolite connecting them to the rest of the reconstruction, to avoid too many dead-end metabolites caused by spontaneous reactions. Examples include the spontaneous degradation of dihydroxyacetone phosphate (DHAP) and glyceraldehyde-3-phosphate (G3P) to form methylglyoxal [37] as well as metal-catalyzed reactions such as Fenton chemistry (H_2O_2 and Fe^{2+} reacting to generate hydroxyl radicals) [38].

1.3.1.5 Nongrowth Associated ATP Maintenance

Nongrowth associated ATP maintenance (NGAM) reactions account for the energy required by the cell to sustain basic functions unrelated to biomass production, such as ion transport, turnover of macromolecules, and maintenance of membrane potential. A fixed lower bound is set on this reaction's flux to reflect the minimum ATP demand necessary for cell viability, based on experimental estimates, for example, 7.6 and 5.9 mmol g^{-1} hr^{-1} for *E. coli* and Chinese hamster ovary (CHO) cells, respectively [39, 40]. This reaction is typically modeled as follows:

$$\mathrm{H_2O + ATP \rightarrow ADP + H^+ + P}$$

Note that growth-associated maintenance (GAM), which accounts for the energy necessary to replicate a cell, e.g. for macromolecular synthesis (e.g. proteins, DNA, and RNA), is included in the biomass reaction outlined in the next section.

1.3.1.6 Biomass Reaction

A fundamental component of GEMs is the addition of an *ad hoc* biomass reaction [41]. The overall cell composition (amino acids, lipids, carbohydrates) is measured, and its constituent components are used as an input to this biomass reaction in the form. The flux through this reaction is used as an approximation of cell growth rate; as such, the molar composition of the biomass inputs sums up to one gram of dry cell weight (gDCW). The resulting units of this reaction are then h, allowing comparison with experimentally measured cell growth rates.

$$\sum_{i \in I^{\mathrm{biomass}}} S_{i,\mathrm{biomass}} X_i + \mathrm{ATP} + \mathrm{H_2O} \rightarrow \mathrm{biomass} + \mathrm{ADP} + \mathrm{H^+} + \mathrm{P}$$

where X_i is a biomass precursor (e.g. amino acid, lipid, carbohydrate) and $S_{i,\mathrm{biomass}}$ is the stoichiometric coefficient for precursor *i*. The biomass reaction also requires an amount of ATP and produces ADP, hydrogen ions, and phosphate.

1.3.2 Network Validation

Validation of the model structure is an essential step to ensure that the GEM accurately represents the biological system it is intended to capture. Structural validation aims to detect errors such as missing reactions, incorrect associations, or stoichiometric imbalances, which could lead to inaccurate predictions.

1.3.2.1 Manual Screening

Manual screening involves meticulously curating each reaction in the model to verify its accuracy, ensuring that reaction stoichiometry and the GPR associations are correct and reflect the known biology of the organism. This step is especially important because computational algorithms used to construct GEMs may introduce errors or make assumptions that lead to inaccurate or incomplete GPR associations. For example, the manually curated genome-scale reconstruction of the metabolic network of *Bacillus megaterium*, a microorganism widely used in industrial biotechnology, is a prime case where manual curation applied after automated reconstruction ensured accurate reflection of the organism's metabolism and high predictive accuracy [42]. Correcting errors at this stage prevents propagation of inaccuracies in subsequent analyses, ultimately improving the predictive power and utility of the model.

1.3.2.2 Screening for Dead-End Reactions and Blocked Metabolites

A critical step in GEM validation is the identification and resolution of dead-end reactions and blocked metabolites. Dead-end reactions are those that either lack substrates or fail to produce products that can be utilized by subsequent reactions. Blocked metabolites are compounds that cannot be consumed or produced by any reaction in the network. These issues typically arise due to incomplete or inaccurate model construction and can significantly distort model predictions by preventing key metabolic pathways from functioning [43].

The set of blocked reactions in a metabolic network is identified by determining whether each reaction can carry any flux under given conditions. This is done by maximizing the flux for each reaction while maintaining the network's stoichiometry. If the maximum flux for a reaction is zero, it is considered unusable or blocked. Although they may appear functional in a static view of the network, these reactions are inactive in practice due to incomplete or inconsistent pathway integration [44].

Flux variability analysis (FVA) is a powerful technique for uncovering such issues. An extension of flux balance analysis (FBA), FVA examines the range of flux values that reactions can carry while optimizing for a specific cellular objective [45]. FVA is particularly useful for identifying reactions that, despite being part of the network, are blocked or inactive in certain conditions due to gaps in pathway connectivity or stoichiometric errors.

Another commonly used technique is FastGapFill [46], which identifies reactions that must be added to restore the functionality of blocked pathways in a way that minimizes alterations to the existing network. FastGapFill operates by proposing the inclusion of missing reactions based on data from external reaction databases (e.g. KEGG) or from closely related organisms. The algorithm works under the assumption that the number of changes should be minimal to preserve the integrity of the model, only adding reactions that are most likely to occur naturally in the organism.

1.3.2.3 Infinite Loops

Infinite loops refer to metabolic cycles that can carry flux indefinitely without the net consumption or production of metabolites (Figure 1.2). These loops often arise from structural errors, such as incorrect reaction stoichiometries, missing thermodynamic constraints, or improperly connected reactions. These cycles can lead to implausible outcomes. For example, the model might predict a perpetual flux through ATP-generating pathways without any corresponding substrate consumption, violating basic thermodynamic laws [47].

Infinite loops typically occur in linear programming-based methods such as FBA, where reactions are optimized for specific objectives (e.g. biomass production or ATP generation). In the absence of appropriate constraints, the model may inadvertently favor these loops.

Detecting and eliminating infinite loops is essential for ensuring the model behaves realistically. Several techniques are used to identify and address infinite loops in genome-scale models. Much like dead-end reactions, infinite loops can also be detected by analyzing flux variability across different environmental conditions. Using FVA, reactions that carry flux under unlikely or unrealistic scenarios, pointing to the presence of infinite loops, can be identified. Reactions with unusually wide ranges of flux under varying conditions are often indicative of these erroneous cycles.

Adding thermodynamic constraints is one of the most effective ways to prevent infinite loops. Thermodynamically constrained FBA (tFBA) [48] imposes directionality constraints on irreversible reactions and ensures that reactions with unfavorable Gibbs free energy do not carry flux. Loopless-COBRA (COnstraint-Based Reconstruction and Analysis) is an extension of traditional FBA that specifically aims to eliminate flux through futile cycles or loops [49]. It introduces additional constraints that prevent reactions from forming internal cycles that can carry flux without contributing to the objective function (e.g. biomass production). Finally, manual identification of highly active reactions and pathways is another approach to detect potential infinite loops that fail to consume or produce net metabolites. Additionally, it can help identify dead-end reactions, which are often linked to infinite loops.

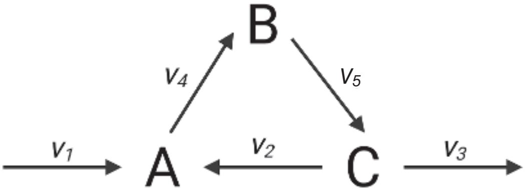

Figure 1.2 A depiction of an infinite cycle. If uptake of metabolite *A*, v_1, is constrained to an experimental value, the demand flux of metabolite *C*, v_3, is limited to this uptake i.e. $v_1 = v_3$. *However,* fluxes v_2, v_4 and v_5 can take any value and still produce a feasible solution. This becomes an issue when these predicted fluxes are important as part of metabolic analysis, for example if the user wants to find a value for conversion of *A* to *B*, v_4.

1.3.2.4 Leaks and Siphons

Leaks and siphons can lead to biologically inaccurate predictions, particularly in the context of energy and metabolite balances [50]. Leaks occur when a model predicts the loss of metabolites or energy carriers, such as ATP or nicotinamide adenine dinucleotide hydride (NADH), without accounting for their regeneration or the presence of corresponding inputs. This leads to the unrealistic depletion of key molecules, disrupting the energy balance and resulting in misleading simulations, such as predicting cell death or metabolic failure when, biologically, the organism could still thrive.

Siphons are pathways or sets of reactions in the model where certain metabolites accumulate or are depleted without affecting the overall balance of the system. These siphons represent “traps” where flux is directed through reactions, but the corresponding consumption and production of metabolites are not balanced, leading to unbounded fluxes or unrealistic metabolite accumulation. Siphons are often problematic in energy metabolism, where they can allow energy carriers (e.g. ATP, NADH) to be produced or consumed without appropriate regulation, distorting the model’s predictions of cellular growth or metabolic output.

One approach to identifying leaks is performing mass and charge balance checks across all reactions. Metabolites in a properly balanced model should not disappear or be generated without corresponding inputs and outputs. This method helps detect reactions that may lack stoichiometrically correct inputs (substrates) or outputs (products), which can lead to leaks where energy or metabolites are lost or accumulated.

A simple but effective approach to testing for leaks involves the addition of synthetic exchange reactions to the model. These synthetic reactions are designed to “leak” specific metabolites or energy carriers from the system, mimicking the effect of a metabolic leak. By adding and constraining these reactions, researchers can test if energy or key metabolites are leaving the system inappropriately. If the model shows flux through these synthetic reactions, it indicates the presence of a leak, requiring further investigation into the reactions involved.

The fast leak test function within the COBRA toolbox [51] systematically checks if a metabolic model can produce metabolites from nothing when all exchange reactions are closed, meaning no uptake is allowed. It identifies any leaking metabolites by maximizing the flux through each exchange reaction. If secretion flux is detected, it signals a leak. Optionally, the function can also test for leaks through demand reactions. It returns a list of leaking metabolites, the closed model, and the flux vector for exchange reactions, helping pinpoint and resolve problematic reactions.

1.4 From Functional Network to Model

Once a functional network has been constructed and validated, it can be transformed into a computational model capable of simulating cellular behavior. This requires formulating the network as a mathematical problem, typically using *constraint-based* methods such as FBA. These approaches enable the prediction of metabolic flux distributions under defined conditions and objectives, providing a powerful tool for integrating high-throughput datasets to understand cell behavior.

