Nature Reviews Genetics | AOP, published online 16 January 2014; doi:10.1038/nrg3643

Constraint-based models predict metabolic and associated cellular functions

Aarash Bordbar, Jonathan M. Monk, Zachary A. King and Bernhard O. Palsson

Abstract | The prediction of cellular function from a genotype is a fundamental goal in biology. For metabolism, constraint-based modelling methods systematize biochemical, genetic and genomic knowledge into a mathematical framework that enables a mechanistic description of metabolic physiology. The use of constraint-based approaches has evolved over ~30 years, and an increasing number of studies have recently combined models with high-throughput data sets for prospective experimentation. These studies have led to validation of increasingly important and relevant biological predictions. As reviewed here, these recent successes have tangible implications in the fields of microbial evolution, interaction networks, genetic engineering and drug discovery.

Understanding the genotype–phenotype relationship is at the core of the life sciences. For the latter half of the twentieth century, the reductionist approaches of genetics, biochemistry and molecular biology focused on the elucidation of biological components that underlie this fundamental relationship. These approaches have provided detailed understanding of individual components, but they do not address the systemic interactions of biological and environmental components that underlie phenotypes. Technological advances have now enabled high-throughput methods to comprehensively characterize biological components simultaneously. The cost of such data generation has decreased exponentially and the amount of data generated has become more abundant, which enables biologists to view and study cells as systems of interacting components.

To cope with the rapidly growing number of high-dimensional data sets, sophisticated data analysis methods are needed. Diverse approaches that range from stochastic kinetic models to statistical Bayesian networks have been applied, and each of these approaches has differing rationales and advantages (TABLE 1). One of these approaches is constraint-based reconstruction and analysis that is applied to genome-scale metabolic networks. Reconstructed genome-scale metabolic networks contain curated and systematized information about the known small metabolites and metabolic reactions of a cell type, which is based on its annotated genome and on experimental literature[1,2]. Genome-scale metabolic networks can be converted to a mathematically consistent format, which is known as the stoichiometric matrix (BOX 1). This matrix is the central component of a constraint-based model (CBM), which can be queried by an ever-growing set of modelling methods[3] (BOX 2). CBMs have been primarily built for metabolic networks, including multicellular metabolic interactions[4–8]. CBMs have also been built for signalling[9,10], transcriptional regulation[11] and macromolecule synthesis[12].

This Review illustrates how CBMs have recently provided the foundation for formulating genome-scale mechanistic predictions of metabolic physiology that are now being used in a prospective manner to elucidate new biological knowledge and understanding. We begin with a brief description of the four-phase history of the development of CBM applications. We then present studies that show the recent progress of integrating high-throughput data sets with the mechanistic and functional context of CBMs to predict metabolic phenotypes, and we emphasize the implemented workflows, limitations of the approach and opportunities for further development.

Foundational developments

Constraint-based analysis has been applied to biochemical reaction networks for more than 25 years. To put these developments into context, we exhaustively searched the literature using Web of Knowledge to collect research articles that use CBMs for interpreting

Department of Bioengineering, University of California, San Diego, 9500 Gilman Dr, La Jolla, California 92093–0412, USA.
Correspondence to B.O.P.
e-mail: palsson@ucsd.edu

doi:10.1038/nrg3643
Published online
16 January 2014

Table 1 | **A comparison of modelling and analysis techniques for high-throughput data**

Method	Model systems	Parameterization	Typical prediction type	Advantages	Disadvantages	Refs
Stochastic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	• Mechanistic • Dynamic • Captures biological stochasticity and biophysics	• Computationally intensive • Difficult to parameterize • Challenging to model multiple timescales	106
Deterministic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	• Mechanistic • Dynamic	• Computationally intensive • Difficult to parameterize	107
Constraint-based modelling	Genome-scale metabolism	Network topology, and uptake and secretion rates	Metabolic flux states and gene essentiality	• Mechanistic • Large scale • No kinetic information is required	• No inherent dynamic or regulatory predictions • No explicit representation of metabolic concentrations	3,104
Logical, Boolean or rule-based formalisms	Signalling networks and transcriptional regulatory networks	Rule-based interaction network	Global activity states and on–off states of genes	Can model dynamics and regulation	Biological systems are rarely discrete	108
Bayesian approaches	Gene regulatory networks and signalling networks	High-throughput data sets	Probability distribution score	• Non-biased • Can include disparate and even non-biological data • Takes previous associations into account	• Statistical • Issues of over-fitting • Requires comprehensive training data	109, 110
Graph and interaction networks	Protein–protein and genetic interaction networks	Interaction network that is based on biological data	Enriched clusters of genes and proteins	• Incorporates prior biological data • Encompasses most cellular processes	• Dynamics are not explicitly represented	111, 112
Pathway enrichment analysis	Metabolic and signalling networks	Pathway databases (for example, KEGG, Gene Ontology and BioCyc)	Enriched pathways	• Simple and quick • Takes prior knowledge into account	• Biased to human-defined pathways • Non-modelling approach	73

KEGG, Kyoto Encyclopedia of Genes and Genomes.

Metabolite overflows
Biological phenomena whereby the rate of substrate use by a cell for growth is lower than the rates of uptake and conversion of the substrate, which results in production of side metabolites (for example, acetate in *Escherichia coli*).

