Trends in Biotechnology

CellPress

Feature Review

Flux sampling and context-specific genome-scale metabolic models for biotechnological applications

Devlin C. Moyer[1,2], Justin Reimertz[1], Juan I. Fuxman Bass[1,2,3,*], and Daniel Segrè[1,2,3,4,5,*]

Genome-scale metabolic models are used in fields ranging from metabolic engineering to drug discovery and microbiome design. Although these models are often used to predict putatively optimal states, some applications, including modeling human tissues for drug development and microbial communities for synthetic ecology, may require sampling the whole space of feasible fluxes to obtain distributions of biologically relevant states. Additionally, many applications involve using transcriptomic or proteomic data to predict fluxes for specific tissues, diseases, or patients. We revisit different methods used toward these goals and focus on their limitations and challenges, providing guidelines on how to avoid some of the shortcomings of existing approaches and highlighting conceptual barriers that will require new methodologies and offer opportunities for future development.

Highlights

Genome-scale metabolic modeling is a growing area of computational biology, rich in biotechnology applications, including the study of human metabolism for drug development and the design of synthetic microbial communities for health environmental and engineering purposes.

Incorporating omics data into genome-scale metabolic models is an important avenue for improved predictive accuracy. We revisit and categorize the major challenges that still limit the applicability of these approaches, pointing to opportunities for future research.

Predicting distributions of all possible fluxes, rather than optimal flux vectors, is a valuable and underused approach for incorporating uncertainty and capturing phenotypic diversity of metabolic states. Multiple tools are available for generating these distributions, but special care must be taken to obtain meaningful results.

Genome-scale metabolic models

Understanding cellular metabolism is critical to studying multiple biological processes, including disease etiologies and treatments [1,2], microbial community organization and dynamics [1,3], and metabolic rewiring caused by environmental and genetic perturbations [2]. These phenomena are fundamental to several biotechnological pursuits, including drug discovery [4], metabolic engineering of microbes for production of commercially or medically valuable compounds [5], and the design or control of synthetic microbial communities for biomedical or environmental applications [1,6,7]. Yet, untangling the complex interdependence between the thousands of reactions present in a given organism constitutes an ongoing challenge. This is partially due to the fact that direct measurements of enzyme expression, metabolite abundances and reaction rates [or **fluxes** (see Glossary)] are generally time-consuming, expensive, and only helpful for quantifying a small portion of the enzymes, metabolites, and reactions in any given organism or cell type [1,8,9]. **Genome-scale metabolic models (GSMMs)**, mathematical representations of the networks of all biochemical reactions known to occur in a given cell type or organism, have been used alongside direct measurements of metabolism to help interpret those measurements and to predict all metabolic fluxes in a cell, including those that are difficult to measure directly [10]. By using algorithms such as **flux balance analysis (FBA)** [10], GSMMs have been applied in multiple contexts, including predicting the impacts of gene knockouts to identify novel drug targets [4,11], guiding metabolic engineering efforts to generate genetically modified microbes that produce commercially and/or medically valuable compounds [5,12], simulating metabolic interactions between different cells within microbial communities [7,10,13] or multicellular organisms [14], and investigating fundamental questions about the evolution of metabolism [15].

[1]Bioinformatics Program, Faculty of Computing and Data Science, Boston University, Boston, MA 02215, USA
[2]Department of Biology, Boston University, Boston, MA 02215, USA
[3]Biological Design Center, Boston University, Boston, MA 02215, USA
[4]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
[5]Department of Physics, Boston University, Boston, MA 02215, USA

*Correspondence:
fuxman@bu.edu (J.I. Fuxman Bass) and
dsegre@bu.edu (D. Segrè).

Most applications of GSMMs involve predicting fluxes through all reactions in the GSMM of interest [10]. In particular, in order to make it computationally feasible to predict fluxes through all reactions in a cell without needing kinetic parameters, the fluxes predicted from GSMMs are generally steady-state fluxes [i.e., fluxes that satisfy mass balance (or flux balance) constraints], where the total flux consuming each metabolite is equal to the total flux producing that metabolite [10]. In addition to the steady-state constraints, it is also common to constrain the reversibility of reactions and the uptake of nutrients to reflect the molecular composition of the environment. In general, multiple vectors of fluxes (where each vector encodes one flux for each reaction in the GSMM) satisfy all of these constraints, collectively constituting the feasible space of a GSMM [16,17]. Although it is common to refer to a single vector of steady-state fluxes as a 'distribution of fluxes' through a GSMM, in the present work, we only use the phrase 'distribution of fluxes' to refer to the set of all possible steady-state fluxes that a reaction within a GSMM can sustain. Another common practice, especially in the context of metabolic engineering, is to try to identify a single vector of steady-state fluxes that is 'optimal' [i.e., it leads to the maximization (or minimization) of a given linear combination of fluxes **(objective function)**], usually the cellular steady-state growth rate [5,10,18]. Notably, in most GSMMs, even the search for steady-state fluxes associated with maximal growth rate typically yields multiple equivalently optimal solutions **(alternative optima)** [16,18]. Many different reactions (and linear combinations of reactions) have been used as objective functions, but few manage to single out a unique optimal set of fluxes [16]. GSMMs typically include also a collection of fictitious reactions that act as sources of specific individual metabolites, representing availability and uptake of nutrients from the extracellular environment [6]. Most cells are in principle capable of consuming a variety of nutrients that may not be present in a given environment. It is therefore critical to ensure that the constraints on fluxes through nutrient uptake reactions are accurately reflecting the availability of environmental molecules before attempting to predict fluxes through the rest of the GSMM.

Although GSMMs have been used successfully in multiple contexts, several challenges limit their applicability and accuracy beyond relatively simple and well-studied scenarios. Here, we review some of the most significant challenges and existing efforts to address them. Note that reaching high accuracy in the inference of a particular GSMM from the corresponding genome constitutes a significant challenge of its own, which has been thoroughly discussed elsewhere [10] and is not discussed here. The first challenge we address is appropriately integrating transcriptomic, proteomic, or other omic data into a GSMM in order to accurately predict metabolic fluxes in specific cell types and conditions [10,19]. It has been shown that in several instances, fluxes predicted using omic integration methods display accuracies comparable with predictions made without incorporating context-specific omic data at all, for reasons that remain unclear [20,21]. The second challenge is obtaining biologically meaningful results from algorithms designed to predict the distributions of all possible fluxes through each reaction in a GSMM. As mentioned above, most reactions in most GSMMs are capable of sustaining a whole distribution of steady-state fluxes, even after imposing objective functions. Predicting these distributions can provide insight into the precision/uncertainty of the predicted fluxes through each reaction and may even be a more realistic reflection of the metabolic adaptability and phenotypic diversity of real cells [16]. Algorithms capable of predicting these distributions can be difficult to use and easy to misuse and may produce outputs that are hard to interpret [16]. This review discusses the details of these challenges with using GSMMs, provides a critical assessment of the performance of existing approaches, and discusses how various choices made when applying GSMMs to different biotechnological applications (Box 1) affect the quality and utility of their predictions, potentially informing future improvement efforts.

Glossary

Allosteric regulation: modulation of enzyme activity induced by the binding of a small molecule to the enzyme at a location other than its active site.
Alternative optima: vectors of steady-state fluxes that are all equivalently optimal with respect to a given objective function (e.g., maximum growth rate).
Catalytic activity: the capacity of an enzyme to catalyze one or more specific chemical reactions.
Catalytic rate: the maximum rate at which an enzyme can transform substrate(s) to product(s), often denoted as k_{cat}.
Convergence: for an MCMC algorithm, convergence is achieved when distributions of possible flux solutions from different sampling chains are all approximately identical to each other.
Discretize: convert a continuous value into a discrete/categorical value.
Flux: the rate of a reaction, generally expressed in millimoles per gram of dry weight of biomass per hour. In FBA and other stoichiometric modeling approaches, fluxes are usually assumed to be at steady state.
Flux balance analysis (FBA): an approach for predicting possible steady-state fluxes for all reactions in a metabolic network (from a GSMM) based on the stoichiometric coefficients of the reactions and an objective function (see below), generally under the assumption that metabolism operates close to an optimum.
Flux sampling: using an MCMC algorithm to generate a representative subset of solutions from the set of all possible solutions to a GSMM. These algorithms provide a distribution of possible fluxes through each reaction, unlike FBA, which provides a single optimal flux for each reaction.
Flux variability analysis (FVA): a variant of FBA in which the minimum and maximum steady-state fluxes sustainable by each reaction are calculated.
Genome-scale metabolic model (GSMM): a formal representation of the network of all metabolic reactions that occur in a particular organism or cell. The core information stored in a GSMM is the stoichiometric matrix, whose rows correspond to metabolites and columns to reactions, and each entry corresponds to the stoichiometric

Creating context-specific GSMMs

GSMMs for multicellular organisms generally represent the metabolism of generic cells of the target organism rather than the metabolism of any particular cell type, organ, or tissue [22]. Similarly, some GSMMs for microbial organisms represent the combined metabolic capacities of all strains of that organism rather than just a single strain (e.g., *Saccharomyces cerevisiae* [23]). Several methods have been developed to use omic data to 'extract' GSMMs of specific strains or cell types, termed 'context-specific' GSMMs (Figure 1A) [10]. Context-specific models serve two major purposes. First, they constitute repositories of knowledge that integrate genomic information (all possible reactions encoded in that genome) with developmental stage or environmental knowledge (i.e., the context), which is typically derived from a readout of gene expression information. This knowledge can be used to qualitatively assess the metabolic capabilities of a given cell type in a way that otherwise would be very challenging. Second, by using these context-specific models as the starting point for FBA predictions, one can generate quantitative predictions of all cellular fluxes in specific tissues, cells, or conditions. This approach can be especially relevant when attempting to identify novel drug targets using GSMMs of diseased cells or pathogens or simulating metabolic interactions between different cell types within a community, tissue, biofilm, or tumor [6,10,14]. Most methods for creating context-specific GSMMs primarily use transcriptomic or proteomic data from the strain or cell type of interest to limit the maximum flux allowed through reactions catalyzed by lowly expressed enzymes (Figure 2). It has been shown through specific examples that many of these methods may produce context-specific GSMMs whose predicted fluxes do not match experimental data better than predictions obtained from generic GSMMs; yet, no specific feature or combination of features of any method has been conclusively determined to be responsible for the limited accuracy of context-specific model predictions [20,21]. In this section, we review the assumptions underlying many of these methods,

coefficient of that row's metabolite in that column's reaction.