1.4.1 Flux Balance Analysis

FBA is the most widely used method to analyze metabolic networks [52, 53]. FBA calculates the flow of metabolites through the system in a linear optimization program, thereby making it possible to make metabolic flux predictions, including growth rate, recombinant protein production, accumulation of toxic metabolites, media uptake/secretion, effects of genetic engineering, influence of enzymes on phenotype, and much more. The principle of FBA is a material balance on all metabolites within the GEM. The (molar) material balance on metabolite A would be:

$$\frac{\mathrm{d}[A]}{\mathrm{dt}} = \sum_{j \in J} S_{A,j} v_j$$

where $[A]$ is the concentration of metabolite A in units of mmol per gram dry cell weight (gDCW), t is the time (hr), $S_{A,j}$ is the stoichiometric coefficient of A in reaction j, from the stoichiometric matrix, and v_j is the flux through reaction j in units of mmol gDCW^{-1} hr^{-1}.

In kinetic models of metabolism, reaction flux v_j is replaced with kinetic expressions, leading to a system of nonlinear ordinary differential equations (ODEs), with metabolite concentrations as the free variables. Instead, FBA makes a pseudo-steady-state assumption, meaning the concentration of metabolites is constant, hence a steady state with no accumulation. This assumption is appropriate in many conditions, as metabolic flux far exceeds accumulation [54], permitting cell culture to be divided into periods of pseudo-steady state. This leads to the elimination of the left-hand side and v_j becomes the free variable, leading to the general equation:

$$\frac{\mathrm{d}[C_i]}{\mathrm{dt}} = \sum_{j \in J} S_{ij} v_j = 0 \text{ for } i \in I$$

where C_i is the intracellular concentration of metabolite i and S_{ij} is the stoichiometric coefficient of metabolite i in reaction j. With this formulation, the number of equations is equal to the number of metabolites, i, and the number of variables is equal to the number of reactions, j. Since most metabolites participate in many reactions, the number of variables far exceeds the number of equations. This means there are multiple degrees of freedom and the system is underdetermined. In general, an underdetermined system of linear equations has an infinite number of solutions. However, in optimization problems that are subject to linear equality constraints, only one of these solutions is relevant, namely the one giving the highest or lowest value of an objective function.

In its most general form, FBA is a constrained linear optimization problem, with an objective function to maximize (or minimize) flux through a particular reaction. The most frequently used reaction in the objective function is the biomass reaction. The assumption behind a biomass objective function is that the cell is allocating all resources toward growth; hence, the cell will prioritize placing fluxes in reactions that will lead to the highest biomass production.

The reaction fluxes are also subject to upper and lower bounds, v_j^{LB} and v_j^{UB}. The selection of these bounds is a key factor in making accurate predictions using FBA. This generalized form of FBA is shown below: e

$$\text{maximise } v_{\text{objective}}$$

$$\sum_{j \in J} S_{i,j} v_j = 0 \text{ for } i \in I$$

$$v_j^{\text{LB}} \leq v_j \leq v_j^{\text{UB}} \text{ for } j \in J$$

1.4.2 Flux Variability Analysis

Solutions to the (FVA) problem are not unique since multiple flux distributions can yield the unique optimal objective value. As a variation on FBA, FVA captures the feasible range of reaction fluxes that satisfy an optimal objective value [45, 55]. The method calculates the minimum and maximum allowable fluxes through each reaction using the linear program:

$$\text{maximise (and minimise)}\, v_j$$

$$\sum_{j \in J} S_{i,j} v_j = 0 \text{ for } i \in I$$

$$v_j^{\text{LB}} \leq v_j \leq v_j^{\text{UB}} \text{ for } j \in J$$

$$v_{\text{objective}} = v_{\text{objective}}^{\text{optimal}}$$

For a network with N reactions, $2N$ optimization problems would be solved. The feasible region of all fluxes is described by an N-dimensional polytope. FVA is a convenient way to inscribe the smallest box around this feasible region. FVA cannot find the N-dimensional polytope itself since changing the flux of a reaction for the same objective function typically requires that the remaining fluxes in the network change as well. Figure 1.3 shows the feasible region and FVA solution space.

FVA can also be used to find blocked reactions in the metabolic network for a given environmental condition, as discussed in Section 1.3.2. These are reactions for which the minimum and maximum values identified by FVA are both zero. Blocked reactions can be caused by reaction constraints and also due to gaps in the metabolic network, for example, metabolites for which there is no production or consumption path in the network (dead-end metabolites). Identifying these blocked reactions is useful for spotting dead-end metabolites and also reducing the FBA search space by pre-setting the value of these reactions to zero.

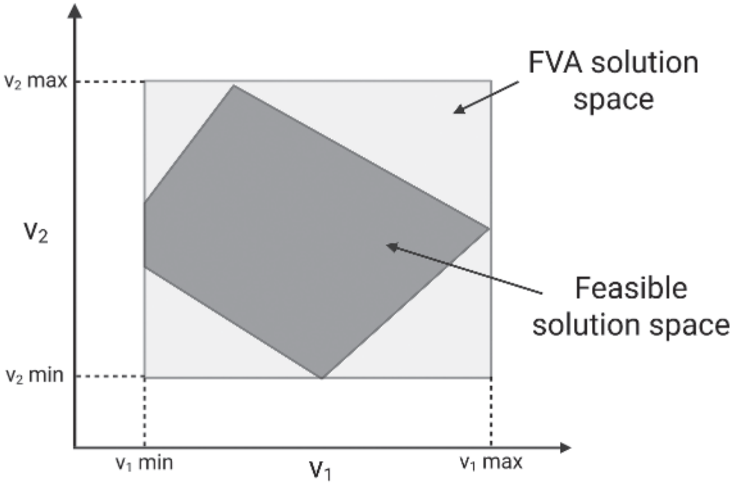

Figure 1.3 FVA solution space.

1.4.3 Flux Sampling

While FBA provides a flux distribution that maximizes (or minimizes) a defined objective, and FVA gives the range of feasible fluxes for each reaction, both are limited in that they do not characterize the full solution space. In reality, the space of possible flux distributions satisfying the steady state mass balance and other biological constraints is vast and underdetermined. Flux sampling methods aim to explore this space by generating many feasible solutions, providing a more comprehensive and probabilistic view of the model's behavior [56–58].

Flux sampling algorithms such as Hit-and-Run [59], Artificial Centering Hit-and-Run [60], and Coordinate Hit-and-Run with Rounding [61] operate by iteratively drawing flux vectors that satisfy the network constraints to fall within the defined flux bounds. This produces a representative sample of the solution space, rather than a single point estimate, enabling statistical analysis of fluxes across different pathways and conditions.

The key advantage of flux sampling is its ability to reveal the range and distribution of possible flux values without being tied to a specific optimization objective [56] as shown in Figure 1.4. This is particularly useful in situations where the biological objective is unclear or where the system does not necessarily prioritize growth, such as in stress conditions, nondividing cells, or engineered strains with altered metabolism. Sampling can highlight which pathways are consistently active, identify flux correlations or trade-offs between competing pathways, and uncover alternative metabolic strategies that achieve similar phenotypes.

Flux sampling also allows integration with ML approaches by generating high-dimensional datasets of flux states that can be clustered, classified, or

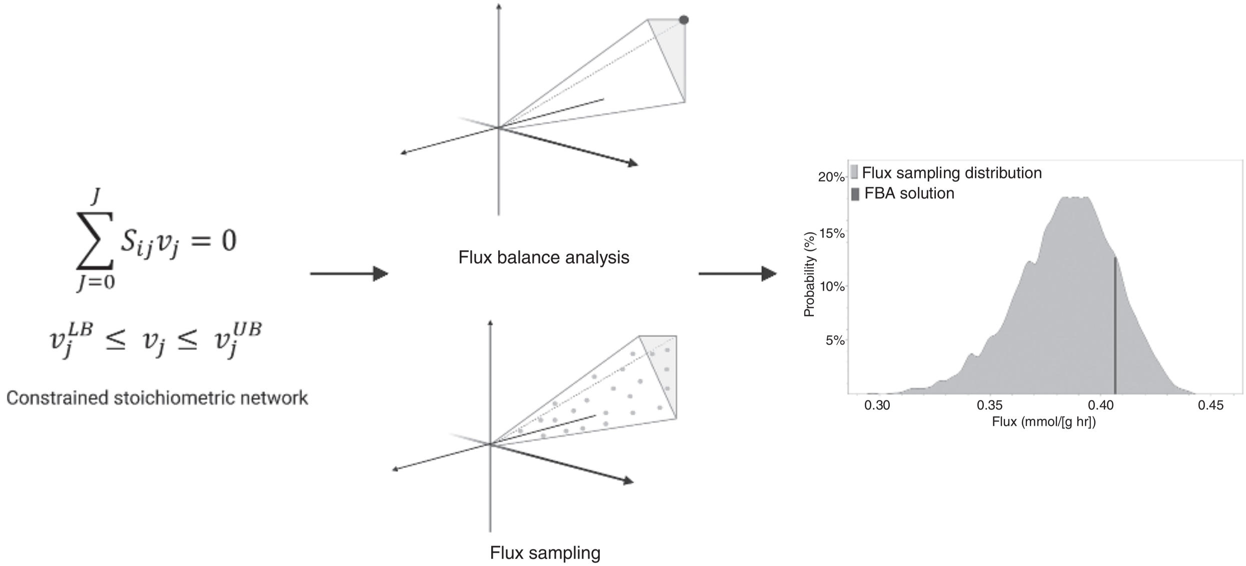
$\sum_{J=0}^{J} S_{ij} v_j = 0$
$v_j^{LB} \leq v_j \leq v_j^{UB}$
Constrained stoichiometric network
Flux balance analysis
Flux sampling
Flux sampling distribution
FBA solution
Probability (%)
20%
15%
10%
5%
0.30
0.35
0.40
0.45
Flux (mmol/[g hr])

Figure 1.4 Flux sampling vs FBA approaches to finding a flux solution. FBA optimizes for a particular metabolic objective, yielding a single flux distribution, while flux sampling samples the feasible solution space without the assumption of an objective. As a result, flux sampling generates a probability distribution and provides deeper insights into metabolism. This example flux distribution is for GAPDH (glyceraldehyde 3-phosphate dehydrogenase) with $N = 5{,}000$.

regressed against phenotypic traits [62–64]. This can be particularly valuable in systems with complex regulation or when trying to predict emergent properties from large design spaces.