Metabolic fluxes
The rates of turnover or movement of metabolites through a reaction or a pathway.

Objective functions
The particular variables, or metabolic reactions, that are being maximized or minimized for by the linear programme. In flux-balance analysis, the objective function is often a pseudoreaction for biomass generation that represents cellular growth.

and predicting biological phenotypes. We collected 645 articles that were published from 1986 to 15 June 2013. The articles are available with short descriptions of their contributions to the research field at the literature on model-driven analysis website hosted by the University of California, San Diego Systems Biology Research group. An analysis of this literature shows that the history of CBMs can be divided into four phases.

Initial studies (1986–1998). CBMs were initially used to determine theoretical pathway yields and metabolite overflows[13,14]. Experimental metabolic fluxes and growth rates were shown to be consistent with fluxes that were computed on the basis of optimization of cellular objective functions, including minimal production of reactive oxygen species (ROS) for hybridoma cells[15] and maximal growth rate for laboratory strains of *Escherichia coli*[16]. Concurrently, algorithms — such as Elementary Flux Modes[17] and Extreme Pathways[18] — were developed[19] to exhaustively calculate metabolic pathways in CBMs for analysis of network topology[20] and for uses in metabolic engineering[21]. The quantitative match between CBM predictions and measured cellular behaviour opened up the possibility of predicting phenotypes from a biochemically reconstructed network.

Building genome-scale networks (1999–2004). The ability to sequence whole genomes[22] made it possible to formulate CBMs at the genome scale and allowed representation of the complete metabolic gene content in the assessment of phenotypic functions[23]. Importantly, metabolic reactions in a CBM could be directly linked to the genotype of the target cell, which allowed prediction of the consequences of gene knockouts[24,25]. These genome-scale models facilitated the study of the global organization of cellular behaviour, such as pathway structure[26], adaptive evolution end points[27], metabolic fluxes[28] and bacterial evolution[29,30].

Integrating omic data (2005–2009). As the generation of 'omic' data became cheaper and as larger data sets appeared, researchers began to incorporate these data sets into CBMs[31] (FIG. 1). Initially, the metabolic network was used as a scaffold to interpret transcriptional changes[32,33] in a manner that is similar to pathway enrichment analysis (FIG. 1a). Subsequently, omic data were used more directly by further constraining individual metabolic reactions to increase the context specificity of CBMs[34,35] (FIG. 1b).

Maturing to predictive practice (2010–present). These efforts resulted in highly curated and validated genome-scale models that are now enabling the research community to obtain meaningful predictions of biological functions. This Review is focused on this most recent phase in the field of CBM development.

We first discuss the latest evaluations of the assumptions of constraint-based modelling. Second, we discuss the integration of genome-scale data sets — specifically, omic data and biomolecular interaction data — with CBMs. Third, we focus on how discrepancies between model predictions and experimental data allow targeted experimentation that leads to biochemical discovery. Fourth, translational applications of constraint-based modelling, including metabolic engineering and drug target discovery, are discussed. Finally, we focus on recent advances of integrating CBMs with other modelling approaches to increase their predictive scope.

Refining objectives

The first constraint-based method for biological predictions was flux-balance analysis (FBA). Its formulation is rooted in the hypothesis that a cell is 'striving' to achieve a metabolic objective (BOX 2). Studies have shown that, by optimizing the assumed cellular objectives of growth[27] and energy use[36,37], one can predict metabolic fluxes in microorganisms. Other studies have questioned the universality of the objective function of biomass growth for predicting relevant metabolic fluxes[38–40].

Do cells maximize growth rate? To identify the objective function that best predicts experimental data on growing cells, one study[41] greatly expanded the initial assessment of appropriate objective functions for FBA by compiling 44 metabolic flux analysis data sets of *in vivo* flux distributions for *E. coli*, and the researchers evaluated the ability of a reduced CBM to predict these measurements using dozens of single and combined candidate cellular objective functions (FIG. 1c). The best representation for the *in vivo* fluxomic data sets was a Pareto surface that is defined by a combination of three objectives: maximizing biomass generation, maximizing ATP generation and minimizing reaction fluxes across the network; that is, the minimization is a proxy for the most efficient use of the proteome[42]. Using flux variability analysis (FVA) (BOX 2), the authors found that there is some 'slack' in metabolic reaction fluxes when the cell is operating close to but not on the Pareto surface. In fact, they also observed that the *in vivo* flux distributions were slightly sub-optimal. The authors showed that this sub-optimality is most likely an evolutionary adaptation that allows rapid adjustment to environmental perturbations. In this study, metabolic flux analysis simulations were limited to central carbon metabolism. Future studies are therefore needed to determine whether the optimality principles that have been derived in this study will hold for other metabolic subsystems that are studied using different

Metabolic pathways
In the context of this Review, sets of pathways that are calculated by metabolic network-based pathway analysis tools such as Extreme Pathways and Elementary Flux Modes.

Metabolic engineering
The practice of improving cellular production of target compounds of interest by modifying and optimizing genetic, regulatory and environmental parameters of cellular metabolism.

Genome-scale models
The formulation, using mathematical models, of genome-scale metabolic network reconstructions. They are synonymous with constraint-based models in the context of this Review.