Markov chain Monte Carlo (MCMC): a class of algorithms for approximating the distribution of all possible solutions to a system of equations. Each solution is ideally computed independently of the previously computed solution.

Objective function: a linear combination of reaction fluxes in a GSMM, which is assumed to be maximized (or minimized) by the cell. A common objective function used in FBA is the maximization of the biomass production flux.

Sampling chain: a collection of solutions derived from a single execution of an MCMC algorithm applied to a specific system of equations.

Thermodynamically infeasible cycle: a loop of reactions where each reaction has a nonzero steady-state flux, but there is zero overall net production or consumption of metabolites. Because nonzero flux through each reaction is thermodynamically possible only if that reaction has a negative free energy change, such a loop must have a nonzero net free energy change, which is thermodynamically possible only if there is a net production or consumption of metabolites.

Box 1. Biotechnological applications of GSMMs

Metabolic engineering

GSMMs can serve useful roles at several stages of different kinds of metabolic engineering projects [5]. They can be used to facilitate efforts to create novel microbial strains that efficiently produce specific compounds of interest by simulating the impacts of various knockouts on the production of the target compound (Figure IA) quickly and inexpensively before generating real-life strains with those knockouts [5]. Multiple tools also leverage GSMMs to predict a minimal set of exogenous enzymes to add to a given organism to enable production of a given compound [5]. GSMMs of strains that have already been engineered can help identify further manipulations (i.e., additional exogenous enzymes or knockouts) that would increase production of the target compound and help explain counterintuitive phenotypes of such strains (e.g., knockouts that were expected to improve production of the target compound that had no measurable effect) [137]. GSMMs can also be used to identify essential nutrients for particular organisms, which can inform the design of chemically defined minimal media, as well as media that are optimized to sustain production of specific compounds while maintaining a given growth rate [138].

Drug targeting

GSMMs of pathogenic cells (e.g., cells in tumors or parasitic organisms) can be used to identify potential drug targets by predicting which knockouts are likely to significantly disrupt the metabolism of those cells, especially their ability to sustain growth (Figure IB) [4]. These disruptions can, in principle, include both reducing the production of essential metabolites and increasing the production of toxic metabolites. In combination with GSMMs of healthy cells, one can also predict the selectivity of (potential) drugs that target metabolic enzymes by simulating their impact on the metabolism of healthy cells [4].

Community engineering

GSMMs of multiple different cell types can be combined into models of metabolic interactions within communities of different cell types, including microbial communities [7,13], interactions between host cells and microbiomes [13], tumor microenvironments [85], and different tissues of multicellular organisms [14] (Figure IC). Such models can be used to predict the consequences of perturbations to these communities, such as the impact that adding a particular drug or metabolite will have on the relative abundance of each cell type or the impact of adding or removing certain members of synthetic communities on the metabolic phenotypes of the other members [7].

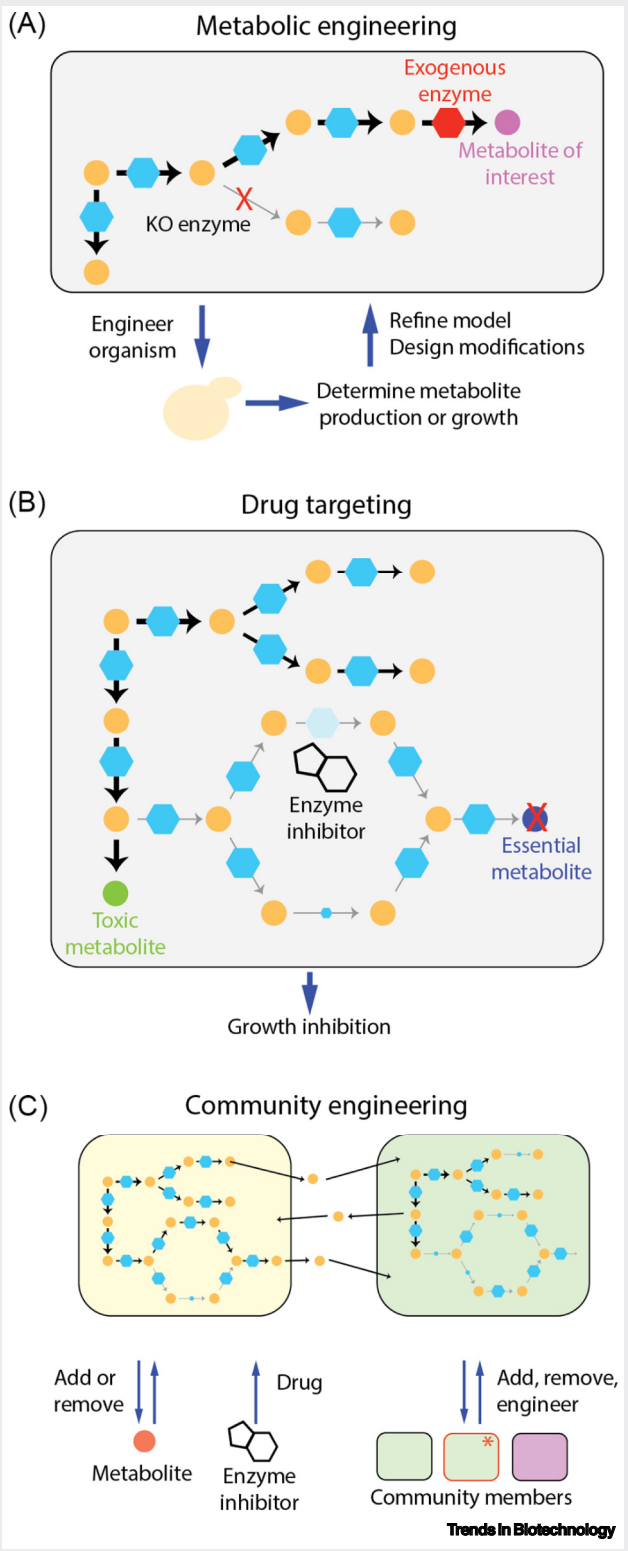

Figure I. Biotechnological applications of genome-scale metabolic models. Illustrations of how genome-scale metabolic models can be applied to (A) metabolic engineering projects focused on increasing production of particular metabolites of interest, (B) identifying potential metabolic drug targets, and (C) engineering metabolic interactions within communities of cells. See box text for more details. Abbreviation: KO, knockout.