However, flux sampling is computationally intensive, especially for large-scale GEMs with thousands of reactions. Efficient sampling requires advanced linear algebra methods and careful preprocessing to ensure numerical stability and uniform coverage of the solution space. Furthermore, interpreting the resulting high-dimensional data can be challenging and often requires dimensionality reduction or statistical summarization. Unlike FBA, which offers a clear optimization criterion, flux sampling does not indicate which distribution is "best" or "most likely" in a biological sense; it just shows what could happen, given the model constraints.

Unlike FBA, which identifies a single flux distribution by optimizing a defined objective function (e.g. maximizing growth), flux sampling generates a large ensemble of feasible flux states without prioritizing any particular solution. As such, the results are inherently nondeterministic, reflecting the full range of possible metabolic behaviors permitted by the model constraints. Further analysis is required to extract biological insight from these distributions, such as identifying frequently active reactions, flux correlations, or dominant metabolic pathways that may represent physiologically relevant states [65].

1.5 From Model to *In Silico* Predictions

Once a functional and validated GEM is constructed, it can be used as a powerful *in silico* tool to simulate and analyze cellular behavior. This step marks the transition from a basic biochemical network model to a predictive platform capable of generating biological insights. *In silico* predictions can guide metabolic engineering, strain design, feed optimization, and hypothesis generation across various biotechnological applications discussed in Section 1.6. However, the quality and reliability of these predictions are highly dependent on the quality of constraints applied to the model, the selected optimization strategy, and the integration of relevant biological data. This section discusses tools and techniques to integrate omics data, systems knowledge, ML, and other biological networks to generate reliable predictions.

1.5.1 Constraints

A metabolic model represents all the possible metabolic capabilities of an organism under any condition. In order to make accurate, condition-specific predictions, biological constraints must be applied to reduce the feasible solution space. These constraints are applied to the same models defined in Section 1.4 (FBA, flux sampling) but provide more context and refinement to predictions by defining which reactions are likely to be active under the given experimental or physiological context. The more biologically realistic the constraints, the closer the predicted fluxes will be to true cellular behavior.

Various omics datasets can be used to impose these constraints, as summarized in Table 1.1. Metabolomics data, when converted to uptake and secretion rates, directly constrain exchange reactions and are commonly used in all FBA approaches. Transcriptomics allows inference of reaction activity via gene expression levels, with methods like MADE [67], MOOMIN [68], and SPOT [81] converting expression data into condition-specific models through GPR associations. Thermodynamic constraints, using Gibbs free energy change, assign directionality to reactions and prevent infeasible loops; approaches like TFA [48], CycleFreeFlux [78], and loopless-COBRA [49] incorporate these principles.

Enzyme kinetic and physical capacity constraints account for cellular limits on abundance and catalytic capacity of enzymes; methods such as FBAwMC [72], MOMENT [73], and GECKO [76] reflect proteome constraints, crowding, and energetic budgets, thereby moving beyond stoichiometry to better capture cellular priorities.

1.5.2 Objective Function

In optimization-based models (of which FBA is the most popular), the objective function is required to achieve a predict flux distribution, yet the choice of objective function is not always clear. Objective functions are a useful component of metabolic models, as they can guide flux toward realistic and appropriate solutions in an underdetermined solution space. However, a poor choice can lead to flux predictions moving away from *in vivo* reaction rates and reducing model validity.

The most commonly applied objective in metabolic models is biomass maximization, based on the assumption that cells naturally evolve to prioritize growth. However, in some conditions (especially under stress, nutrient limitation, or engineered phenotypes) and will more complex cell types, this assumption may not hold. Omics data can be integrated into metabolic models to help guide objective functions to generate more realistic solutions. Similar to the constraining methods discussed above, these can use transcriptomics, proteomics, to guide fluxes that best fit with the omics data provided.

Furthermore, objectives can be inferred directly from the data rather than assumed or guided by the data. Methods such as ObjFind [82], BOSS [83], invFBA [84], and SCOOTI [85] use experimental fluxes, omics data, or single-cell profiles to mathematically deduce plausible objective functions that are consistent with observed phenotypes. These approaches help overcome the limitations of assuming a fixed objective like biomass, particularly in nonproliferative cells or systems with poorly defined cellular priorities. Multiobjective formulations and pareto-based methods also allow simultaneous consideration of competing biological goals, such as balancing growth with energy efficiency or survival, providing a more nuanced and flexible modeling framework.

Examples of these methods are summarized in Table 1.2.

1.5.3 Validating In Silico Predictions

The utility of a GEM lies in its ability to make accurate predictions of biological behavior under defined conditions. Predictive accuracy can be assessed through

Table 1.1 Examples of methods using omics data and biological knowledge to constrain metabolic models.

Method/name	Description	Original application	Reference
Standard FBA	Uptake reactions are bound by experimental measurements.	Growth prediction in *E. coli*.	Varma and Palsson [53]
E-Flux (Expression-Flux)	Directly binds reactions by gene expression.	Prediction of drug impacts on mycolic acid production in *Mycobacterium tuberculosis*.	Colijn et al. [66]
MADE (Metabolic Adjustment by Differential Expression)	MADE uses the statistical significance of change in expression between conditions to determine if the genes are switched on or off.	Model the metabolic adjustments seen in the transition from fermentative to glycerol-based respiration in *S. cerevisiae*.	Jensen and Papin [67]
MOOMIN (Mathematical explOration of Omics data on a MetabolIc Network)	MOOMIN uses a Bayesian posterior probability of differential gene expression to assess if genes are switched on or off.	Model diauxic shift from glucose to glycerol in *S. cerevisiae*.	Pusa et al. [68]
LBFBA (Linear Bound FBA)	Finds a linear relationship between gene expression and reaction flux. Uses this relationship to bind reactions in new conditions.	Predicting intracellular fluxes in *E. coli* and *S. cerevisiae*.	Tian and Reed [69]
mCADRE (Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation)	Creates context-specific GEMs by ranking and pruning reactions from the network.	Creating context-specific GEMs for 126 human tissue and cell types.	Wang et al. [70]
ccFBA (carbon-constrained FBA)	Constrains each reaction by a mass balance on the total available carbon.	Genome-scale metabolic model of CHO cells.	Lularevic et al. [71]
TFA (Thermodynamic-based Flux Analysis)	Reaction directionality is determined by Gibbs free energy change.	Genome-scale metabolic model of *E. coli* cells.	Henry et al. [48]

(Continued)

Table 1.1 (Continued)

Method/name	Description	Original application	Reference
FBAwMC (FBA with Molecular Crowding)	Reactions are constrained by maximum volume of enzymes inside the cell or compartments.	Prediction of growth rate of wild-type and mutant *E. coli* cells.	Beg et al. [72]
MOMENT (MetabOlic Modeling with ENzyme kineTics)	Reactions are constrained by maximum concentration of enzymes.	Predict growth rates and intracellular reactions of *E. coli*.	Adadi et al. [73]
ecFBA (enzyme capacity FBA)	Reactions are constrained by protein abundance, using enzyme turnover numbers and molecular weights.	Prediction of central carbon metabolism fluxes in CHO cells and lactate metabolism.	Yeo et al. [74]
CAFBA (Constrained Allocation FBA)	Biosynthetic costs associated with growth are accounted for through a single additional genome-wide constraint.	Predictions on the rate of acetate excretion and growth in *E. coli*.	Mori et al. [75]
GECKO (GEM with Enzymatic Constraints using Kinetics and Omics data)	Including enzymes as part of GEM, thus limiting fluxes by protein abundance.	Describing *S. cerevisiae* phenotypes under high enzymatic pressure conditions, such as alternative carbon sources, stress, or overexpressing a specific pathway.	Sánchez et al. [76]
RBA (Resource Balance Analysis)	Accounts for the distribution of proteins across translation apparatus, enzymatic reactions, and production of microcontents as a limitation on cell growth.	Built for purely bacterial applications, no results were presented in the original paper.	Goelzer and Fromion [77]
Multiscale Models of Metabolism and Macromolecular Expression (ME)	Integration of macromolecular machinery into metabolic network.	Prediction of growth in *E. coli*.	Thiele et al. [35]

Loopless-COBRA	Applies mixed-integer linear programming to prevent cyclic fluxes without thermodynamic data.	Impact on the solution space of an *H. pylori* network.	Schellenberger et al. [49]
CycleFreeFlux	Post-process flux distributions to remove cyclic fluxes.	Impact on the solution space of an *E. coli* network in glucose-limited aerobic medium.	Desouki et al. [78]
Regulatory FBA	Incorporates gene regulatory logic (e.g. Boolean rules) to simulate regulatory control of metabolic fluxes.	Gene expression and metabolic adaptation in *E. coli* and *M. tuberculosis*.	Covert et al. [79]
NEXT-FBA (Neural-net EXtracellular Trained Flux Balance Analysis)	Correlates exometabolomic data with ^{13}C-labeled intracellular fluxomic data to compute upper and lower bounds for intracellular reaction fluxes for constraining GEMs.	Constraining CHO cell GEM for improved accuracy of intracellular flux distribution.	Morrissey et al. [80]

Table 1.2 Examples of metabolic objective methodologies, including those guided or inferred from omics data.