Pathway enrichment analysis
A high-throughput data analysis technique to understand more global changes in an experiment by grouping individual measurements of biological components (for example, genes and proteins) into a context that is based on various pathway databases (for example, Kyoto Encyclopedia of Genes and Genomes, BioCyc and Gene Ontology).

Metabolic flux analysis
An experimental approach to identify metabolic fluxes using isotopically labelled metabolites and computational software that reconciles experimental data with network topology.

Flux distributions
Sets of calculated flux values for all reactions in a constraint-based model.

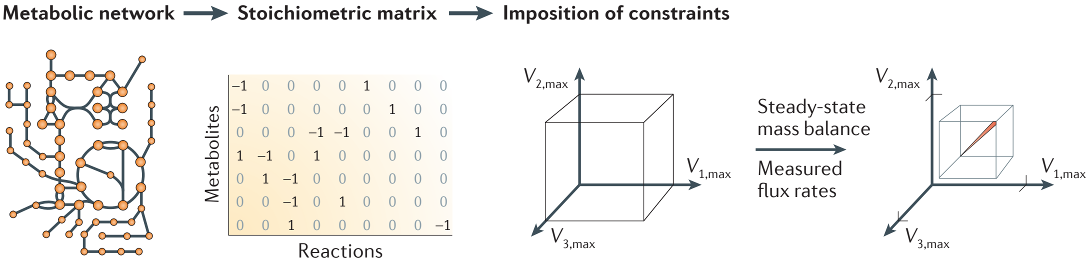
Pareto surface
The space that is formed when multiple objective functions are modelled at once; it represents a set of optimal solutions, in which increasing the value of one of the objectives results in a trade-off with other objective values.

Box 1 | Constraint-based modelling: motivation and definition

The functional capabilities of biological systems are constrained by their genetics and environment, and by physico-chemical laws. For example, most natural environments are limited in nitrogen or phosphate. In addition, the rate of photosynthesis is a function of latitude as the incident flux of photons changes. In the 1960s, Daniel Atkinson realized that solvent capacity was a limitation in all cells, as cells tend to consist of 70% water and 30% biomass[100]. In 1973, Paul Weisz showed that most intracellular processes operate at rates that are close to the limits of diffusion[101]. These and other myriad constraints under which cells operate and evolve have been summarized[102].

We can now systematically reconstruct metabolic and other biochemical reaction networks (see the figure). Metabolic networks are analogous to flow networks, in which metabolites (shown as circles) 'flow' through the network in a manner that is similar to liquids flowing in a pipe. These flows, and thus the state of a network, are subject to myriad constraints. The network can be converted into a mathematical format known as the stoichiometric matrix for computation. Rather than deriving a single solution, constraint-based models have an associated solution space (shown as a box) in which all feasible phenotypic states exist given the imposed constraints. This allows one to simultaneously account for the many processes that act on and in cells.

Metabolite flow is constrained by, among other things, the network topology (for example, the connection of metabolites) and a steady-state assumption (for example, the assumption that internal metabolites must be produced and consumed in a mass-balanced manner). It is also constrained by the known upper bounds (also known as capacities; for example, $V_{1,\max}$) and lower bounds of individual reaction fluxes. Imposing such constraints 'shrinks' the solution space to a more biologically relevant region. The challenges in constraint-based modelling lie in identifying and imposing the necessary and dominant constraints to define a solution space, as well as in probing the solution space in a manner such that physiologically relevant fluxes or phenotypes are determined.



$$\begin{array}{c|ccccccccc} & & & & & & & & & \\ & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ \text{Metabolites} & 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ \hline & \multicolumn{9}{c}{\text{Reactions}} \end{array}$$

Figure is modified, with permission, from REF. 3 © (2012) Macmillan Publishers Ltd. All rights reserved.

labelled metabolites. Nonetheless, CBMs can use optimality principles to predict the approximate growth state and the 'hedging' functions that keep the cells from fully reaching the predicted optimal states, which are yet to be delineated. Such hedging functions are expected to vary from strain to strain and from organism to organism on the basis of their evolutionary history.

Moving beyond the assumption of growth optimality. Although the prediction of optimal growth rates has historically received much attention, most of the recent studies that are highlighted in this Review do not assume optimality of cellular growth. Researchers have been increasingly adopting alternative unbiased approaches — such as Markov chain Monte Carlo (MCMC) sampling, omic data integration and metabolic pathway analyses[3] — that are not subject to assumptions of optimality.

Contextualizing omic data

The constraint-based modelling framework is amenable to simultaneous integration of a range of omic data types[31] (FIG. 1). In particular, omic data have been used both to constrain calculated flux distributions and as a comparison and validation tool for model predictions. Such omic data integration has enabled context-specific studies of the metabolism of an organism and, in the following cases, the studies of enzyme promiscuity and pathogenesis.