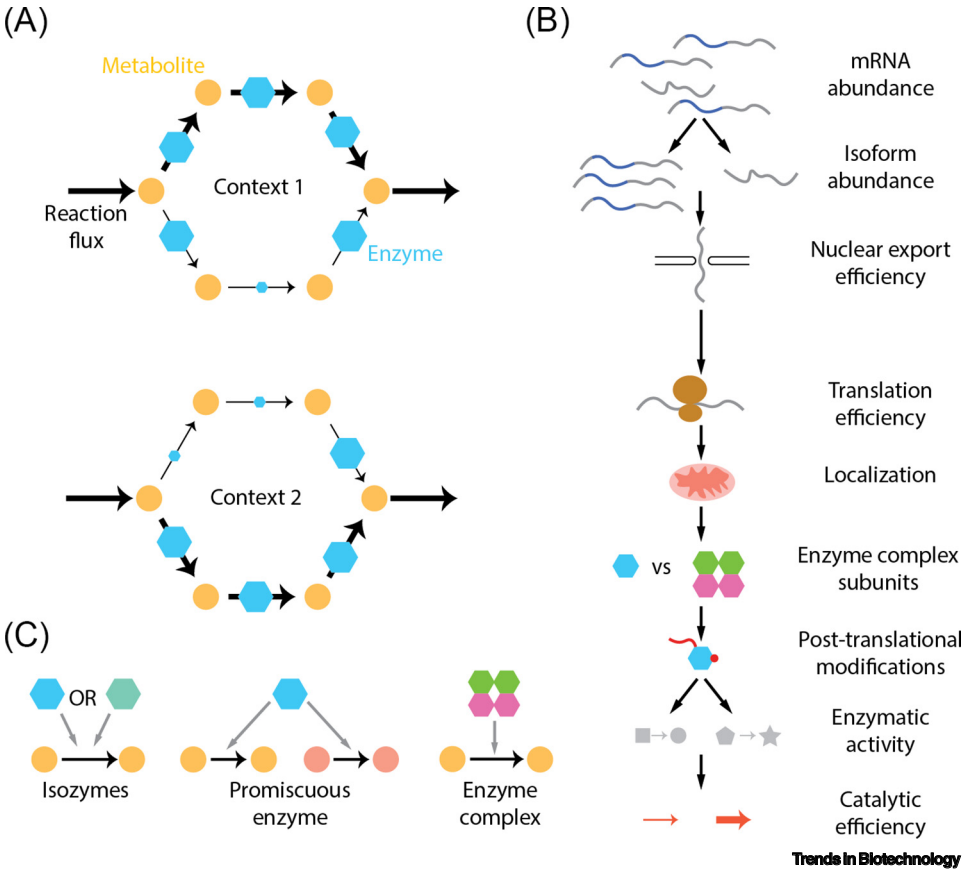

Figure 1. Factors relevant to creating context-specific genome-scale metabolic models (GSMMs). (A) Schematic representations of two different GSMMs representing two cells capable of catalyzing the same set of reactions, but with different fluxes through each reaction because of differences in the abundances of the corresponding enzymes in each cell. Circles represent metabolites, arrows represent metabolic reactions, hexagons over arrows represent metabolic enzymes, the size of each arrow represents the flux through the reaction, and the size of each hexagon represents the abundance of the enzyme. Note that some reactions with low fluxes are associated with highly abundant enzymes because flux through an upstream reaction was limited by a lowly abundant enzyme. (B) The effective abundance of enzymes available to catalyze a particular reaction can differ from the total abundance of that enzyme in the cell and the abundance of the mRNA encoding that enzyme. Alternative splicing can lead to different mRNA isoforms that are exported from nuclei with different efficiencies and translated into different numbers of proteins per mRNA. Alternative splicing can also affect the subcellular localization of the corresponding protein, as well as which post-translational modifications it receives, both of which may also depend on the regulatory state of the cell as a whole and can alter the catalytic activities and rates of each protein isoform. Each individual protein may only be catalytically active as part of an enzyme complex comprised of multiple copies of that protein and/or proteins derived from different genes. Some enzymes may be catalytically active both as monomers and as subunits of enzyme complexes but may have different catalytic activities or rates as monomers than they do as subunits. (C) Examples of one-to-many and many-to-one mappings between enzymes and reactions. One reaction may be capable of being independently catalyzed by multiple different enzymes (isozymes), one enzyme may be capable of catalyzing multiple different reactions (promiscuous enzymes), and some reactions are catalyzed by complexes of multiple separate protein subunits encoded by different genes.

assess their biological plausibility, and identify the most biologically plausible approach and inherent limitations for each step in the process of creating a context-specific GSMM.

Measuring or estimating enzyme abundance

Because enzyme abundance varies significantly between cell types or strains and can constrain metabolic fluxes, most methods for creating context-specific GSMMs focus on incorporating

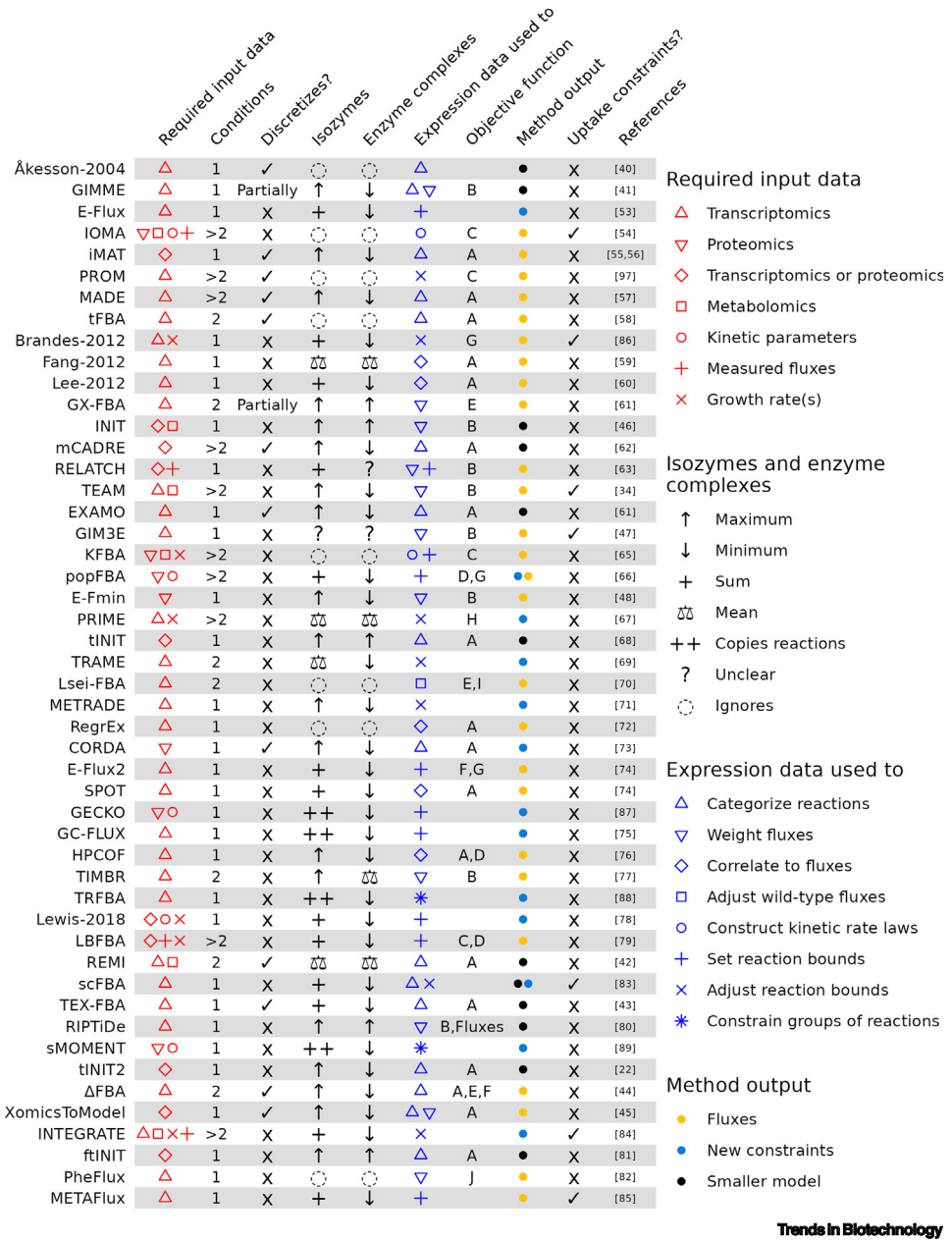

Figure 2. Characteristics of methods for creating context-specific genome-scale metabolic models (GSMMs). Methods are arranged chronologically by the date of publication of the first paper about the method (i.e., the first reference in the 'References' column), with the oldest method in the uppermost row. Methods with multiple shapes in the 'Required input data' column require all of the indicated types of input data. Methods with a '1' in the 'Conditions' column can create context-specific GSMMs with a single set of the required input data, whereas methods with a '2' or '>2' require multiple sets of each type of required input data, each gathered in a different condition. A '✓' in the 'Discretizes?' or 'Uptake constraints?' column indicates that the corresponding method discretizes the input data or constrains nutrient uptake fluxes, respectively. An 'x' in the 'Discretizes?' or 'Uptake constraints?' column indicates that the corresponding method does not discretize the input data or constrain nutrient uptake fluxes, respectively. The 'Isozyme' and 'Enzyme complex' columns indicate how methods assign expression levels to reactions catalyzed by multiple isozymes or enzyme complexes composed of multiple subunits: using the maximum, minimum, sum, or mean expression level of all isozymes or complex subunits or by creating copies of reactions

(Figure legend continued at the bottom of the next page.)

enzyme abundance data. However, proteomic approaches for quantifying enzyme abundance are expensive, time-consuming, and often produce accurate measurements for only a fraction of the enzymes present [9]. In contrast, transcriptomic techniques are cheaper, faster, and capable of precisely measuring the abundance of nearly all mRNAs. Consequently, transcriptomic data are generally much more available than proteomic data, and most methods for creating context-specific GSMMs use mRNA abundance as a proxy for enzyme abundance (Figure 2).

Although there is some correlation between transcript and protein abundance [24], several important biological phenomena contribute to discrepancies in the two values (Figure 1B) [20,25]. Some transcripts are translated into proteins thousands of times more efficiently than others [25,26], and alternative splicing of the same pre-mRNA can affect the likelihood of nuclear retention/export of the mature mRNA [27], as well as the enzymatic activity and/or subcellular localization of enzymes translated from it [28–30] and the likelihood of post-translational modifications (such as acetylation and glycosylation) that can affect enzymatic activity [31–33]. When using GSMMs to predict how fluxes change over time (as opposed to predicting steady-state fluxes) [7,34], it is also important to consider that the aforementioned processes may take meaningfully different amounts of time for different genes. Various tools have been developed to quantify the relative abundances of alternative mRNA isoforms [35]; predict nuclear export [27] and translation efficiencies of arbitrary mRNAs [26]; and predict the enzymatic activity [36], subcellular localization [37], and probable post-translational modifications [38] of the corresponding proteins. In principle, such tools could be leveraged to mitigate the problems with using transcript abundances as proxies for enzyme abundances; yet, few methods for creating context-specific GSMMs employ such tools [39].

Several methods for tailoring generic GSMMs to specific contexts involve discretizing enzyme or transcript abundances to designate certain reactions as 'inactive,' 'lowly expressed,' or 'highly expressed' (Figure 2). This has occasionally been explained as a strategy for mitigating noise [20] or various other limitations of microarray data [40,41], but it is unclear why more recent methods developed to use RNA-sequencing data involve discretization (Figure 2) [42–45]. Many methods that do not **discretize** expression data transform or normalize expression levels (e.g., by using logarithms of transcript abundances [46,47] or by normalizing all expression levels to the highest expression level [47,48]). Although these transformations may reflect standard practice in expression data analysis, it is not clear that they represent biologically grounded ways of encoding the effect of mRNA abundance on fluxes.