Method	Description	Original application	Reference
pFBA (parsimonious FBA)	Minimizes the sum of all fluxes in the network.	Applied to *E. coli* to better understand genotype-phenotype reactionship.	Lewis et al. [86]
IMAT (Integrative Metabolic Analysis Toolbox)	Categorizes reactions as low/moderate/high expression. MILP maximizes the number of reactions whose activity is consistent with their expression state.	Method applied to a toy model as an example.	Zur et al. [28]
SPOT (Simplified Pearson cOrrelation with Transcriptomic data)	Uses a Pearson product-moment correlation to maximize the link between gene expression and flux.	Predicting intracellular flux distribution in *E. coli* and *S. cerevisiae*.	Kim et al. [81]
GIMME (Gene Inactivity Moderated by Metabolism and Expression)	GIMME minimizes usage of low-expression reactions while keeping a metabolic objective (e.g. biomass and productivity) above a certain threshold.	Produce context-specific metabolic networks for *E. coli* for several different conditions. Produce genome-scale models for particular human cells.	Becker and Palsson [27]
MOMA (Minimization Of Metabolic Adjustment)	Predicts post-perturbation fluxes by minimizing the distance from the wild-type state (typically using Euclidean distance in flux space).	Predicting flux distributions in *E. coli* mutants with disrupted pyruvate kinase.	Segrè et al. [87]
PROM (Probabilistic Regulation Of Metabolism)	Integrates gene regulatory networks and transcriptomics into GEMs by estimating the probability that a gene is active and adjusting flux bounds accordingly.	Context-specific modeling in *E. coli* and *M. tuberculosis* to predict condition-specific metabolic states.	Chandrasekaran and Price [88]

ObjFind	Finds the objective function that best matches experimental flux data by assigning metabolite weights.	*E. coli* in aerobic and anaerobic conditions to compare metabolic objectives.	Burgard and Maranas [82]
invFBA (inverse FBA)	Two-step optimization problems that minimize the error between measured and predicted objective functions in linear, quadratic.	Applied fluxes measured in the central carbon metabolism of ancestral and evolved *E. coli* strains.	Zhao et al. [84]
BOSS (Biological Objective Solution Search)	Searches over randomly generated objectives to find the one best matching experimental fluxes.	*S. cerevisiae* central metabolic network to evaluate which objective reaction it infers.	Gianchandani et al. [83]
SCOOTI (Single-Cell Optimization Objective and Trade-off Inference)	Infers cell-specific metabolic objectives and trade-offs by integrating GEMs with bulk or single-cell omics data.	Studying embryogenesis and cell cycle with single-cell transcriptomics and flux models.	Lin et al. [85]
CellFIE (Cellular Functions InferencE)	Summarize flux solutions of transcriptomics-based context-specific models with metabolic subsystems to provide scores for metabolic tasks.	Determining ranking of cellular tasks for various human tissues.	Richelle et al. [89]
ParTi (Pareto Task Inference)	Identifies transcriptomic "archetypes" and links them to metabolic tasks using Pareto fronts.	Modeling cancer transcriptomes to infer key metabolic tasks and drug sensitivity.	Hart et al. [90]

comparisons with experimental data across multiple axes, depending on the context of the model and the application. In the following sections, we outline several key benchmarks for evaluating model performance.

1.5.3.1 Growth Rate Predictions

A key initial benchmark for model validation is to assess predictive accuracy and to compare the predicted growth rate (through biomass reaction maximization) to experimentally measured growth rates under matched culture conditions. This provides a direct readout of how well the model reflects the overall metabolic capacity of the cell and the appropriateness of the biomass reaction. Deviations between predicted and observed growth can signal poor model formulation, missing constraints, and inaccurate biomass reaction construction.

1.5.3.2 Amino Acid Auxotrophies

Amino acid auxotrophies (where an organism is unable to synthesize an amino acid that it needs for growth) can be assessed by limiting the exchange reactions for specific amino acids and observing whether the model becomes infeasible when running an FBA biomass maximization. This simulates conditions where the organism is unable to produce a particular amino acid, reflecting its dependency on external sources for growth. Accurate models should correctly predict known auxotrophies, where the absence of an essential nutrient renders the organism nonviable.

For example, in CHO cells, auxotrophies for amino acids like proline, cystine, and arginine have been reported. In one study, the authors evaluated several CHO models by restricting the exchange reactions for these amino acids [91]. The models were tested to see if growth would be inhibited when proline, cystine, or arginine was unavailable, reflecting known auxotrophies. Accurate prediction of these dependencies demonstrates that the model successfully captures the metabolic requirements of CHO cells, making it a useful tool for optimizing culture conditions or improving strain robustness in industrial applications.

1.5.3.3 Gene Essentialities

Gene essentiality analysis is a critical method for validating genome-scale models by determining whether the inactivation of a gene leads to cell death, as predicted by the model. This is achieved by deleting reactions associated with a particular gene and testing the model's viability. The process simulates the knockout of genes that encode essential enzymes or metabolic functions, allowing researchers to assess whether the model accurately predicts the organism's survival under such conditions.

To validate these predictions, *in silico* gene knockouts are often compared to experimental results, such as those generated through CRISPR knockout screens. CRISPR technology enables precise gene deletions *in vivo*, providing experimental data on which genes are essential for cell survival [92]. A robust model will correctly predict these essential genes, while discrepancies may indicate gaps in the model, such as missing pathways or incomplete regulatory information. For example, the iML1515 model of *E. coli* showed 93.4% accuracy in predicting gene essentiality when tested under minimal media containing 16 different carbon sources [93]. This high level of agreement between computational predictions and experimental data emphasizes the quality of this model and the utility of gene deletion screens in validating genome-scale models and identifying areas where models can be improved.

1.5.3.4 Known Host Traits

Models can also be benchmarked by their ability to reproduce known metabolic or physiological characteristics of the host organism. For example, the Crabtree effect in yeast or the Warburg effect in cancer cells is a well-documented trait that should be captured by accurate GEMs. Additionally, uptake and secretion profiles of amino acids, glucose, or waste products can be compared directly to extracellular metabolomic data. The more traits the model can replicate without overfitting, the more confidence one can have in its biological fidelity.

1.5.3.5 Intracellular Predictive Accuracy

A crucial aspect of genome-scale models is their ability to predict intracellular fluxes, providing insights into the metabolic state of the cell at a genome-wide level. These predictions enable researchers to understand cellular metabolism in detail, facilitating the design of industrial processes and the engineering of optimized strains. To ensure that these predictions are reliable, it is essential to validate the model's ability to accurately predict intracellular fluxes.

One of the most effective methods for this validation is through comparisons with experimental ^{13}C-metabolic flux analysis (MFA) data. By incorporating ^{13}C-labeled substrates and tracking their incorporation into metabolic products, researchers can derive actual flux distributions within the cell, providing a robust dataset for comparison with GEM predictions [94, 95].

Due to the inherent issues with multiple optimal solutions using FBA to simulate metabolism in GEMs, flux sampling is recommended as a method for generating intracellular flux predictions (see Section 1.4.3). Flux sampling generates a range of possible solutions, allowing researchers to assess the variability of flux distributions rather than relying on a single optimal solution. These sampled flux distributions can then be compared with ^{13}C-MFA results to assess the accuracy of the model's predictions [96].

This approach has been successfully applied in the literature for industrially relevant cells. In one body of work, researchers compared the flux predictions from CHO cell GEMs with ^{13}C-MFA data, providing insights into the model's reliability in predicting intracellular metabolic activity [91]. These comparisons are critical for ensuring the model's utility in guiding cell line engineering and optimizing biotechnological processes and must be considered a part of any model evaluation workflow.

1.5.4 Toward Multilayer, Multiscale Metabolic Networks

While GEMs have provided a powerful framework for modeling metabolism, they only capture one layer of cellular function. In reality, biological systems operate through an interplay of multiple processes, including transcription, translation,

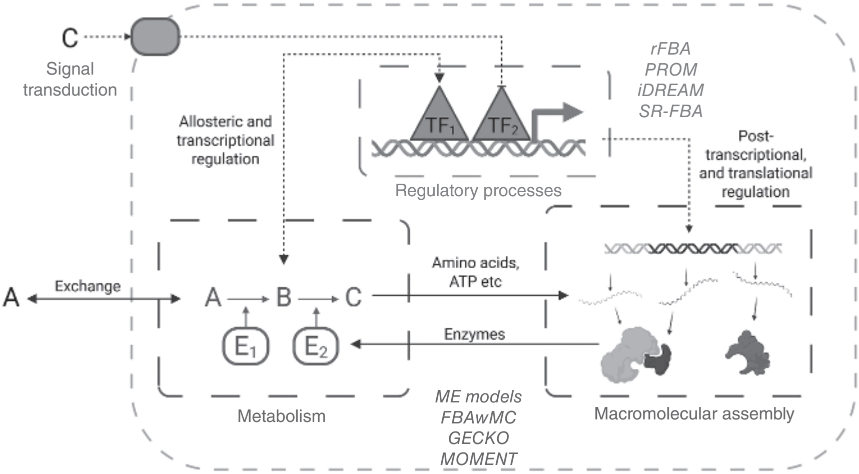

Figure 1.5 Overview of multi-layer modelling of Cellular Function. Core metabolic networks can be constrained by regulatory processes through TFs. ME models consider the interaction between transcription and translation machinery, the resources needed to synthesize them, and how they interact with metabolic function. Simpler resource allocation methods like GECKO consider enzyme kinetics and abundance, but not the full macromolecular machinery. Signal transduction influences network function. In this example, the accumulation of metabolite *C* in the extracellular environment may trigger a cascade to activate TF2 and prevent the expression of the gene required to catalyze the *B*→*C* reaction to prevent overproduction of metabolite *C*. Diagram created in BioRender.

signaling, and regulation. To move toward a more complete understanding of cell behavior, efforts are focused on developing multilayer and multiscale models that integrate these biological processes with metabolic networks, as illustrated in Figure 1.5.