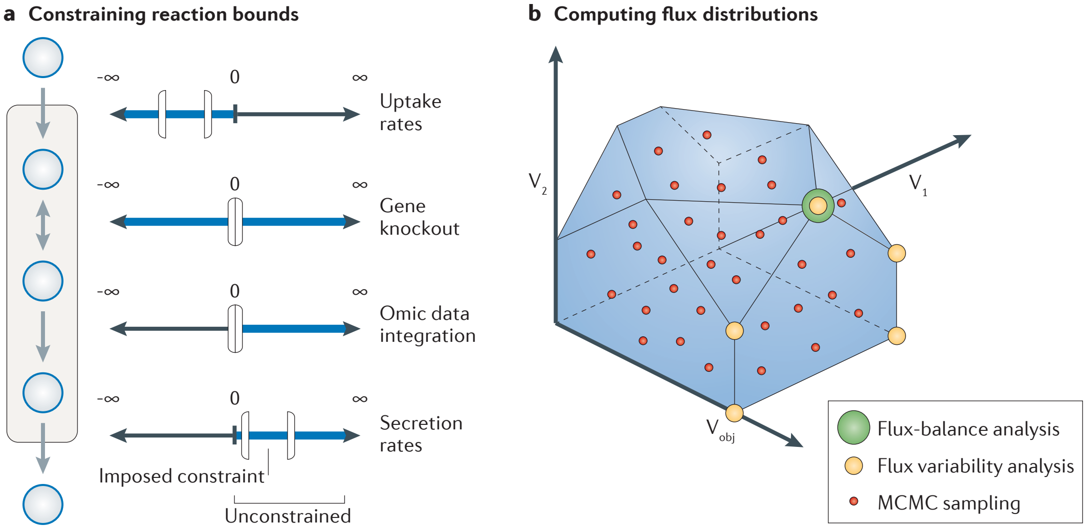
Why are some enzymes specific and some promiscuous? It is thought that ancestral enzymes were promiscuous and inefficient, and that they have evolved to become catalytically efficient and specific[43]. However, it is not well understood why such evolution took place for some enzymes but not for others. To address this question, one study[44] classified proteins in the *E. coli* CBM[45] into two groups (that is, specialist and generalist enzymes) on the

Box 2 | Constraint-based modelling: introduction to methods for analysis

Constraint-based models (CBMs) have been widely deployed; see REF. 3 for an extensive description of the developed methods. However, most of these techniques are based on two key components: the constraints on the biological system and the analysis method to predict fluxes (see the figure).

Constraining metabolic models. The toy model (see the figure, part **a**) contains metabolites (shown as circles) that are converted by reactions (shown as arrows). Each reaction has a range of potential flux values, which can be constrained (shown as sliders). The imposition of constraints defines the associated solution space of the CBM. Most methods modify the metabolic reaction bounds for model parameterization. Simple constraints include fixing cellular input and output ranges on the basis of uptake and secretion of metabolites[56], as well as carrying out genetic knockouts by setting the associated bounds of the reactions to zero[24]. More advanced techniques include modifying reaction bounds on the basis of mRNA and protein expression data, either by setting the bounds to zero for reactions that correspond to absent transcripts and proteins[34,35] (see the figure, part **a**) or by linearly adjusting the bounds on the basis of transcript and protein abundances[103].

Determining flux distributions. After model parameterization, fluxes are calculated. A simplified solution space is depicted with two reactions (V_1 and V_2) and an objective function (V_{obj}) (see the figure, part **b**). The standard approach of flux-balance analysis[104] either maximizes or minimizes the flux of a user-defined reaction (that is, the objective function) using linear programming (shown by the green circle). Another common approach is flux variability analysis[105], in which the maximum and minimum fluxes through each reaction are iteratively computed when the flux of the objective function is typically constrained to its maximum value (shown by yellow circles). Finally, Markov chain Monte Carlo (MCMC) sampling computes many candidate flux distributions (shown as red dots) that provide a probability distribution for the fluxes. This approach is unbiased, as no assumption of an objective is required.



Central carbon metabolism
The metabolic pathways and reactions that convert sugars into the metabolic precursors that are required for growth. It is typically comprised of glycolysis, pentose phosphate pathways and the tricarboxylic acid cycle.

Solution space
The range of all feasible values for variables in a constraint-based model, which represents all potential metabolic reaction flux distributions on the basis of the given constraints.