In summary, predicted fluxes from context-specific GSMMs are generally more likely to be biologically plausible when made with continuous and not discretized expression levels; proteomic rather than transcriptomic data whenever possible; and/or tools such as GeTPRA [39], rMATS-

catalyzed by isozymes and associating each isozyme's expression level with a different copy. A value of 'Ignores' in the 'Isozymes' or 'Enzyme complex' column signifies that the paper(s) that introduced the corresponding method did not mention if or how the method handled expression levels of isozymes or enzyme complex subunits, whereas a value of 'Unclear' signifies that the paper(s) mentioned that the method accounts for isozymes or enzyme complexes but limited or no explanation is provided. Methods with multiple shapes in the 'Expression data used to' column use expression data in all of the indicated ways. Objective function abbreviations: 'A' = maximize agreement (which includes but is not limited to correlation) between predicted fluxes and expression data; 'B' = maximize or minimize the weighted sum of all predicted fluxes; 'C' = minimize extent to which predicted fluxes violate their bounds; 'D' = minimize the L1-norm of all predicted fluxes; 'E' = minimize the total change in all predicted fluxes between conditions; 'F' = minimize the L2-norm of all predicted fluxes; 'G' = maximize the predicted flux through the biomass reaction; 'H' = minimize the change in predicted flux through the biomass reaction between conditions; 'I' = maximize the total predicted production of ATP; 'J' =

(See figure legend at the bottom of the next page.)

turbo [35], MEW [26], and BUSCA [37], etc. [36,38], to account for alternative splicing, variable translation efficiency, etc. if transcriptomic data are used in place of proteomic data. As discussed below, an accurate estimate of enzyme abundance is only one of the pieces of information necessary to make biologically plausible predictions about the magnitude of metabolic fluxes.

Mapping enzyme abundance to reactions

Another major obstacle to incorporating expression data into GSMMs is determining how to map enzyme expression levels to reactions, because many metabolic reactions can be independently catalyzed by multiple enzymes (isozymes), some can be catalyzed only by complexes of multiple protein subunits encoded by different genes, and some enzymes can catalyze multiple different reactions (multifunctional enzymes) (Figure 1C) [49–52]. Existing methods for creating context-specific GSMMs have taken different approaches to mapping expression levels to reactions (Figure 2). Most methods associate a single expression level with each reaction [34,40–48,53–86], which requires integrating multiple expression levels into a single value for all reactions with isozymes or those catalyzed by enzyme complexes.

The most common approach to integrating the expression levels of isozymes is to use only the expression level of the most expressed isozyme for each reaction (Figure 2) [34,41,44,48,55–57,62,64,71,73,76,77]. A less popular alternative approach is to make copies of all reactions associated with isozymes, one copy per isozyme, and associate each copy with the expression level of a different isozyme [75,87–89]. Although the biological plausibility of both approaches is comparable, different enzymes catalyze their reactions at different rates, so many copies of a slow enzyme can have the same effective catalytic capacity as few copies of a fast enzyme. Because the catalytic rates of different enzymes can vary by several orders of magnitude [25], the biological plausibility of the expression levels assigned to each reaction, especially reactions catalyzed by isozymes, depends strongly on whether the expression levels are weighted by catalytic rates.

The most common approach to deriving an expression level for an enzyme complex is to use the minimum expression level of all core subunits (Figure 2) [34,41,43–45,48,53,55–57,60,62,64,66,69,71,73–76,78,79,83–89]. This may accurately represent enzyme complexes that require all of their core subunits to have **catalytic activity** [90], but it misrepresents heteromeric complexes with different numbers of each subunit present within each complex [91,92]. For example, the human pyruvate dehydrogenase complex generally has three E2 subunits for each E3 subunit [93], so the number of complete complexes could be limited by the expression level of the E2 subunit even if the expression level of the E3 subunit was lower. GSMMs generally indicate which reactions are catalyzed by enzyme complexes using gene-protein-reaction (GPR) rules, which are strings of gene names (or symbols or other identifiers) separated by 'and' when they are subunits of the same complex or 'or' when they are isozymes [91]. One approach to incorporating the stoichiometry of enzyme complex subunits is to extend GPRs to also include copy numbers, such as 'A*1 and B*2' (representing a complex of one copy of subunit A and two copies of subunit B), so that one can divide the expression level of each enzyme associated with a particular reaction by its copy number before determining the minimum expression level to assign to the reaction [94]. Another approach is to add each enzyme to the reaction(s) it catalyzes as a 'reactant' and use its subunit copy number as its stoichiometric coefficient in the reaction, then add a reaction that creates each enzyme 'metabolite' whose maximum flux is set with that enzyme's expression level (usually divided by its **catalytic rate**) [87,91,95]. Neither approach is necessarily more or less biologically plausible than the other, but the second approach also addresses potential

issues with the representation of isozymes as described in the previous paragraph, whereas the first approach addresses only enzyme complexes.

Most methods for creating context-specific GSMMs constrain all reaction fluxes associated with each multifunctional enzyme separately, ignoring the fact that they share a limited pool of that enzyme (Figure 2). This approach is at odds with the general principle behind using enzyme abundance to influence the predicted fluxes through reactions in GSMMs: the amount of enzyme available to catalyze each reaction can limit how much flux it can sustain. The few methods that account for multifunctional enzymes do so by adding each enzyme to the reactions it catalyzes as a 'reactant' and limiting the availability of each enzyme 'reactant' with its expression level [87–89,91]. Although this enables a more accurate representation of multifunctional enzymes, these methods create copies of all reactions with isozymes in order to add a different enzyme 'reactant' to each copy, and the total number of reactions in a GSMM significantly influences the computational resources required to perform downstream analyses.

Altogether, a biologically plausible way to address the many-to-many mapping between enzymes and reactions is to account for the existence of isozymes and multifunctional enzymes by creating copies of reactions associated with isozymes, as done in GC-Flux [75], GECKO [87], and sMOMENT [89]; constrain all groups of reactions that can be catalyzed by the same enzyme together, as done in GECKO [87], sMOMENT [89], and TRFBA [88]; and use databases such as the Complex Portal [92] to account for enzyme complex subunit stoichiometry [87,94].

The relationship between enzyme abundance and reaction fluxes

Although the abundance of the enzyme(s) that can catalyze a particular reaction is one of the factors limiting its maximum possible flux, it is not the only relevant factor. The relative concentrations and chemical potentials of the products and reactants also significantly influence reaction fluxes – the abundance of an enzyme has no influence on the flux through a reaction if its substrates are not present [96]. Furthermore, the extent to which changes in enzyme abundance influence reaction fluxes depends on each enzyme's catalytic rate, which can vary by seven orders of magnitude [25]. However, few methods for incorporating expression data into GSMMs also incorporate catalytic rates (e.g., by weighting the abundances of enzymes with their catalytic rates) [54,65,66,78,87], metabolite concentrations [47,65], or any thermodynamic parameters [42,43]. Many methods that ignore metabolite concentrations and catalytic rates also force predicted fluxes to correlate to enzyme abundances as much as possible, an assumption that oversimplifies the implications of Michaelis-Menten kinetics (Figure 2) [34,40–42,44–47,55–64,68,70,72–74,76,77,80–82,97].

Many methods assume that a reaction cannot carry flux if none of the associated enzymes appear to be expressed in the given expression dataset [40–42,46,53,55–57,62,64,68,80,81,83,85,87]. This neglects the possibility of false negatives, which is especially concerning in light of the fact that many housekeeping genes – genes that are constitutively expressed at similar levels in most or all contexts – are known to be expressed at relatively low levels [98]. In addition, recent papers have shown that particular enzymes are capable of catalyzing many more reactions than they were initially known to [51,99]. Thus, permanently blocking the flux through a reaction in a GSMM because none of the enzymes it is currently associated with are expressed may artificially rule out the possible role of yet to be uncovered secondary catalytic activities of other enzymes. Furthermore, some reactions can occur at non-negligible rates in the absence of any enzymes, such as the (de)hydration of carbonic acid [100], so completely blocking flux through such reactions just because the enzymes that can catalyze them are absent is risky. Blocking fluxes through reactions associated with enzymes that do not appear to be expressed has been suggested to be a contributing factor to the low accuracy of predicted fluxes made by context-specific GSMMs produced

with existing methods [20]. One suggested way to avoid these issues is to allow all reactions in a GSMM to sustain some relatively small amount of flux, regardless of enzyme abundances [83,86]. Note that the magnitudes of fluxes predicted through reactions in GSMMs can vary by several orders of magnitude, so this minimal flux should be treated as a parameter of the model whose value should be carefully optimized.

A reason why most methods do not incorporate catalytic rates is that relatively few enzymes in any single organism have had their catalytic rates measured. Although experimentally measuring catalytic rates has proved difficult to scale, several recent machine learning methods have been developed that can accurately predict the catalytic rates of most enzymes [101,102]. Some of these methods [101] can also predict secondary catalytic activities, so incorporating them into future methods for constraining GSMMs with omic data may help address multiple obstacles to predicting accurate fluxes.

Allosteric regulation of enzymes by metabolites is known to have significant influence on metabolic fluxes. For multiple reasons, however, allosteric effects are difficult to represent in GSMMs and to incorporate into algorithms for flux predictions [96,103]. First, the basic information about which metabolites regulate which enzymes is available for only relatively few highly studied organisms [65], and the experiments necessary to measure such interactions in other organisms require significant investments of time and resources [103]. Even with the knowledge of which allosteric interactions occur in the cell(s) of interest, modeling the impact they have on metabolic fluxes requires experimentally measured enzyme concentrations and either metabolite concentrations or thermodynamic parameters for all reactions [65]. Furthermore, accurately representing the relationship between the concentrations of allosteric regulators and the fluxes through reactions catalyzed by the enzymes they regulate involves kinetic parameters that have not been measured and are difficult to measure for most interactions in most organisms. This relationship is also challenging because predicting steady-state fluxes provides no information about steady-state metabolite concentrations. In addition to the challenges with obtaining the required input data, accounting for the frequently nonlinear relationships between the concentrations of allosteric regulators and the activities of the enzyme(s) they regulate significantly increases the computational complexity of flux prediction algorithms [65].