1.5.4.1 Integrating Gene Regulatory Networks

While metabolic networks offer a great method to integrate omics data to simulate cell behavior and assess genetic perturbations, they do not capture any regulatory interactions. For example, *E. coli* has hundreds of TFs that can turn genes on or off in a combinatorial fashion. In more complex organisms, these regulatory processes extend to a myriad of posttranslational and posttranscriptional interactions that govern cellular behavior.

The first approach to incorporate transcriptional regulation into metabolic models, regulatory FBA (rFBA) [79] combines Boolean regulatory networks with FBA. Regulatory constraints (e.g. presence or absence of TFs) are used to turn reactions on or off, enabling simulation of regulatory effects on metabolism under dynamic environmental conditions. Though simplistic in its logic layer, rFBA provides a tractable method to include gene regulation and has been used in *E. coli*, *M. tuberculosis*, and *S. cerevisiae*.

PROM [88] improves upon rFBA by relaxing the Boolean logic and estimating the probability that a regulatory gene/TF is active and then adjusting the flux bounds of associated reactions accordingly. This probabilistic framework has been applied to various organisms, including *M. tuberculosis*, offering condition-specific metabolic predictions based on transcriptional profiles. Other methods include IDREAM [97], SR-FBA [98], iFBA [99], and TRFBA [100].

1.5.4.2 Integrating Transcription and Translation

Beyond transcriptional regulation, transcription and translation interact with metabolism through their dependence on metabolic precursors (e.g. RNA, amino acids, ATP) for macromolecule synthesis, while the resulting proteins catalyze metabolic reactions. Capturing this dependency can impose additional constraints and offer deeper insight into metabolic networks.

ME models address this by integrating GEMs with the transcriptional and translational machinery required to produce metabolic enzymes [35, 36]. These models impose global, growth-related regulatory constraints on metabolism and have outperformed standalone GEMs in predicting phenotypes such as growth, metabolic fluxes, and gene expression levels [36]. ME models do not explicitly capture specific regulatory mechanisms, which is an important area for development of multiscale ME models.

1.5.4.3 Integrating Signaling Networks

Cells respond to changing environments by adjusting gene expression and metabolic activity, a process tightly coordinated by signal transduction pathways. These pathways relay external cues to intracellular targets, influencing regulatory and metabolic networks. Integrating signaling into metabolic and regulatory models can provide context and impose additional constraints, improving predictive accuracy. However, reconstructing intracellular signaling networks remains a major challenge, with large-scale maps available for only a limited number of organisms [99, 101]. Since independent characterization of these pathways is both time-consuming and resource-intensive, a more scalable approach is to leverage high-throughput datasets to infer how environmental changes influence cellular coordination.

1.5.4.4 Multicellular and Multitissue Models

Biological function often emerges not from isolated cells but from coordinated interactions across tissues and cell types. As such, there is growing interest in extending GEMs beyond the single-cell level to simulate multicellular and multitissue systems. One approach involves connecting multiple tissue-specific GEMs via shared extracellular compartments, enabling simulation of metabolite exchange across organs [102]. This strategy has been applied in whole-body metabolic models of humans, where liver, muscle, adipose, and gut models interact to study systemic metabolism, such as in diabetes or fasting responses [103].

In the context of biotechnology, coculture microbial cell cultures have been studied [104], as well as combining mammalian and microbe GEMs to study host–microbiome interactions [105]. Cancer research has leveraged tumor–stromal metabolic models to capture nutrient competition and microenvironmental effects [106]. Frameworks such as the human metabolic atlas and resources like the virtual metabolic human database have facilitated the construction of such models by providing tissue-specific expression data and curated reaction sets [107, 108].

1.5.4.5 Multiscale Bioreactor Models

In the above sections, "multiscale" has referred to layering biological processes to gain deeper insight into each layer individually. In the context of biotechnology, it may be prudent to add an additional scale, the bioreactor. To link molecular models with population or bioreactor-scale behavior, multiscale models integrate intracellular networks with higher-level process models describing growth, nutrient consumption, and bioreactor conditions [109–113]. To make these simulations computationally tractable, reduced-order models or simplified GEMs are often used in place of full-scale metabolic reconstructions. Dynamic FBA (dFBA) approaches [114] integrate time-dependent environmental changes with FBA simulations over discrete intervals. Model reduction techniques, such as elementary mode analysis or using core metabolic subnetworks, allow faster simulation while retaining essential metabolic features [115]. As cells exhibit heterogeneity in metabolic activity in cell culture due to asynchronous growth, cell cycle states, and location, population balance models can be used [116].

1.6 From Predictions to Actionable Insights in Biotechnology

Metabolic networks and constraint-based modeling offer powerful tools for simulating metabolism, but their true value lies in enabling real-world applications across industrial biotechnology. This chapter section focuses on translating *in silico* predictions into practical strategies for metabolic engineering, process design, and bioprocess control. To further enhance these strategies, the integration of mechanistic and data-driven approaches is increasingly adopted.

Hybridization of network-based and data-driven tools allows the biological context and structure of network approaches to be combined with the pattern recognition strengths of data-driven methods, offering enhanced predictive power and interpretability. Broadly, hybrid approaches are used in two ways: either to inform and constrain network models using data-driven predictions or to use outputs from network models as features for data-driven algorithms (summarized in Figure 1.6). These two strategies provide complementary routes to deepen biological insight and improve model utility in biotechnology. In the following section, we highlight several key examples that demonstrate their application.

1.6.1 Metabolic Engineering

GEMs formalize biological knowledge and enable the design and quantitative comparison of metabolic engineering strategies *in silico*. They are a smart test bed for screening such strategies efficiently before selecting a subset for experimental investigation as they capture the complexity of cellular systems and the intricate network

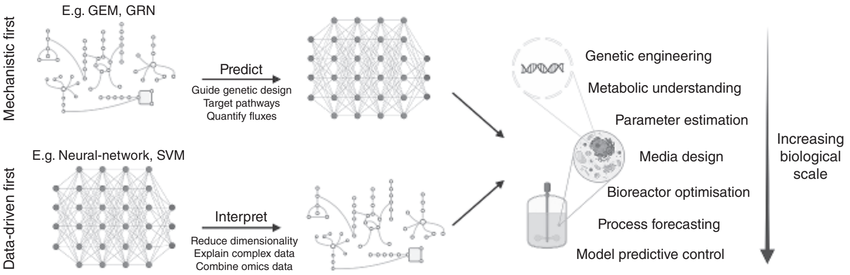

Figure 1.6 Schematic summarizing how data-driven and mechanistic approaches can be hybridized to generate insights in biotechnology. Hybridization can be mechanistic first, meaning mechanistic predictions (such as flux from a GEM) can be inputted to data-driven tools. Alternatively, data-driven tools can be used to contextualize and interpret complex data to inform mechanistic models. Both approaches have yielded actionable insights in biotechnology. Diagram created in BioRender.

of biochemical reactions that are candidates for manipulation. By integrating large datasets and creating custom GEMs, researchers can then apply a variety of novel optimization algorithms for target identification as detailed in Table 1.3, reducing the trial-and-error of experimental methods and accelerating the development of improved strains.

These algorithms primarily use mixed-integer linear programming (MILP) to identify candidate reactions (and therefore genes) for manipulation, e.g. knockout, knockdown, or overexpression. For example, OptKnock [117] uses a bilevel optimization framework to suggest gene knockouts that link the overproduction of a desired product (metabolic or recombinant protein) to cellular growth. A number of other tools have expanded on this concept [118], with subsequent algorithms incorporating knock-ins of non-native functions from extensive reaction databases [126]. Most of these efforts were originally developed for and applied to microbial cell systems, which are described by smaller stoichiometric networks. Moving to plant or mammalian cell systems, where the networks expanded from 100s to 1,000s of equations, makes the application of these algorithms more computationally demanding [127].

1.6.2 Cell Line Development and Metabolic Profiling

In parallel to algorithmic approaches, researchers have also used GEMs to guide cell line development in a data-driven fashion. For example, tailored CHO cell GEMs solved using pFBA have been used to identify key metabolic reactions that distinguish low- and high-producing cell lines [128]. The authors identified glucose-6-phosphate dehydrogenase (G6PD) and PDH as potential biomarkers, as well as a bottleneck at malate dehydrogenase (MDH) in low producers, in line with previously published work [129]. Similarly, Huang and Yoon [130] studied transcriptomic and metabolomic data using a CHO GEM to identify the metabolic pathways that determine recombinant protein productivity in two clones grown in batch cultures. Their analysis confirmed the importance of G6PD and PDH in high

Table 1.3 Summary of key optimization algorithms for strain design.

Algorithm	Description	Reference
OptKnock	Suggests gene knockouts based on a bilevel optimization framework that couples the desired overproduction target (i.e. the biopharmaceutical) to growth.	Burgard et al. [117]
OptReg	Extension of OptKnock that suggests reactions for regulation (up, down, etc.) based on a bilevel optimization framework that couples the fluxes of two different reactions.	Pharkya and Maranas [118]
OptForce	Extension of OptReg that contrasts the metabolic flux patterns observed in an initial strain and a strain overproducing the chemical at the target yield.	Ranganathan et al., [119]
GDLS	Uses logical search to look for multiple paths to improve the production of metabolites.	Lun et al. [120]
OptORF	Identifies optimal number of metabolic and regulatory gene knockouts/overexpressions to couple production with growth.	Kim and Reed [121]
k-OptForce	Identifies regulators in transcription factors and metabolic genes by accounting for gene expression in its objective function.	Chowdhury et al. [122]
OptRam	Extension of OptForce that uses kinetic equations to better calculate steady-state fluxes of metabolic networks.	Shen et al. [123]
gcOpt	Predicts coupling strength of two metabolic reactions.	Alter and Ebert [124]
OptCouple	Analyzes both genetic engineering strategies with process engineering solutions for the overproduction of a metabolite.	Jensen et al. [125]
OptStrain	Assesses knock-ins of non-native functionalities from a comprehensive database of reactions.	Pharkya et al. [126]

Source: Antonakoudis et al., [11] / with permission of Elsevier.

producers and revealed a tricarboxylic acid (TCA) cycle bottleneck at succinyl-CoA synthetase. Yusufi et al. [131] validated GEM predictions with metabolomic data, identifying metabolites associated with energy metabolism and oxidative phosphorylation, the concentration of which is elevated in protein-producing CHO cells. In a study across different bioreactor scales ranging from 10to 1,000 L, Vodopivec et al. [132] employed a GEM to analyze metabolomics data and drive our understanding of the differences in metabolic states that occur during upstream process scale-up. Kol et al. [133] used a CHO cell GEM that includes the secretory pathway to quantify the secretion cost of abundant host cell proteins (HCPs). They used the results to rank HCP as knockout targets and subsequently implemented sequential knockouts of burdensome HCPs achieving a 40–70% reduction in HCP content in the supernatant. In addition to improving cell growth and productivity, these modifications generated clean feedstock for downstream purification steps.