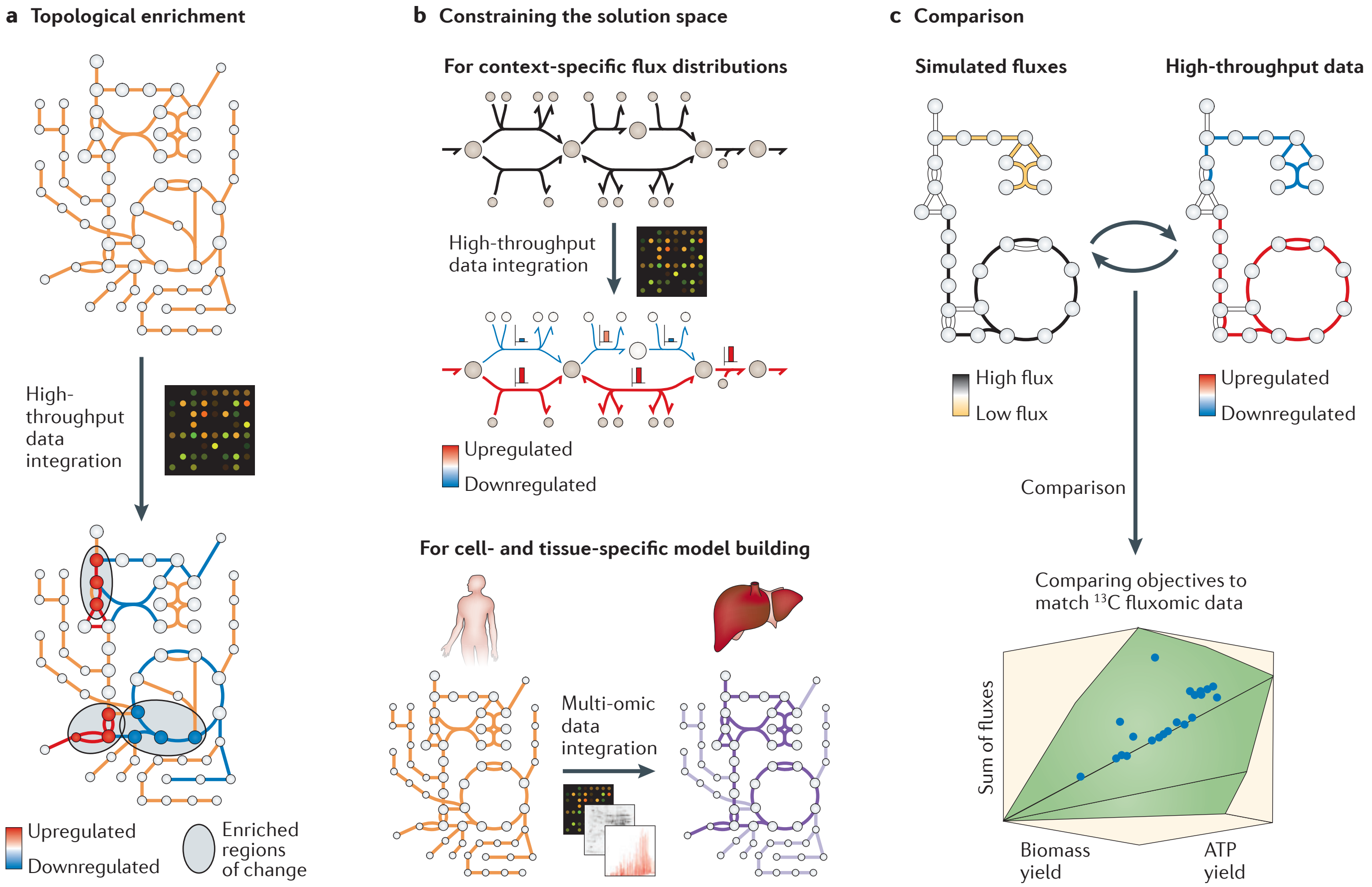

a Topological enrichment
High-throughput data integration
Upregulated
Downregulated
Enriched regions of change
b Constraining the solution space
For context-specific flux distributions
High-throughput data integration
Upregulated
Downregulated
For cell- and tissue-specific model building
Multi-omic data integration
c Comparison
Simulated fluxes
High-throughput data
High flux
Low flux
Upregulated
Downregulated
Comparison
Comparing objectives to match 13C fluxomic data
Sum of fluxes
Biomass yield
ATP yield

Figure 1 | **The multiple uses of high-throughput data in constraint-based models.** Constraint-based modelling can be used to interpret and augment omic data sets by using an underlying cellular network that has been biochemically validated. Metabolites are represented by circles. **a** | Similarly to pathway enrichment analysis and interaction networks, high-throughput data can be integrated with the metabolic network topology to determine enriched regions and even significantly perturbed metabolites[32]. **b** | Omic data add an additional layer of constraints for reaction fluxes. One study[48] integrated expression profiling data to determine context-specific flux distributions (pathway shown in red), which increases the fidelity of the data (represented as bars) as well as the accuracy of flux predictions (upper panel). In addition, two other studies[77,78] used omic data to build cell- and tissue-specific models of human metabolism by removing unexpressed reactions (shown as discoloured reactions) from the global human metabolic network (lower panel). Differences in these networks can be exploited to learn unique features of each network. **c** | Constraint-based analysis predictions can be compared and validated against high-throughput data sets. One study[41] compared flux-balance analysis solutions of different objectives against ^{13}C fluxomic data to find a combination of objectives that best fit the *in vivo* fluxes.

basis of the number of reactions that are catalysed by each enzyme. The authors showed that reactions that are associated with specialist enzymes are more likely to be essential on the basis of growth phenotypes of knockout strains, and that these reactions are more likely to carry a high and variable flux across hundreds of *in silico* environmental conditions on the basis of MCMC sampling (BOX 2). Large, disparate data sets were used to validate simulations. Gene essentiality predictions were validated by comparison with a gene deletion collection[46]. An analysis of kinetic parameters from the Braunschweig enzyme database (BRENDA)[47] showed that specialist enzymes have higher *in vitro* catalytic activity (that is, higher turnover number (k_{cat})) and higher substrate affinity (that is, lower Michaelis constant (K_m)). Omic data sets revealed that specialist enzymes are more tightly regulated at multiple levels, which is indicated by transcriptional and post-translational modifications, as well as by small-molecule-mediated control. Although enzyme promiscuity has not been fully elucidated and might not be fully captured in the model, the CBM is nevertheless the best self-consistent representation of known metabolic reactions and enzymes in *E. coli*. Consequently, these predictions provided a direction to integrate and interpret disparate data sets with the CBM, thereby validating a genome-scale hypothesis that the evolution of an enzyme towards specificity and catalytic efficiency is dependent both on the function of the enzyme in its metabolic network context and on its evolutionary response to selection pressures.