An additional important aspect of creating context-specific GSMMs is the choice of constraints on fluxes through nutrient uptake reactions. Relatively small changes in the composition of cell culture media have been observed to lead to significant changes in metabolic phenotypes of cultured cells [96]. Therefore, it is important to use as much information as possible about nutrient availability and uptake to constrain allowable fluxes through these nutrient uptake reactions, especially when simultaneously using expression data to constrain predicted fluxes [21,80,85]. Even though the precise chemical composition of culture media or other cellular environments (e.g., tumor microenvironments) is often unknown, GSMMs may contain uptake reactions for a variety of drugs or other xenobiotics that can safely be assumed to be absent in most contexts, so it is often possible to determine that at least some uptake reactions should be assumed to have zero flux. Relatively few existing methods for incorporating expression data into GSMMs also constrain uptake fluxes [34,47,54,66,83,84], which is potentially attributable to the fact that transporters are more likely to be misannotated than enzymes [104]. Ideally, one would either set all nutrient uptake fluxes to experimentally measured values [21] or set the bounds on predicted values using both measured concentrations of environmental metabolites and kinetic parameters of all relevant transporters [7], but even just preventing uptake of metabolites known to be unavailable in the condition of interest has been shown to improve prediction accuracy [80,85].

Predicting distributions of possible fluxes

As mentioned above, most reactions in most GSMMs under a given environmental condition are capable of sustaining a distribution of steady-state fluxes (rather than a single possible flux). From a mathematical perspective, this is a consequence of the fact that a typical FBA problem is underdetermined [i.e., the constraints imposed on the network (in the form of flux balance or upper or lower bound to specific fluxes) define a convex polyhedron of possible flux states (the feasible space), which are all, in principle, possible]. From a biological perspective, the goal of predicting a single set of fluxes to be compared with experimental data has historically represented the standard go-to approach to narrowing down possible states within the feasible space. However, researchers have increasingly started recognizing the possible biological relevance of the full set of possible fluxes compatible with available constraints and, in parallel, addressing the challenging mathematical question of how to obtain an accurate representation of this space. Although, through linear optimization, standard FBA typically reports a single vector of possible fluxes through all reactions in a GSMM at a time, we focus here on alternative algorithms, which either systematically explore the possible range of individual fluxes or use random sampling of points within this space (**flux sampling**) in order to characterize the shapes of these distributions of all possible fluxes (Figure 3A). It is important to note that, although objective functions are often presented as a necessary prerequisite to predicting fluxes through reactions in a GSMM [4,10,16], the alternatives shown in Figure 3 do not require users to specify an objective function. This makes these alternatives particularly useful in the numerous contexts in which the modeled cells lack a clearly defined metabolic objective, such as most cells in multicellular organisms [4] or bacteria growing in nutrient-poor or fluctuating environments [18]. In many cases, it is also possible to use such algorithms to characterize the space of alternative optima for a particular objective function [16,17]. These distributions of possible fluxes can also be used to quantify the uncertainty of the predicted flux through each reaction [16].

Flux variability analysis (FVA) [105] can be used to compute the minimum and maximum possible flux that each reaction can sustain (Figure 3A). However, each reaction's maximum and minimum flux is computed independently, neglecting flux–flux correlations. For example, it may not be possible for two reactions to simultaneously sustain the maximum fluxes that FVA predicts for each of them. Other algorithms that predict the entire distribution of possible fluxes (as opposed to just the ranges) through each reaction (Figure 3A) can offer more nuanced insights into the relationships between the fluxes through different reactions and the impacts of genetic or environmental perturbations [16,17]. Algorithms for computing these distributions generally involve much more sophisticated mathematics and computational resources than FVA (Table S1 in the supplemental information online) [16,106–108], especially when applied to large GSMMs, such as those of human cells or communities of multiple cell types [14,22]. In the context of metabolic engineering, such algorithms can also enable more accurate predictions of the range of phenotypes exhibited by cultures of engineered microbes than just predicting the theoretical maximum fluxes through individual reactions. Furthermore, these distributions of possible fluxes may accurately represent the metabolic versatility and adaptability of individual cells or the phenotypic heterogeneity of populations of cells, as opposed to merely representing noise or errors in the modeling approach [109]. Many methods for creating context-specific GSMMs are also compatible with algorithms for predicting distributions of possible fluxes and can help ensure that those distributions are biologically plausible.

Convergence diagnostics

Most algorithms for predicting distributions of possible fluxes through reactions in GSMMs are **Markov chain Monte Carlo (MCMC)** algorithms, which combine all reactions' distributions into a single multivariate distribution and iteratively generate samples from that distribution,

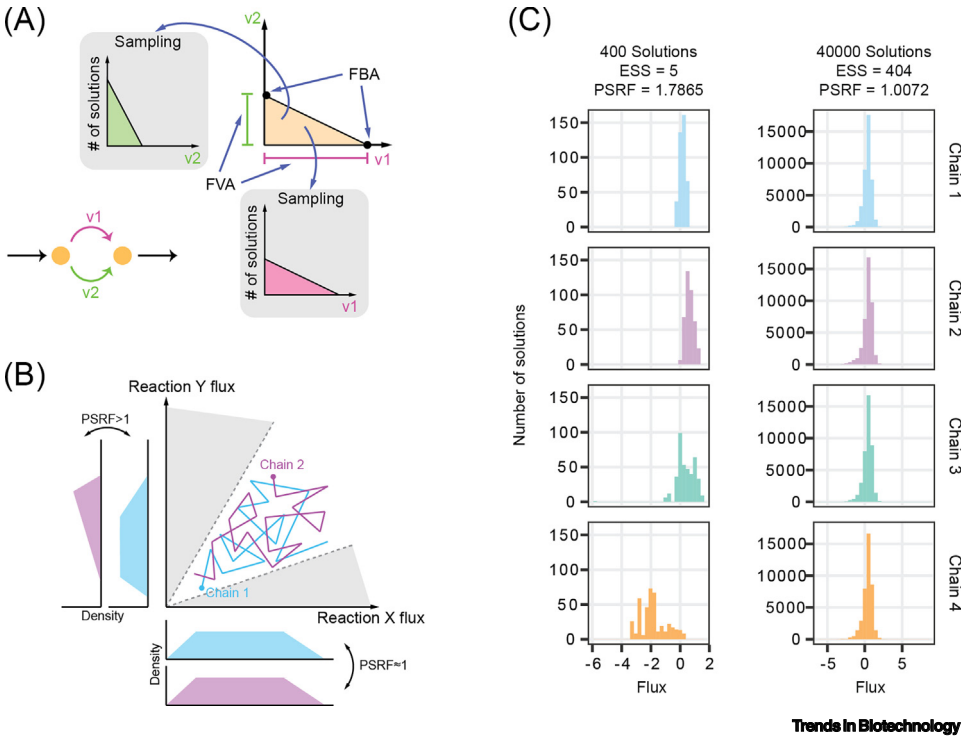

Figure 3. Determining distributions of all possible fluxes through genome-scale metabolic models (GSMMs). (A) Schematic illustrating how different algorithms for predicting fluxes from GSMMs work, specifically flux balance analysis (FBA), flux variability analysis (FVA), and sampling algorithms. Each point in the plot in the upper right represents a set of fluxes through the toy metabolic network in the lower left, where the *x*-axis represents the flux through reaction v_1 and the *y*-axis represents the flux through reaction v_2. The shaded area indicates the space of possible steady-state fluxes through the network, subject to a constraint on the maximum possible flux through one of the unlabeled reactions to the left or right of v_1 and v_2 in the toy metabolic network. FBA can predict the maximum or minimum possible flux through a linear combination of reactions, including the maximum or minimum possible flux through a single reaction (points at corners of the solution space labeled 'FBA'). By computing the maximum and minimum possible fluxes through all reactions, FVA predicts the ranges of possible fluxes through each reaction (pink and green lengths labeled 'FVA' along axes of upper right plot). By computing many points within the solution space, sampling algorithms can predict the full distributions (as opposed to just the ranges) of possible fluxes through each reaction (pink and green polygons in shaded boxes labeled 'Sampling' below and to the left of the upper right plot). Although the full solution space (upper right plot) of this toy model with only two reactions is easy to visualize, creating histograms of the sampled fluxes through each reaction (plots in shaded boxes labeled 'Sampling') is a useful way to approximately visualize the shapes of the many-dimensional solution spaces of GSMMs of real organisms. (B) Schematic illustrating how sampling algorithms predict distributions of all possible fluxes through reactions in a GSMM by computing a representative sample of all possible configurations of flux through the GSMM. The plot is analogous to the one in (A), representing the space of all possible configurations of fluxes through a GSMM (the feasible space), where the shaded areas represent configurations of fluxes that are not allowed by constraints on the fluxes through the reactions in the GSMM. The output of a sampling algorithm can be visualized as a 'chain' of connected points in the feasible space. A way to test if a particular sampling chain constitutes a representative subset of all possible configurations of fluxes through a GSMM is to run the same sampling algorithm on the same GSMM multiple times and use statistics such as the potential scale reduction factor (PSRF) to quantify the extent to which the different chains have converged to the same distribution of possible fluxes for an individual reaction. A rank-normalized PSRF greater than 1.01 may indicate that the chains have not converged [120]. The two sampling chains appear to have converged to the same marginal distribution of fluxes through reaction X but have not converged for reaction Y. (C) Distributions of possible fluxes through the D-lactate dehydrogenase reaction (LDH_D) in iML1515 [139] from four sampling chains generated by the Cobrapy [122] implementation of OptGP [111] with the default thinning factor of 100 and a thinned sample size of 40 000. The marginal distributions from the first 400 (thinned) samples from each chain are shown in the left column, and the right column shows the distributions from all 40 000 (thinned) samples. The rank-normalized effective sample size (ESS) and PSRF were computed separately for only the first 400 and all 40 000 samples produced by all four chains using the posterior R package [120]. The authors of the posterior R package, who also defined the rank-normalized ESS and PSRF statistics, recommend sampling until reaching an ESS of at least 400 and a PSRF of less than 1.01 [120].