1.6.3 Media and Feed Design

Multiomics datasets and derived GEMs have also found application in the design of cell culture media and supplements. GEMs can be used to elucidate the impact of changes in the culture environment on intracellular flux distribution and therefore identify bottlenecks for cell growth or the production of a metabolite/recombinant product. This understanding can then be used to propose improved process strategies, including media and feed formulations. An example of such an application is the work of Huang et al., in which a GEM together with transcriptomic data were used to identify metabolic pathways that are upregulated in high-producing CHO cell clones [134]. Their analysis highlighted branched-chain amino acids as a key source of TCA cycle intermediates, which led to the design of a new cell culture feed rich in valine and leucine. The feed was tested experimentally and was demonstrated to result in a significant increase in recombinant protein titer. A similar study used a CHO-DG44 GEM to reveal a link between folate supplementation in media and lactate secretion. This mechanism balances increased lactate production by suppressing TCA cycle flux with reduced lactate production through decreased glutamine/asparagine anaplerosis [135]. GEMs have also been used to understand the metabolism of asparagine and aspartate for feed ratio optimization [136], to design potential feed supplements for recombinant protein titer optimization [137], and to reduce the accumulation of toxic metabolic byproducts like ammonia by improving feed formulation [138]. Additionally, GEMs have been used to identify succinate dehydrogenase (SDH) as a bottleneck in early cell culture, leading to media reformulation to include coenzyme q10 (ubiquinone) for enhanced FAD+ regeneration [139].

The analysis of large fluxomic datasets resulting from metabolic models can be facilitated by statistical or ML methods toward new knowledge generation. To this end, Ramos et al. analyzed the results of a CHO-K1 GEM with principal component analysis (PCA) to design an efficient growth medium that minimized by-product accumulation [140]. Previously, isotope labeling experimental datasets analyzed using small-scale metabolic models to infer the flux of key metabolic pathways have been further interrogated using a variety of ML techniques. For example, Wu et al. developed a platform that applied support vector machines (SVMs), k-nearest neighbors, and decision trees to literature data from nearly 100 ^{13}C-FMA studies on heterotrophic bacteria [141]. The platform predicts fluxomes as a function of bacterial species, substrate type, growth rate, oxygen conditions, and cultivation methods. Nandi et al. [142] extended the methodology to include gene sequencing and expression, network topology, and flux-based features. By also accounting for environmental factors, the model was able to capture the minimal set of genes that are essential in any given environment.

NEXT-FBA [80] is another recent example of using ML methods to understand complex biological data. NEXT-FBA uses neural networks trained on metabolomic and fluxomic data to understand the relationship between a cell's internal and external dynamics. This understanding is used to compute constraints on GEMs that create more refined and accurate predictions. NEXT-FBA was used to identify the causes of ammonia accumulation in a CHO cell process, determining feed and metabolic engineering process improvements.

1.6.4 Gene Essentiality

Understanding how genetic perturbations lead to phenotypic changes is a key application of metabolic networks. Combined data-driven and mechanistic approaches have been used to identify essential genes. For example, Plaimas et al. used SVMs on an *E. coli* FBA network with genomic and transcriptomic data to identify essential enzymes [143]. Acencio and Lemke extended this by applying decision trees to topological, compartmental, and functional features, improving gene essentiality prediction and revealing phenotype determinants [144].

Graph-based learning is another emerging data-driven-first approach, where metabolic reaction networks – constructed from empirical data or draft reconstructions – are treated as graphs and analyzed using graph neural networks (GNNs). Nodes may represent metabolites or reactions, and edges capture biochemical or topological relationships. By training GNNs on these graphs with omics data as node or edge features, researchers can predict reaction essentiality, flux control points, or context-specific pathway activation [145]. This approach retains biological structure without requiring a fully curated GEM, making it particularly useful for less-characterized organisms or exploratory studies where network topology and data integration are the primary drivers of insight.

1.6.5 Kinetic Parameter Estimation

Another application of data-driven-first applications is parameter estimation. A major barrier to developing large-scale kinetic or enzyme-constrained network models is the lack of high-quality kinetic parameters. As an early solution, data-driven methods were used to predict kinetic values to parameterize network models. For instance, a statistical regression method was used to estimate kinetic parameters within appropriate experimental ranges [146]. More recently, ML algorithms were trained on a dataset incorporating enzyme biochemistry, protein structure, and network context to predict enzyme turnover numbers, used in proteome-constrained models [147, 148].

1.6.6 Process Monitoring and Forecasting

A key focus of the biopharmaceutical industry has been a move from quality by testing to quality by design (QbD) in process development and manufacturing (ICH Q8 guideline). Mathematical model-based process monitoring, forecasting, and control is an important part of the QbD approach. Models designed for cell culture control must be complex enough to understand the link between critical process parameters (CPPs) and both the key performance indicators (KPIs) and the critical quality attributes (CQAs) in the process, but as the same time simple enough to allow fast action by the controller. Mathematical models for mammalian cell culture process control can be data-driven, based on simple mechanistic relations, or more complex metabolic models [11, 149].

GEMs offer a unique perspective on process control as they link CPPs and KPIs/CQAs by modeling cell metabolism, rather than treating cells as a black box for the conversion of substrates to products using a set of macroscopic reactions without considering intracellular behavior. GEMs can offer insights that macroscopic models cannot and therefore improving forecasting and control. CHO cell GEMs have been used to predict amino acid concentrations by applying minimization of uptake rate objective combined with forecasted viable cell densities [150]. In addition, Schinn et al. combined CHO cell GEM flux sampling with statistical models to describe time-course amino acid consumption [151], while Antonakoudis et al. used GEM-computed fluxes of nucleotide sugars as inputs to an ML model that predicted the glycosylation profile of the protein product [152].

However, several limitations hinder the practical use of network models in process control. Their complexity often exceeds what is necessary for actionable control, increasing computational demand without proportional benefit. The underdetermined nature of these models leads to broad solution spaces, reducing predictive precision unless constrained appropriately. Furthermore, the reliance on steady-state assumptions limits forecasting accuracy over longer culture durations, necessitating dynamic modeling approaches. Lastly, GEMs alone cannot capture the impact of key CPPs such as temperature, pH, and toxic metabolite accumulation on metabolism. To overcome these issues, simplification strategies, dynamic extensions, and integration with empirical or data-driven methods (such as reinforcement learning or hybrid MPC frameworks) are essential for effective implementation.

In this direction, metabolic predictions from a small-scale CHO cell metabolic model have been integrated into dynamic models of production bioreactors [109, 112]. Finally, Gopalakrishnan et al. (2024) [110] proposed the use of a data-driven classifier for categorizing cells into "growth" or "production" states, each with distinct metabolic objectives. Using this term, a GEM was embedded within a system of ODEs to predict metabolite uptake and secretion throughout a fed-batch CHO cell culture, successfully capturing the coexistence of two cellular states.

References

1 Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2: 342–372.

2 Kitano, H. (2002). Systems biology: a brief overview. *Science* 295: 1662–1664.

3 Barabási, A.L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101–113.

4 Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nat. Rev. Genet.* 14: 333–346.

5 Pal, S., Mondal, S., Das, G. et al. (2020). Big data in biology: the hope and present-day challenges in it. *Gene Rep.* 21: 100869.

6 Lewis, N.E., Nagarajan, H., and Palsson, B.O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10: 291–305.

7 Palsson, B. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press.

8 Thiele, I. and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5: 93–121.

9 Fang, X., Lloyd, C.J., and Palsson, B.O. (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* 18: 731–743.

10 Zielinski, D.C., Patel, A., and Palsson, B.O. (2020). The expanding computational toolbox for engineering microbial phenotypes at the genome scale. *Microorganisms* 8: 2050.

11 Antonakoudis, A., Barbosa, R., Kotidis, P., and Kontoravdi, C. (2020). The era of big data: genome-scale modelling meets machine learning. *Comput. Struct. Biotechnol. J.* 18: 3287, 3300.

12 Passi, A., Tibocha-Bonilla, J.D., Kumar, M. et al. (2022). Genome-scale metabolic modeling enables in-depth understanding of big data. *Metabolites* 12: 14.

13 Francke, C., Siezen, R.J., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* 13: 550–558.

14 Gu, C., Kim, G.B., Kim, W.J. et al. (2019). Current status and applications of genome-scale metabolic models. *Genome Biol.* 20: 121.

15 Strain, B., Morrissey, J., Antonakoudis, A., and Kontoravdi, C. (2023). Genome-scale models as a vehicle for knowledge transfer from microbial to mammalian cell systems. *Comput. Struct. Biotechnol. J.* 21: 1543–1549.

16 Ashburner, M., Ball, C.A., Blake, J.A. et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25: 25–29.

17 Barrett, A.J. (1995). Enzyme Nomenclature. Recommendations 1992. *Eur. J. Biochem.* 232: 1.

18 Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27–30.

19 Placzek, S., Schomburg, I., Chang, A. et al. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* 45: D380–D388.

20 Jankowski, M.D., Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2008). Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* 95: 1487–1499.