What is the role of metabolism in pathogenesis? Intracellular pathogens adapt metabolism to their host environment during pathogenesis. One study[48] generated transcriptional profiling data of pathogenic intracellular growth to investigate the relationship between metabolism and pathogenesis of *Listeria monocytogenes*. The researchers analysed the data both through traditional pathway enrichment analysis (TABLE 1) and through integration with a CBM of *L. monocytogenes*. They used the iMAT algorithm[34] that computes a flux distribution which best uses reactions that are associated with upregulated genes and which avoids using reactions that are associated with downregulated genes (FIG. 1b), thereby predicting differential reaction use between conditions. By comparing pathway enrichment analysis of the transcription data with and without iMAT, the authors found that iMAT increased accuracy in representing known changes in intracellular growth because both the CBM and the computed flux states contextualize the expression data. In this way, incorrectly upregulated transcripts — either due to a false-positive measurement or due to post-transcriptional regulation — are algorithmically 'corrected' if the rest of the associated pathway is inactive (FIG. 1b) and vice versa. The higher predictive accuracy helped the authors to focus their experiments on highly active pathways, which were then experimentally confirmed by generating conditional knockout strains. Prospective experiments that were based on the identified pathways showed that limiting concentrations of branched-chain amino acids induced virulence activator genes and elucidated the role of amino acid metabolism in pathogenesis. Thus, analysis of transcription data is often hindered by the low signal-to-noise ratio and by the limitation that post-transcriptional regulation is not captured in these data sets. These limitations can be ameliorated for metabolic transcripts by the contextualization of the iMAT algorithm and CBMs.

Characterizing interaction networks

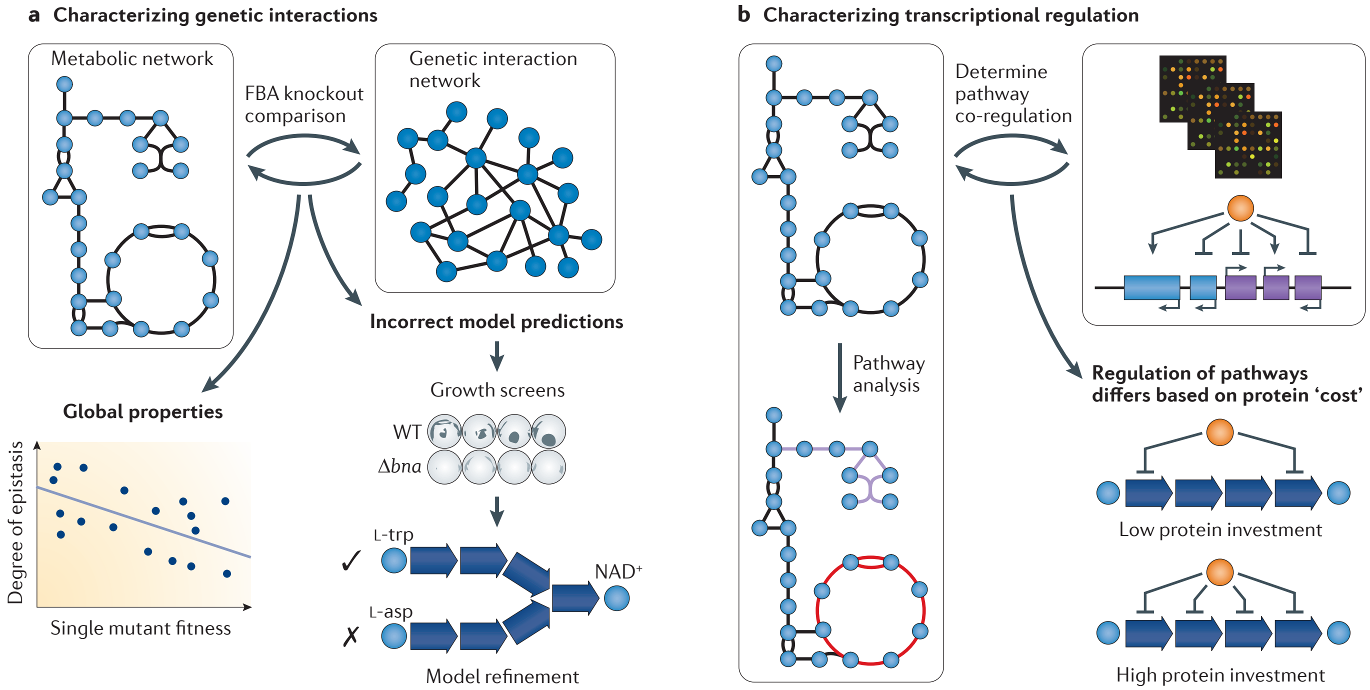
Recent work shows that CBMs can be used to place interaction networks of diverse biological components into context and to interpret these networks. Interaction networks describe the phenomenological interactions between different biomolecules, including genes[49], proteins[50] and transcription factors[51]. Such interaction networks are information dense and highly valuable, but they cannot generally be formulated into a modelling framework for prediction of physiological functions. In the following examples, the mechanistic information in CBMs is used in conjunction with biomolecular interaction networks to derive principles that underlie cellular organization.