where each sample is a vector containing a flux value for each reaction in the GSMM (Table S1 in the supplemental information online) [106,110–114] (although a few alternative algorithmic approaches have been used [108,115,116]). Specifically, most of these algorithms are hit-and-run algorithms that compute one sample, compute a direction and distance to travel to arrive at a different point in the multivariate distribution, and repeat (Figure 3B; Table S1 in the supplemental information online) [117]. Different algorithms take different approaches to determining where the next sample should be relative to the current one [106,112], transforming the multivariate distribution so that fewer iterations are required to generate a representative subset of samples [112,118], or biasing certain points as more or less likely to be sampled [114,119]. A key consideration when using MCMC algorithms is determining how many samples are 'enough' in order to constitute a representative subset of all possible samples. A common way to assess this is to run a particular algorithm on the same input multiple times to construct multiple separate '**sampling chains**,' then compare the distributions from the different chains and assess how similar they are (Figure 3C). A number of statistical approaches have been devised to measure the extent to which the different sampling chains have **converged** to the same distribution (which is assumed to be the underlying distribution the samples are being generated from), such as the potential scale reduction factor [120]. It is not generally possible to precisely predict how many samples will be necessary before a particular MCMC algorithm converges on a particular input, so directly assessing the extent to which the output of an MCMC algorithm has converged is critical [16,120].

Although implementations of newer algorithms such as CRHMC [106] and LooplessFluxSampler [113] automatically compute one or more **convergence** diagnostics, implementations of many earlier algorithms do not [110–112], leaving it up to each individual user to perform such tests. In particular, the implementations of ACHR [110], OptGP [111], and CHRR [112] available in the COBRA MATLAB toolbox [121] and/or the Cobrapy Python package [122], the two most widely used software packages for manipulating GSMMs, do not automatically report any measures of convergence along with their output. Many published papers that used MCMC algorithms on the solution spaces of GSMMs do not report any measures of convergence (Table 1), so it is unclear if their predicted distributions of fluxes are sufficiently representative of the GSMMs' solution spaces to provide conclusive biological interpretations (Figure 3C) [16].

Note that tests of convergence are generally performed separately on each marginal distribution (i.e., each individual reaction's distribution of sampled fluxes), and the distributions for some reactions may take many more iterations to converge than others. If the distribution for a particular reaction has converged but others have not, the converged distribution should maintain the same shape, mean, and so forth as additional samples are computed, so it is not strictly necessary for all reactions' distributions to converge before one can obtain meaningful results. However, computing additional samples after distributions appear to have converged is generally recommended to ensure that the distributions remain stable, because tests for convergence are technically tests for the presence of certain warning signs of nonconvergence, and passing one or all does not guarantee convergence [16,120].

Thermodynamically infeasible cycles

A major obstacle to obtaining biologically meaningful predicted fluxes in both regular FBA and sampling is the potential presence of cycles of reactions that can sustain thermodynamically infeasible fluxes (Figure S1 in the supplemental information online) [113,123]. Many techniques have been developed to avoid predicting meaningless fluxes through these **thermodynamically infeasible cycles** [113,123], including some of the aforementioned methods for incorporating context-specific omic data into GSMMs that incorporate chemical potentials of metabolites or standard Gibbs free energy changes of reactions [42,43,78].

Table 1. Papers that have used sampling algorithms to predict fluxes from GSMMs[a]

Application	Algorithm	Convergence assessment	Refs
Enzymopathies of glycolytic enzymes in human red blood cells	Rejection	Qualitatively judged stability of marginal distributions	[17,124]
Organization of fluxes throughout entire *Escherichia coli* metabolic network	HR	None	[125]
Human mitochondrial diseases	ACHR	Unclear	[110,126]
Central carbon metabolism in *E. coli*	HR	Qualitatively judged stability of marginal distributions	[125,127]
Metabolic impacts of knockouts in *Saccharomyces cerevisiae*	ACHR	None	[110,128]
Changing carbon sources and/or knocking out genes in *S. cerevisiae*	CB	Compared with experimentally measured fluxes	[129]
Gene knockouts in *E. coli*	OptGP	None	[65,111]
Adipocyte metabolism in lean and obese patients	CB	None	[129,130]
Metabolic division of labor between different *Arabidopsis thaliana* tissues	ACHR	None	[110,131]
Enzyme usage in wild-type and succinate-producing *E. coli*	ACHR	None	[91,110]
Endothelial cell metabolism in patients with sepsis who survived or died	OptGP	None	[132]
Redox metabolism in head and neck squamous cell carcinomas in smokers	OptGP	None	[78,111]
Central carbon metabolism before and after chilling *A. thaliana*	ACHR, OptGP, and CHRR	Raftery & Lewis, Geweke, and IPSRF[b] diagnostics	[16,110–112]
Phosphate depletion in *Streptomyces coelicolor*	CB	None	[129,133]
Metabolism of volatile organic compounds in strains of *S. cerevisiae* used for wine-making	OptGP	None	[111,134]
Metabolic heterogeneity of breast cancer	OptGP	None	[84,111]
Metabolism during metastasis of ovarian cancer	OptGP	None	[111,135]
Metabolic interactions between plant-associated *Pseudomonas* strains	OptGP	None	[111,136]
Metabolic interactions in human gut microbiome	CRHMC	None	[13,106]

[a]Papers are arranged in chronological order by date of publication.
[b]IPSRF, interval-based scale reduction factor.

One of the only algorithms for predicting distributions of all possible fluxes through GSMMs that avoids thermodynamically infeasible cycles is LooplessFluxSampler [113]. Notably, LooplessFluxSampler is comparable in speed to the fastest algorithms that do not avoid thermodynamically infeasible cycles [16,106–108,113]. Although LooplessFluxSampler is guaranteed to converge for all reactions only if they can simultaneously have nonzero fluxes without forming any thermodynamically infeasible cycles, which is generally only true of relatively small models of core metabolic pathways, few other algorithms for predicting distributions of possible fluxes are guaranteed to converge at all, yet still wind up converging for most GSMMs [113]. Overall, LooplessFluxSampler seems an excellent choice for predicting fluxes from GSMMs.

Concluding remarks

GSMMs are an approachable and efficient way of keeping track of the current knowledge about the biochemical pathways present in individual organisms and can serve as the starting point for

Outstanding questions

Predicting how transcript abundances relate to enzyme abundances and how enzyme abundances relate to fluxes at the genome scale is still challenging. Can these predictions be improved, such as by integrating additional data and combining mechanistic modeling with data-driven approaches?

Is it possible to robustly incorporate small-molecule regulatory effects on metabolic enzymes, such as allosteric regulation, into genome-scale metabolic models, despite the inability of steady-state models to predict concentrations of metabolites and enzymes? Is it possible to infer these regulatory effects from existing data without having to perform expensive and time-consuming experiments, or do new types of data need to be gathered?

To what extent will approaches for creating context-specific genome-scale metabolic models based on bulk omic data be applicable to single-cell data? Will fundamentally different approaches be required to obtain accurate predictions when using single-cell data?

Which biotechnological and biomedical applications of genome-scale metabolic models are more sensitive to errors in the structure of the model, which are more sensitive to the choices made when incorporating context-specific omic data, and which are more sensitive to the choice of algorithm used to predict fluxes?

predicting metabolic fluxes on the basis of fundamental mechanistic principles and simplifying assumptions. As the field has progressed, the interest in applying GSMMs to high-stakes applications, such as personalized medicine and microbiome engineering, have revealed both the power and the limitations of this approach. Other than the process of GSMM reconstruction itself, we believe that the two most challenging obstacles to using GSMMs as predictive tools, both rooted in the simplifying assumptions of steady-state metabolic models, are the difficulty of tailoring predicted fluxes to particular contexts of interest by incorporating measured enzyme, transcript, or metabolite abundances and the existence of multiple possible predicted fluxes for each reaction.

Predicting metabolic fluxes through an entire cellular metabolic network often involves making a series of assumptions, some of which may imply trade-offs between realism and computational feasibility. We discussed a number of key assumptions made by existing approaches to creating and predicting fluxes from GSMMs and identified cases where it should be possible to use more biologically plausible assumptions without sacrificing computational scalability. We also highlighted a number of problematic assumptions for which there are no apparent alternatives that are simultaneously more realistic and computationally tractable (see Outstanding questions). Although neither method addresses all of the issues we raise with existing approaches to predicting fluxes with GSMMs, version 3.0 of GECKO [87] and LooplessFluxSampler [113] collectively address most of the limitations we identify with other methods.

It is worth noting that the most biologically plausible approach we describe for accounting for isozymes and enzyme complexes when creating context-specific GSMMs significantly increases the total number of reactions in the GSMM, especially for metabolic models that simultaneously represent multiple different cell types and their interactions [6,10,14]. Attempts to extend GSMMs to also model allosteric regulation [65] may exacerbate this issue further. It is unclear if large models and long computation times are unavoidable for accurately modeling cellular metabolism or if an alternative modeling paradigm might scale more efficiently without compromising prediction accuracy. An additional complication with integrating multiple different types of data into a single GSMM is the difficulty of identifying a particular experimental condition for which all desired data types are available. It is possible that novel machine learning algorithms (along the lines of those mentioned above for predicting kinetic parameters [101,102]) may alleviate some of these issues with data availability. Even if all of the above issues are resolved for context-specific GSMMs constructed using bulk omic data, recent papers have identified additional obstacles with using single-cell omic data to create context-specific GSMMs, and it is unclear how well approaches designed for bulk data will translate to single-cell data (see Outstanding questions) [19,83,85]. Altogether, the limitations and challenges of current approaches presented in this review should be seen as opportunities for exploring extended and alternative methods, for rigorously assessing the biological/biochemical plausibility of modeling assumptions, and for developing novel strategies for creating context-specific models.