21 Caspi, R., Billington, R., Fulcher, C.A. et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46: D633–D639.

22 Henry, C.S., Dejongh, M., Best, A.A. et al. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28: 977–982.

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

24 Karp, P.D., Paley, S.M., Krummenacker, M. et al. (2009). Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings Bioinf.* 11: 40–79.

25 Gopalakrishnan, S., Joshi, C.J., Valderrama-Gómez, M. et al. (2023). Guidelines for extracting biologically relevant context-specific metabolic models using gene expression data. *Metab. Eng.* 75: 181–191.

26 Richelle, A., Chiang, A.W.T., Kuo, C.C., and Lewis, N.E. (2019). Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. *PLoS Comput. Biol.* 15: e1006867.

27 Becker, S.A. and Palsson, B.O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4: e1000082.

28 Zur, H., Ruppin, E., and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. *Bioinformatics* 26: 3140–3142.

29 Schultz, A. and Qutub, A.A. (2016). Reconstruction of tissue-specific metabolic networks using CORDA. *PLoS Comput. Biol.* 12: e1004808.

30 Babu, M.M., Luscombe, N.M., Aravind, L. et al. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14: 283–291.

31 Blais, A. and Dynlacht, B.D. (2005). Constructing transcriptional regulatory networks. *Genes Dev.* 19: 1499–1511.

32 Covert, M.W., Knight, E.M., Reed, J.L. et al. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429: 92–96.

33 Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13: 966–967.

34 Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M. et al. (2014). Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol Hepatol Bed Bench.* 7: 17.

35 Thiele, I., Fleming, R.M.T., Que, R. et al. (2012). Multiscale Modeling of metabolism and macromolecular synthesis in *E. Coli* and its application to the evolution of codon usage. *PLoS One* 7: e45635.

36 O'Brien, E.J., Lerman, J.A., Chang, R.L. et al. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* 9: 693.

37 Allaman, I., Bélanger, M., and Magistretti, P.J. (2015). Methylglyoxal, the dark side of glycolysis. *Front. Neurosci.* 9: 23.

38 Held, K.D., Sylvester, K.C., Hopcia, K.L., and Biaglow, J.E. (1996). Role of Fenton chemistry in thiol-induced toxicity and apoptosis. *Radiat. Res.* 145: 542–553.

39 Selvarasu, S., Ow, D.S.W., Lee, S.Y. et al. (2009). Characterizing *Escherichia coli DH5α* growth and metabolism in a complex medium using genome-scale flux analysis. *Biotechnol. Bioeng.* 102: 923–934.

40 Széliová, D., Štor, J., Thiel, I. et al. (2020). Inclusion of maintenance energy improves the intracellular flux predictions of CHO. *PLoS Comput. Biol.* 17: e1009022.

41 Feist, A.M. and Palsson, B.O. (2010). The biomass objective function. *Curr. Opin. Microbiol.* 13: 344–349.

42 Aminian-Dehkordi, J., Mousavi, S.M., Jafari, A. et al. (2019). Manually curated genome-scale reconstruction of the metabolic network of *bacillus megaterium* DSM319. *Sci. Rep.* 9: 18762.

43 Price, N.D., Reed, J.L., and Palsson, B. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2: 886–897.

44 Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. (2003). An expanded genome-scale model of Escherichia coli *K-12* (iJR904 GSM/GPR). *Genome Biol.* 4: R54.

45 Mahadevan, R. and Schilling, C.H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5: 264–276.

46 Thiele, I., Vlassis, N., and RMT, F. (2014). FASTGAPFILL: efficient gap filling in metabolic networks. *Bioinformatics* 30: 2529–2531.

47 Beard, D.A., Liang, S.D., and Qian, H. (2002). Energy balance for analysis of complex metabolic networks. *Biophys. J.* 83: 79–86.

48 Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophys. J.* 92: 1792–1805.

49 Schellenberger, J., Lewis, N.E., and Palsson, B. (2011). Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys. J.* 100: 544–553.

50 Fleming, R.M.T., Haraldsdottir, H.S., Minh, L.H. et al. (2023). Cardinality optimization in constraint-based modelling: application to human metabolism. *Bioinformatics* 39: btad450.

51 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: 639–702.

52 Orth, J.D., Thiele, I., and Palsson, B.O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28: 245–248.

53 Varma, A. and Palsson, B.O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 60: 3724–3731.

54 Srivastava, S. and Chan, C. (2008). Application of metabolic flux analysis to identify the mechanisms of free fatty acid toxicity to human hepatoma cell line. *Biotechnol. Bioeng.* 99: 399–410.

55 Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinf.* 11: 489.

56 Schellenberger, J. and Palsson, B. (2009). Use of randomized sampling for analysis of metabolic networks. *J. Biol. Chem.* 5457–5461.

57 Herrmann, H.A., Dyson, B.C., Vass, L. et al. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Syst. Biol. Appl.* 5: 32.

58 Wiback, S.J., Famili, I., Greenberg, H.J., and Palsson, B. (2004). Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J. Theor. Biol.* 228: 437–447.

59 Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res.* 32: 1296–1308.

60 Kaufman, D.E. and Smith, R.L. (1998). Direction choice for accelerated convergence in hit-and-run sampling. *Oper. Res.* 46: 84–95.

61 Haraldsdóttir, H.S., Cousins, B., Thiele, I. et al. (2017). CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics* 33: 1741–1743.

62 Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* e1007054.

63 King, Z.A., Lloyd, C.J., Feist, A.M., and Palsson, B.O. (2015). Next-generation genome-scale models for metabolic engineering. *Curr. Opin. Biotechnol.* 35: 23–29.

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

65 Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* 6: e1000859.

66 Colijn, C., Brandes, A., Zucker, J. et al. (2009). Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* 5: e1000489.

67 Jensen, P.A. and Papin, J.A. (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics.* 27: 541–547.

68 Pusa, T., Ferrarini, M.G., Andrade, R. et al. (2020). MOOMIN – mathematical explOration of'omics data on a MetabolIc network. *Bioinformatics.* 36: 514–523.

69 Tian, M. and Reed, J.L. (2018). Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics.* 34: 3882–3888.

70 Wang, Y., Eddy, J.A., and Price, N.D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* 6: 153.

71 Lularevic, M., Racher, A.J., Jaques, C., and Kiparissides, A. (2019). Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. *Biotechnol. Bioeng.* 116: 2339–2352.

72 Beg, Q.K., Vazquez, A., Ernst, J. et al. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *PNAS* 104: 12663–12668.

73 Adadi, R., Volkmer, B., Milo, R. et al. (2012). Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput. Biol.* 8: e1002575.

74 Yeo, H.C., Hong, J., Lakshmanan, M., and Lee, D.Y. (2020). Enzyme capacity-based genome scale modelling of CHO cells. *Metab. Eng.* 60: 138–147.

75 Mori, M., Hwa, T., Martin, O.C. et al. (2016). Constrained allocation flux balance analysis. *PLoS Comput. Biol.* 12: e1004913.

76 Sánchez, B.J., Zhang, C., Nilsson, A. et al. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* 13: 935.

77 Goelzer, A. and Fromion, V. (2011). Bacterial growth rate reflects a bottleneck in resource allocation. *Biochim. Biophys. Acta, Gen. Subj.* 1810: 978–988.

78 Desouki, A.A., Jarre, F., Gelius-Dietrich, G., and Lercher, M.J. (2015). CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. *Bioinformatics* 31: 2159–2165.

79 Covert, M.W., Schilling, C.H., and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* 213: 73–88.

80 Morrissey, J., Barberi, G., Strain, B. et al. (2025). NEXT-FBA: a hybrid stoichiometric/data-driven approach to improve intracellular flux predictions. *Metab. Eng.* 91: 130–144. https://www.sciencedirect.com/science/article/pii/S1096717625000461.

81 Kim, M.K., Lane, A., Kelley, J.J., and Lun, D.S. (2016). E-Flux2 and sPOT: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. *PLoS One* 11: e0157101.

82 Burgard, A.P. and Maranas, C.D. (2003). Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* 82.

83 Gianchandani, E.P., Oberhardt, M.A., Burgard, A.P. et al. (2008). Predicting biological system objectives de novo from internal state measurements. *BMC Bioinf.* 9: 670–677.

84 Zhao, Q., Stettner, A.I., Reznik, E. et al. (2016). Mapping the landscape of metabolic goals of a cell. *Genome Biol.* 17: 109.

85 Lin, D.-W., Zhang, L., Zhang, J., and Chandrasekaran, S. (2025). Inferring metabolic objectives and trade-offs in single cells during embryogenesis. *Cell Syst.* 16: 101164. https://www.sciencedirect.com/science/article/pii/S2405471224003703.

86 Lewis, N.E., Hixson, K.K., Conrad, T.M. et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* 6: 390.

87 Segrè, D., Vitkup, D., and Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *PNAS* 99: 15112–15117.

88 Chandrasekaran, S. and Price, N.D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *PNAS* 107: 17845–17850.

89 Richelle, A., Kellman, B.P., Wenzel, A.T. et al. (2021). Model-based assessment of mammalian cell metabolic functionalities using omics data. *Cell Reports Methods.* 1: 100040.

90 Hart, Y., Sheftel, H., Hausser, J. et al. (2015). Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* 12: 233–235.

91 Strain, B., Morrissey, J., Antonakoudis, A., and Kontoravdi, C. (2023). How reliable are Chinese hamster ovary (CHO) cell genome-scale metabolic models? *Biotechnol. Bioeng.* 120: 2460–2478.

92 Wang, T., Birsoy, K., Hughes, N.W. et al. (2015). Identification and characterization of essential genes in the human genome. *Science* 350: 1096–1101.

93 Monk, J.M., Lloyd, C.J., Brunk, E. et al. (2017). iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* 904–908.

94 Young, J.D., Walther, J.L., Antoniewicz, M.R. et al. (2008). An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol. Bioeng.* 99: 686–699.