Genetic interaction networks. The theoretical aspects of genetic interactions of metabolic genes in *Saccharomyces cerevisiae* that were derived using CBMs have been studied[52–54]. A recent study[55] used a CBM and experimental data to discover mechanistic principles that underlie global properties of *S. cerevisiae* genetic interaction data (FIG. 2a). First, the authors experimentally and computationally quantified genetic interactions for the genes in the *S. cerevisiae* CBM[56]. Both the experimental data and the computational predictions showed a global property that genes which are associated with low-fitness single mutants share many genetic interactions. They then used the CBM to propose a mechanistic explanation of this phenomenon. The researchers showed that these deleterious gene deletions directly disrupt the production of multiple metabolite precursors that are necessary for cellular growth. Thus, these genes share genetic interactions with other genes that contribute to various aspects of their functionality.

The same researchers found that FBA underpredicts genetic interactions, which can be attributed to the optimality assumption of FBA, to the inherent inability of FBA to capture regulation and data on toxic intermediates, or to an incompletely or incorrectly annotated metabolic network. To determine whether modifying the CBM could increase its predictive power, a machine learning method was implemented to reconcile the two networks by removing reactions, modifying reaction reversibility and altering the biomass function in the CBM. Model refinement identified one of the two NAD^+ biosynthetic pathways from amino acids in the CBM as a source of inaccurate predictions. Through growth experiments using mutant strains, the researchers confirmed that the second biosynthetic pathway was not present in *S. cerevisiae*. This study shows that CBMs can suggest the mechanistic underpinnings of genetic interaction networks and that the comparison of the metabolic and genetic interaction networks can lead to targeted improvements in biochemical knowledge.

Transcriptional regulatory networks. CBMs have aided the characterization of underlying principles of transcriptional regulatory networks for *E. coli* metabolism[57] (FIG. 2b). Previous studies have shown a moderate link between metabolic topology and transcriptional regulation[26,58]. To provide a more detailed analysis, one study[57] calculated potential pathways through metabolic subsystems of the *E. coli* CBM. Metabolic pathway structure has been of great interest[26] because a full enumeration of pathways can describe all possible steady-state metabolic phenotypes. However, the difficulty in computing Elementary Flux Modes for genome-scale networks has hindered their widespread use. A recently developed alternative — Elementary Flux Patterns[59] — calculates metabolic pathways in individual subsystems. This method ignores pathways that traverse multiple metabolic subsystems but is computationally tractable for genome-scale networks. By comparing Elementary Flux Patterns with transcriptional profiling data sets[60], the authors showed that pathways were only moderately co-expressed, but the degree of such expression varied greatly from perfect co-expression to no co-expression. They then showed that transcriptional regulation of pathways is dependent on the 'cost' of producing the associated enzymes and on the required response time. In pathways that contain

Machine learning method
A method that applies statistical methods to discover generalizable rules and patterns in complex data sets.



a Characterizing genetic interactions
Metabolic network
FBA knockout comparison
Genetic interaction network
Incorrect model predictions
Global properties
Degree of epistasis
Single mutant fitness
Growth screens
WT
Δbna
L-trp
L-asp
✓
✗
NAD+
Model refinement
b Characterizing transcriptional regulation
Determine pathway co-regulation
Pathway analysis
Regulation of pathways differs based on protein 'cost'
Low protein investment
High protein investment

Figure 2 | **Predictive case studies in understanding underlying principles of interaction networks.** Many network types are used to represent cellular behaviour. Recent studies have compared the properties of interaction networks against constraint-based models (CBMs) to learn global principles. **a** | One study[55] compared an experimental set of genetic interactions for metabolic genes against interactions that were predicted by flux-balance analysis (FBA). The CBM was able to recapitulate many of the *in vivo* principles. However, there was a high number of incorrect model predictions. Using machine learning techniques, key changes to the metabolic network that would improve model accuracy were identified. Using growth screens, the authors validated that the synthesis of NAD$^+$ from amino acids was only possible from L-tryptophan (L-trp) but not from L-aspartate (L-asp). Δ*bna* refers to any of the genes that are related to the kynurenine pathway, including *bna1*, *bna2*, *bna4* and *bna5*. **b** | Another study[57] calculated metabolic pathways — Elementary Flux Patterns — for the network. Elementary Flux Patterns decompose the metabolic network into distinct functional pathways (shown by different colours). The degree of co-regulation of the genes of each pathway was calculated, which reveals that some pathways are highly correlated, whereas others are not. Variation in co-regulation was attributed to the 'cost' that is needed for building the proteins in a particular pathway.

'expensive' proteins (that is, proteins that are larger in size), transcriptional regulation typically occurs for all enzymes in the pathway, whereas pathways with 'low-cost' proteins are typically transcriptionally regulated only at the first and last enzymes of the pathway. This categorization explained some cases of low co-expression. Thus, by pairing the CBM network topology with transcriptional regulatory networks, this study was able to outline key principles of metabolic regulation for different types of metabolic pathways.