Acknowledgments

This work was partially supported by the National Institutes of Health grants R35 GM128625 awarded to J.I.F.B. and the National Cancer Institute (1R21CA279630-01) to D.S. and J.I.F.B. D.M. was supported by a bioinformatics NIH-funded predoctoral training fellowship (T32GM100842). D.S. further acknowledges funding by the Human Frontiers Science Program (grant number RGP0060/2021), the NIH National Institute on Aging award number UH2AG064704, the NSF Center for Chemical Currencies of a Microbial Planet (C-CoMP publication 073), and NSF-BSF grant 2246707.

Declaration of interests

The authors declare no competing interests.

Supplemental Information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tibtech.2025.07.010.

References

1. Kumar, R. *et al.* (2020) Single cell metabolomics: a future tool to unmask cellular heterogeneity and virus-host interaction in context of emerging viral diseases. *Front. Microbiol.* 11, 1152
2. Bergers, G. and Fendt, S.-M. (2021) The metabolism of cancer cells during metastasis. *Nat. Rev. Cancer* 21, 162–180
3. Dal Co, A. *et al.* (2023) Spatial self-organization of metabolism in microbial systems: a matter of enzymes and chemicals. *Cell Syst.* 14, 98–108
4. Nilsson, A. and Nielsen, J. (2017) Genome scale metabolic modeling of cancer. *Metab. Eng.* 43, 103–112
5. Otero-Muras, I. and Carbonell, P. (2021) Automated engineering of synthetic metabolic pathways for efficient biomanufacturing. *Metab. Eng.* 63, 61–80
6. Diener, C. and Gibbons, S.M. (2023) More is different: metabolic modeling of diverse microbial communities. *mSystems* 8, e0127022
7. Dukovski, I. *et al.* (2021) A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS). *Nat. Protoc.* 16, 5030–5082
8. Antoniewicz, M.R. (2021) A guide to metabolic flux analysis in metabolic engineering: methods, tools and applications. *Metab. Eng.* 63, 2–12
9. Taylor, M.J. *et al.* (2021) Spatially resolved mass spectrometry at the single cell: recent innovations in proteomics and metabolomics. *J. Am. Soc. Mass Spectrom.* 32, 872–894
10. Tarzi, C. *et al.* (2024) Emerging methods for genome-scale metabolic modeling of microbial communities. *Trends Endocrinol. Metab.* 35, 533–548
11. Yizhak, K. *et al.* (2015) Modeling cancer metabolism on a genome scale. *Mol. Syst. Biol.* 11, 817
12. Gleizer, S. *et al.* (2019) Conversion of *Escherichia coli* to generate all biomass carbon from CO_2. *Cell* 179, 1255–1263.e12
13. Gelbach, P.E. *et al.* (2024) Flux sampling in genome-scale metabolic modeling of microbial communities. *BMC Bioinformatics* 25, 45
14. Yilmaz, L.S. *et al.* (2020) Modeling tissue-relevant *Caenorhabditis elegans* metabolism at network, pathway, reaction, and metabolite levels. *Mol. Syst. Biol.* 16, e9649
15. Moyer, D. *et al.* (2021) Stoichiometric modeling of artificial string chemistries reveals constraints on metabolic network structure. *J. Mol. Evol.* 89, 472–483
16. Herrmann, H.A. *et al.* (2019) Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ Syst. Biol. Appl.* 5, 32
17. Price, N.D. *et al.* (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophys. J.* 87, 2172–2186
18. Feist, A.M. and Palsson, B.O. (2010) The biomass objective function. *Curr. Opin. Microbiol.* 13, 344–349
19. Hrovatin, K. *et al.* (2022) Toward modeling metabolic state from single-cell transcriptomics. *Mol. Metab.* 57, 101396
20. Machado, D. and Herrgård, M. (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10, e1003580
21. Bhadra-Lobo, S. *et al.* (2020) Assessment of transcriptomic constraint-based methods for central carbon flux inference. *PLoS One* 15, e0238689
22. Robinson, J.L. *et al.* (2020) An atlas of human metabolism. *Sci. Signal.* 13, eaaz1482
23. Zhang, C. *et al.* (2024) Yeast9: a consensus genome-scale metabolic model for *S. cerevisiae* curated by the community. *Mol. Syst. Biol.* 20, 1134–1150
24. Moulana, A. *et al.* (2018) Gene-specific predictability of protein levels from mRNA data in humans. *bioRxiv*, Published online August 24, 2018. https://doi.org/10.1101/399816
25. Hoppe, A. (2012) What mRNA abundances can tell us about metabolism. *Metabolites* 2, 614–631
26. Nieuwkoop, T. *et al.* (2023) Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* 51, 2363–2376
27. Zuckerman, B. and Ulitsky, I. (2019) Predictive models of subcellular localization of long RNAs. *RNA* 25, 557–572
28. Rouleau, M. *et al.* (2016) Divergent expression and metabolic functions of human glucuronosyltransferases through alternative splicing. *Cell Rep.* 17, 114–124
29. Kozlovski, I. *et al.* (2017) The role of RNA alternative splicing in regulating cancer metabolism. *Hum. Genet.* 136, 1113–1127
30. Lam, P.Y. *et al.* (2022) Alternative splicing and its roles in plant metabolism. *Int. J. Mol. Sci.* 23, 7355
31. Berg-Fussman, A. *et al.* (1993) Human acid beta-glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. *J. Biol. Chem.* 268, 14861–14866
32. Solá, R.J. *et al.* (2007) Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. *Cell. Mol. Life Sci.* 64, 2133–2152
33. Wang, Q. *et al.* (2010) Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* 327, 1004–1007
34. Collins, S.B. *et al.* (2012) Temporal expression-based analysis of metabolism. *PLoS Comput. Biol.* 8, e1002781
35. Wang, Y. *et al.* (2024) rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. *Nat. Protoc.* 19, 1083–1104
36. Watanabe, N. *et al.* (2020) Exploration and evaluation of machine learning-based models for predicting enzymatic reactions. *J. Chem. Inf. Model.* 60, 1833–1843
37. Savojardo, C. *et al.* (2018) BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466
38. Pakhrin, S.C. *et al.* (2021) DeepNGlyPred: a deep neural network-based approach for human N-linked glycosylation site prediction. *Molecules* 26, 7314
39. Ryu, J.Y. *et al.* (2017) Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism. *Proc. Natl. Acad. Sci. U. S. A.* 114, E9740–E9749
40. Akesson, M. *et al.* (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* 6, 285–293
41. Becker, S.A. and Palsson, B.O. (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4, e1000082
42. Pandey, V. *et al.* (2019) Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. *PLoS Comput. Biol.* 15, e1007036
43. Pandey, V. *et al.* (2019) TEX-FBA: a constraint-based method for integrating gene expression, thermodynamics, and metabolomics data into genome-scale metabolic models. *bioRxiv*, Published online January 31, 2019. https://doi.org/10.1101/536235
44. Ravi, S. and Gunawan, R. (2021) ΔFBA-predicting metabolic flux alterations using genome-scale metabolic models and differential transcriptomic data. *PLoS Comput. Biol.* 17, e1009589
45. Preciat, G. *et al.* (2021) XomicsToModel: omics data integration and generation of thermodynamically consistent metabolic models. *bioRxiv*, Published online December 19, 2022. https://doi.org/10.1101/2021.11.08.467803
46. Agren, R. *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.* 8, e1002518
47. Schmidt, B.J. *et al.* (2013) GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics* 29, 2900–2908
48. Song, H.-S. *et al.* (2014) Prediction of metabolic flux distribution from gene expression data based on the flux minimization principle. *PLoS One* 9, e112524
49. Richelle, A. *et al.* (2019) Assessing key decisions for transcriptomic data integration in biochemical networks. *PLoS Comput. Biol.* 15, e1007185
50. Jeanguenin, L. *et al.* (2010) Moonlighting glutamate formiminotransferases can functionally replace 5-