95 Antoniewicz, M.R. (2015). Methods and advances in metabolic flux analysis: a mini-review. *J. Ind. Microbiol. Biotechnol.* 42: 317–325.

96 Quek, L.E., Wittmann, C., Nielsen, L.K., and Krömer, J.O. (2009). Open FLUX: efficient modelling software for 13 C-based metabolic flux analysis. *Microb. Cell Fact.* 8: 25.

97 Wang, Z., Danziger, S.A., Heavner, B.D. et al. (2017). Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. *PLoS Comput. Biol.* 13: e1005189.

98 Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* 3: 101.

99 Covert, M.W., Xiao, N., Chen, T.J., and Karr, J.R. (2008). Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* 24: 2044–2050.

100 Motamedian, E., Mohammadi, M., Shojaosadati, S.A., and Heydari, M. (2017). TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinformatics* 33: 1057–1063.

101 Carrera, J., Estrela, R., Luo, J. et al. (2014). An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Mol. Syst. Biol.* 10: 735.

102 Bordbar, A., Feist, A.M., Usaite-Black, R. et al. (2011). A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst. Biol.* 5: 180.

103 Thiele, I., Sahoo, S., Heinken, A. et al. (2020). Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.* 16: e8982.

104 Zomorrodi, A.R. and Maranas, C.D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* e1002363.

105 Heinken, A., Sahoo, S., Fleming, R.M.T., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes* 4: 28–40.

106 Zhang, C. and Hua, Q. (2016). Applications of genome-scale metabolic models in biotechnology and systems medicine. *Front. Physiol.* 6: 413.

107 Noronha, A., Modamio, J., Jarosz, Y. et al. (2019). The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* 47: D614–D624.

108 Robinson, J.L., Kocabaş, P., Wang, H. et al. (2020). An atlas of human metabolism. *Sci. Signaling* 13: eaaz1482.

109 Monteiro, M., Fadda, S., and Kontoravdi, C. (2023). Towards advanced bioprocess optimization: a multiscale modelling approach. *Comput. Struct. Biotechnol. J.* 21: 3639–3655.

110 Gopalakrishnan, S., Johnson, W., Valderrama-Gomez, M.A. et al. (2024). COSMIC-dFBA: a novel multi-scale hybrid framework for bioprocess modeling. *Metab. Eng.* 82: 183–192.

111 Mahnert, C., Oyarzún, D.A., and Berrios, J. (2024). Multiscale modelling of bioprocess dynamics and cellular growth. *Microb. Cell Fact.* 23: 315. https://doi.org/10.1186/s12934-024-02581-0.

112 Lázaro, J., Jansen, G., Yang, Y. et al. (2022). Combination of genome-scale models and bioreactor dynamics to optimize the production of commodity chemicals. *Front. Mol. Biosci.* 9: 855735.

113 Yatipanthalawa, B.S., Wallace Fitzsimons, S.E., Horning, T. et al. (2024). Development and validation of a hybrid model for prediction of viable cell density, titer and cumulative glucose consumption in a mammalian cell culture system. *Comput. Chem. Eng.* 184: 108648.

114 Mahadevan, R., Edwards, J.S., and Doyle, F.J. (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* 83: 1331–1340.

115 von Kamp, A. and Klamt, S. (2014). Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Comput. Biol.* 10: e1003378.

116 Karra, S., Sager, B., and Karim, M.N. (2010). Multi-scale modeling of heterogeneities in mammalian cell culture processes. *Ind. Eng. Chem. Res.* 49: 7990–8006.

117 Burgard, A.P., Pharkya, P., and Maranas, C.D. (2003). OptKnock: a Bilevel programming framework for identifying gene knockout strategies for microbial Strain optimization. *Biotechnol. Bioeng.* 84: 647–657.

118 Pharkya, P. and Maranas, C.D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.* 8: 1–13.

119 Ranganathan, S., Suthers, P.F., and Maranas, C.D. (2010). OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* 6: e1000744.

120 Lun, D.S., Rockwell, G., Guido, N.J. et al. (2009). Large-scale identification of genetic design strategies using local search. *Mol. Syst. Biol.* 5: 296.

121 Kim, J. and Reed, J.L. (2010). OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst. Biol.* 4: 53.

122 Chowdhury, A., Zomorrodi, A.R., and Maranas, C.D. (2014). K-OptForce: integrating kinetics with flux balance analysis for Strain design. *PLoS Comput. Biol.* 10: e1003487.

123 Shen, F., Sun, R., Yao, J. et al. (2019). Optram: in-silico strain design via integrative regulatory-metabolic network modeling. *PLoS Comput. Biol.* 15: e1006835.

124 Alter, T.B. and Ebert, B.E. (2019). Determination of growth-coupling strategies and their underlying principles. *BMC Bioinf.* 20: 447.

125 Jensen, K., Broeken, V., Hansen, A.S.L. et al. (2019). OptCouple: joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs. *Metab. Eng. Commun.* 8: e00087.

126 Pharkya, P., Burgard, A.P., and Maranas, C.D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 14: 2367–2376.

127 Long, M.R., Ong, W.K., and Reed, J.L. (2015). Computational methods in metabolic engineering for strain design. *Curr. Opin. Biotechnol.* 34: 135–141.

128 Calmels, C., Arnoult, S., Ben Yahia, B. et al. (2019). Application of a genome-scale model in tandem with enzyme assays for identification of metabolic signatures of high and low CHO cell producers. *Metab. Eng. Commun.* 9: e00097.

129 Chong, W.P.K., Reddy, S.G., Yusufi, F.N.K. et al. (2010). Metabolomics-driven approach for the improvement of Chinese hamster ovary cell growth: overexpression of malate dehydrogenase II. *J. Biotechnol.* 147: 116–121.

130 Huang, Z. and Yoon, S. (2020). Identifying metabolic features and engineering targets for productivity improvement in CHO cells by integrated transcriptomics and genome-scale metabolic model. *Biochem. Eng. J.* 159: 107624.

131 Yusufi, F.N.K., Lakshmanan, M., Ho, Y.S. et al. (2017). Mammalian systems biotechnology reveals global cellular adaptations in a recombinant CHO cell line. *Cell Syst.* 4: 530–542.

132 Vodopivec, M., Lah, L., Narat, M., and Curk, T. (2019). Metabolomic profiling of CHO fed-batch growth phases at 10, 100, and 1,000 L. *Biotechnol. Bioeng.* 116: 2720–2729.

133 Kol, S., Ley, D., Wulff, T. et al. (2020). Multiplex secretome engineering enhances recombinant protein production and purity. *Nat. Commun.* 11: 1908.

134 Huang, Z., Xu, J., Yongky, A. et al. (2020). CHO cell productivity improvement by genome-scale modeling and pathway analysis: application to feed supplements. *Biochem. Eng. J.* 160: 107638.

135 Yeo, H.C., Park, S.Y., Tan, T. et al. (2022). Combined multivariate statistical and flux balance analyses uncover media bottlenecks to the growth and productivity of Chinese hamster ovary cell cultures. *Biotechnol. Bioeng.* 119: 1740–1754.

136 Pang, K.T., Hong, Y.F., Shozui, F. et al. (2024). Genome-scale modeling of CHO cells unravel the critical role of asparagine in cell culture feed media. *Biotechnol. J.* http://dx.doi.org/10.22541/au.171105421.18572875/v1.

137 Fouladiha, H., Marashi, S.A., Torkashvand, F. et al. (2020). A metabolic network-based approach for developing feeding strategies for CHO cells to increase monoclonal antibody production. *Bioprocess. Biosyst. Eng.* 43: 1381–1389.

138 Park, S.Y., Choi, D.H., Song, J. et al. (2023). Debottlenecking and reformulating feed media for improved CHO cell growth and titer by data-driven and model-guided analyses. *Biotechnol. J.* 18: 2300126.

139 Hong, J.K., Choi, D.H., Park, S.Y. et al. (2022). Data-driven and model-guided systematic framework for media development in CHO cell culture. *Metab. Eng.* 73: 114–123.

140 Ramos, J.R.C., Oliveira, G.P., Dumas, P., and Oliveira, R. (2022). Genome-scale modeling of Chinese hamster ovary cells by hybrid semi-parametric flux balance analysis. *Bioprocess. Biosyst. Eng.* 45: 1889–1904.

141 Wu, S.G., Wang, Y., Jiang, W. et al. (2016). Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. *PLoS Comput. Biol.* 12: e1004838.

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

143 Plaimas, K., Mallm, J.P., Oswald, M. et al. (2008). Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC Syst. Biol.* 2: 67.

144 Acencio, M.L. and Lemke, N. (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinf.* 10: 290.

145 Hasibi, R., Michoel, T., and Oyarzún, D.A. (2024). Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. *npj Syst. Biol. Appl.* 10: 24.

146 Borger, S., Liebermeister, W., and Klipp, E. (2006). Prediction of enzyme kinetic parameters based on statistical learning. *Genome Inform.* 17: 80–87.

147 Heckmann, D., Lloyd, C.J., Mih, N. et al. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat. Commun.* 9: 5252.

148 Ferreira, M.A.d.M., da Silveira, W.B., and Nikoloski, Z. (2024). Protein constraints in genome-scale metabolic models: data integration, parameter estimation, and prediction of metabolic phenotypes. *Biotechnol. Bioeng.* 121: 915–930.

149 Tsopanoglou, A. and Jiménez del Val, I. (2021). Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. *Curr. Opin. Chem. Eng.* 32: 100691.

150 Chen, Y., Liu, X., Anderson, J.Y.L. et al. (2022). A genome-scale nutrient minimization forecast algorithm for controlling essential amino acid levels in CHO cell cultures. *Biotechnol. Bioeng.* 119: 435–451.

151 Schinn, S.M., Morrison, C., Wei, W. et al. (2021). A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese hamster ovary cell cultures. *Biotechnol. Bioeng.* 118: 2118–2123.

152 Antonakoudis, A., Strain, B., Barbosa, R. et al. (2021). Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells. *Comput. Chem. Eng.* 154: 107471.