formyltetrahydrofolate cycloligase. *J. Biol. Chem.* 285, 41557–41566
51. Zmich, A. *et al.* (2023) Multiplexed assessment of promiscuous non-canonical amino acid synthase activity in a pyridoxal phosphate-dependent protein family. *ACS Catal.* 13, 11644–11655
52. Gu, C. *et al.* (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 121
53. Colijn, C. *et al.* (2009) Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* 5, e1000489
54. Yizhak, K. *et al.* (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* 26, i255–i260
55. Shlomi, T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* 26, 1003–1010
56. Zur, H. *et al.* (2010) iMAT: an integrative metabolic analysis tool. *Bioinformatics* 26, 3140–3142
57. 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
58. van Berlo, R.J.P. *et al.* (2011) Predicting metabolic fluxes using gene expression differences as constraints. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 206–216
59. Fang, X. *et al.* (2012) Modeling phenotypic metabolic adaptations of *Mycobacterium tuberculosis* H37Rv under hypoxia. *PLoS Comput. Biol.* 8, e1002688
60. Lee, D. *et al.* (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* 6, 73
61. Navid, A. and Almaas, E. (2012) Genome-level transcription data of *Yersinia pestis* analyzed with a new metabolic constraint-based approach. *BMC Syst. Biol.* 6, 150
62. Wang, Y. *et al.* (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* 6, 153
63. Kim, J. and Reed, J.L. (2012) RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol.* 13, R78
64. Rossell, S. *et al.* (2013) Inferring metabolic states in uncharacterized environments using gene-expression measurements. *PLoS Comput. Biol.* 9, e1002988
65. Cotten, C. and Reed, J.L. (2013) Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics* 14, 32
66. Labhsetwar, P. *et al.* (2013) Heterogeneity in protein expression induces metabolic variability in a modeled *Escherichia coli* population. *Proc. Natl. Acad. Sci. U. S. A.* 110, 14006–14011
67. Yizhak, K. *et al.* (2014) Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife* 3, e03641
68. Agren, R. *et al.* (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* 10, 721
69. Carrera, J. *et al.* (2014) An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. *Mol. Syst. Biol.* 10, 735
70. Gavai, A.K. *et al.* (2015) Using Bioconductor package BiGGR for metabolic flux estimation based on gene expression changes in brain. *PLoS One* 10, e0119016
71. Angione, C. and Lió, P. (2015) Predictive analytics of environmental adaptability in multi-omic network models. *Sci. Rep.* 5, 15147
72. Robaina Estévez, S. and Nikoloski, Z. (2015) Context-specific metabolic model extraction based on regularized least squares optimization. *PLoS One* 10, e0131875
73. Schultz, A. and Qutub, A.A. (2016) Reconstruction of tissue-specific metabolic networks using CORDA. *PLoS Comput. Biol.* 12, e1004808
74. Kim, M.K. *et al.* (2016) E-Flux2 and SPOT: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. *PLoS One* 11, e0157101
75. Fyson, N. *et al.* (2017) Gene-centric constraint of metabolic models. *bioRxiv*, Published online March 14, 2017. https://doi.org/10.1101/116558
76. Zhang, S.-W. *et al.* (2017) Prediction of metabolic fluxes from gene expression data with Huber penalty convex optimization function. *Mol. BioSyst.* 13, 901–909
77. Blais, E.M. *et al.* (2017) Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. *Nat. Commun.* 8, 14250
78. Lewis, J.E. *et al.* (2018) Genome-scale modeling of NADPH-driven β-lapachone sensitization in head and neck squamous cell carcinoma. *Antioxid. Redox Signal.* 29, 937–952
79. 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
80. Jenior, M.L. *et al.* (2020) Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. *PLoS Comput. Biol.* 16, e1007099
81. Gustafsson, J. *et al.* (2023) Generation and analysis of context-specific genome-scale metabolic models derived from single-cell RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2217868120
82. González-Arrué, N. *et al.* (2023) Phenotype-specific estimation of metabolic fluxes using gene expression data. *iScience* 26, 106201
83. Damiani, C. *et al.* (2019) Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS Comput. Biol.* 15, e1006733
84. Di Filippo, M. *et al.* (2022) INTEGRATE: model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLoS Comput. Biol.* 18, e1009337
85. Huang, Y. *et al.* (2023) Characterizing cancer metabolism from bulk and single-cell RNA-seq data using METAFlux. *Nat. Commun.* 14, 4883
86. Brandes, A. *et al.* (2012) Inferring carbon sources from gene expression profiles using metabolic flux models. *PLoS One* 7, e36947
87. Chen, Y. *et al.* (2024) Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. *Nat. Protoc.* 19, 629–667
88. Motamedian, E. *et al.* (2017) TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinformatics* 33, 1057–1063
89. Bekiaris, P.S. and Klamt, S. (2020) Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics* 21, 19
90. Robinson, D.R.L. *et al.* (2022) Applying sodium carbonate extraction mass spectrometry to investigate defects in the mitochondrial respiratory chain. *Front. Cell Dev. Biol.* 10, 786268
91. Machado, D. *et al.* (2016) Stoichiometric representation of gene-protein-reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS Comput. Biol.* 12, e1005140
92. Meldal, B.H.M. *et al.* (2019) Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* 47, D550–D558
93. Zdanowicz, R. *et al.* (2024) Stoichiometry and architecture of the human pyruvate dehydrogenase complex. *Sci. Adv.* 10, eadn4582
94. Marín de Mas, I. *et al.* (2019) Stoichiometric gene-to-reaction associations enhance model-driven analysis performance: Metabolic response to chronic exposure to Aldrin in prostate cancer. *BMC Genomics* 20, 652
95. Zhang, C. *et al.* (2015) Logical transformation of genome-scale metabolic models for gene level applications and analysis. *Bioinformatics* 31, 2324–2331
96. Hackett, S.R. *et al.* (2016) Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* 354, aaf2786
97. Chandrasekaran, S. and Price, N.D. (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* 107, 17845–17850
98. Eisenberg, E. and Levanon, E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.* 29, 569–574
99. Cao, X. *et al.* (2018) Protein moonlighting elucidates the essential human pathway catalyzing lipoic acid assembly on its cognate enzymes. *Proc. Natl. Acad. Sci. U. S. A.* 115, E7063–E7072
100. Supuran, C.T. (2016) Structure and function of carbonic anhydrases. *Biochem. J.* 473, 2023–2032

101. Li, F. *et al.* (2022) Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* 5, 662–672
102. Boorla, V.S. and Maranas, C.D. (2025) CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nat. Commun.* 16, 2072
103. Link, H. *et al.* (2013) Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat. Biotechnol.* 31, 357–361
104. Price, M.N. *et al.* (2018) Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* 557, 503–509
105. 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
106. Kook, Y. *et al.* (2022) Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. *Adv. Neural Inf. Proces. Syst.* 35, 31684–31696 https://proceedings.neurips.cc/paper_files/paper/2022/file/cdaa7f07b0c5a7803927d20aa717132e-Paper-Conference.pdf
107. Jadebeck, J.F. *et al.* (2023) Practical sampling of constraint-based models: optimized thinning boosts CHRR performance. *PLoS Comput. Biol.* 19, e1011378
108. Braunstein, A. *et al.* (2017) An analytic approximation of the feasible space of metabolic networks. *Nat. Commun.* 8, 14915
109. Bardoscia, M. *et al.* (2015) Phenotypic constraints promote latent versatility and carbon efficiency in metabolic networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 92, 012809
110. Kaufman, D.E. and Smith, R.L. (1998) Direction choice for accelerated convergence in hit-and-run sampling. *Oper. Res.* 46, 84–95
111. Megchelenbrink, W. *et al.* (2014) optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS One* 9, e86587
112. Haraldsdóttir, H.S. *et al.* (2017) CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics* 33, 1741–1743
113. Saa, P.A. *et al.* (2024) LooplessFluxSampler: an efficient toolbox for sampling the loopless flux solution space of metabolic models. *BMC Bioinformatics* 25, 3
114. De Martino, D. *et al.* (2018) Statistical mechanics for metabolic networks during steady state growth. *Nat. Commun.* 9, 2988
115. Keaty, T.C. and Jensen, P.A. (2020) Gapsplit: efficient random sampling for non-convex constraint-based models. *Bioinformatics* 36, 2623–2625
116. Damiani, C. *et al.* (2014) An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes. *Nat. Comput.* 13, 321–331
117. Zabinsky, Z.B. and Smith, R.L. (2013) Hit-and-run methods. In *Encyclopedia of Operations Research and Management Science*, pp. 721–729, Springer
118. De Martino, D. *et al.* (2015) Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding. *PLoS One* 10, e0122670
119. De Martino, D. (2017) Maximum entropy modeling of metabolic networks by constraining growth-rate moments predicts coexistence of phenotypes. *Phys. Rev. E* 96, 060401
120. Vehtari, A. *et al.* (2021) Rank-normalization, folding, and localization: an improved R^ for assessing convergence of MCMC (with Discussion). *Bayesian Anal.* 16, 667–718
121. Heirendt, L. *et al.* (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702
122. Ebrahim, A. *et al.* (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* 7, 74
123. Noor, E. (2018) Removing both internal and unrealistic energy-generating cycles in flux balance analysis. *arXiv*, Published online Marach 13, 2018. https://doi.org/10.48550/arXiv.1803.04999
124. Wiback, S.J. *et al.* (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
125. Almaas, E. *et al.* (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 427, 839–843
126. Thiele, I. *et al.* (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J. Biol. Chem.* 280, 11683–11695
127. Barrett, C.L. *et al.* (2009) Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC Syst. Biol.* 3, 30
128. Mo, M.L. *et al.* (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 3, 37
129. Bordel, S. *et al.* (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* 6, e1000859
130. Mardinoglu, A. *et al.* (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol. Syst. Biol.* 9, 649
131. Gomes de Oliveira Dal'Molin, C. *et al.* (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Front. Plant Sci.* 6, 4
132. McGarrity, S. *et al.* (2018) Metabolic systems analysis of LPS induced endothelial dysfunction applied to sepsis patient stratification. *Sci. Rep.* 8, 6811
133. Sulheim, S. *et al.* (2020) Enzyme-constrained models and omics analysis of *Streptomyces coelicolor* reveal metabolic changes that enhance heterologous production. *iScience* 23, 101525
134. Scott, W.T., Jr. *et al.* (2021) Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts. *Microb. Cell Factories* 20, 204
135. Arora, G. *et al.* (2023) Targeting metabolic fluxes reverts metastatic transitions in ovarian cancer. *iScience* 26, 108081
136. Poncheewin, W. *et al.* (2022) Comparative genome-scale constraint-based metabolic modeling reveals key lifestyle features of plant-associated *Pseudomonas* spp. *bioRxiv*, Published online July 27, 2022. https://doi.org/10.1101/2022.07.26.501552
137. Ren, J. *et al.* (2018) An unnatural pathway for efficient 5-aminolevulinic acid biosynthesis with glycine from glyoxylate based on retrobiosynthetic design. *ACS Synth. Biol.* 7, 2750–2757
138. Song, H. *et al.* (2008) Development of chemically defined medium for *Mannheimia succiniciproducens* based on its genome sequence. *Appl. Microbiol. Biotechnol.* 79, 263–272
139. Monk, J.M. *et al.* (2017) iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* 35, 904–908