Targeted expansion of metabolic knowledge

The studies discussed above focus on integrating CBMs with large-scale data sets to gain mechanistic understanding. However, incomplete knowledge of the metabolism of the target cell leads to inaccurate predictions. One feature of computational models is that incorrect predictions can identify missing or incorrect metabolic knowledge. Thus, the discrepancies between CBM predictions and experimental data have been used to design targeted experiments that correct such inaccuracy in metabolic knowledge[61,62].

Gap-filling
Pertaining to a procedure for targeted expansion of metabolic knowledge, whereby prospective experiments are designed on the basis of discrepancies in experimental data and model predictions.

Discovering new human metabolic capabilities. The initial reconstruction of the global human metabolic network — Recon 1 (REF. 63) — is incomplete owing to gaps in our knowledge of human metabolism. Thus, Recon 1 is missing metabolic reactions. Using an established protocol[62], one study[64] identified such gaps in our knowledge by simulating either production or consumption of every metabolite to assess whether the metabolite was fully connected to the rest of the network. For the metabolites that were not fully connected, a universal database of metabolic reactions[65] was used to predict the fewest reactions that were required to fully connect them. The authors found 73 candidate 'gap-filling' solutions that fully connected the disconnected metabolites, 47 of which were supported by the literature. Focusing on gluconate, which is a disconnected metabolite, the authors experimentally characterized open reading frame 103 on chromosome 9 (C9orf103) as the gene that encodes gluconokinase. This study illustrates how a self-consistent model of metabolism guides researchers to refine experiments to fill in missing gaps in our current knowledge.

***Discovering enzyme functions in* E. coli.** Constraint-based modelling has been used to discover biochemical knowledge about the well-characterized metabolic network of *E. coli*. Through systematic genetic perturbation of *E. coli* central carbon metabolism, one study[66] discovered a novel pathway and previously uncharacterized enzymatic functions. Single, double and triple knockout strains of central metabolic genes were grown on 13 different growth conditions with various carbon sources to determine positive and negative genetic interactions. Concurrently, genetic interactions were predicted using the *E. coli* CBM[45]. After careful removal of false predictions that were due to model assumptions (for example, the inability of FBA to differentiate between major and minor isozymes, as enzyme abundance and kinetic activity are not captured), it was observed that discrepancies that were related to *talAB* interactions in the pentose phosphate pathway could not be reconciled. To determine the cause, the authors generated transcription and metabolite profiling data for the wild-type and knockout strains. A metabolomic analysis identified a new metabolite — sedoheptulose-1,7-bisphosphate — that had not been previously characterized in *E. coli*, which suggests the existence of a novel reaction. Using metabolic flux analysis and *in vitro* enzyme assays, they confirmed that phosphofructokinase carries out the reaction and that glycolytic aldolase can split the seven-carbon sugar into three- and four-carbon sugars. Thus, the detailed analysis of the CBM against data discrepancies found two new catalytic functions of classic glycolytic enzymes.

Designing metabolic phenotypes

CBMs have been used for translational applications, including the design of metabolic phenotypes. In the past ten years, many algorithms have been developed for predicting useful genetic manipulation strategies for metabolic engineering[67,68]. They have also been important in assessing the net energy balance and the level of greenhouse gas emission for bioethanol and biodiesel production[69]. Here, we discuss one recent CBM success in this field of research.

Production of non-natural, commodity chemicals. There has been a push to use biotechnology to produce commodity chemicals. To this end, one study[70] designed an *E. coli* strain that produces 1,4-butanediol (BDO) at high yields. Two key hurdles were overcome using computational methods. First, BDO is not a naturally occurring compound in any organism. The authors used a pathway prediction algorithm[71] that determines the necessary biochemical transformations to convert an endogenous *E. coli* metabolite to BDO. A final pathway was chosen on the bases of thermodynamic feasibility[72], the theoretical yield of BDO (which was determined using FBA), the number of known enzymes for the biochemical transformations (which was determined using pathway databases[73]) and the topological distance of the pathway from central carbon metabolism. Second, when the pathway was introduced, the organism did not produce BDO at high rates; thus, a 'rational' approach to producing a metabolic design was pursued using the *E. coli* CBM[45] and the OptKnock algorithm[74]. A four-knockout strategy that blocked the production of natural fermentation products was chosen to force the strain to balance redox and to channel all carbon flux through BDO production (FIG. 3a). Further genetic manipulations were needed to create the final strain, which included modifying transcription factors, swapping *E. coli* metabolic enzymes with non-native enzymes and optimizing codons. There are many hurdles to designing a production strain, but this study shows that CBMs can have a vital role in accelerating and completing the industrial strain design pipeline to produce non-natural metabolites.

Discovering drug targets

The ability of constraint-based modelling to predict the effects of gene knockouts provides an important tool for drug targeting studies[75]. Three recent experimentally validated studies have discovered novel cancer drug targets and antibiotics.

Exploiting deficiencies in cancer metabolism. There has been renewed interest in studying metabolic alterations that occur in cancer cells[76]. In two studies[77,78], researchers hypothesized that they could use CBMs to determine and exploit the metabolic auxotrophies of cancer cells. The first study[77] used a model-building algorithm[79] that uses cues from transcriptomic data to prune metabolic reactions from Recon 1 (REF. 63) in order to build a 'generic' cancer model (FIG. 1b). They then used FBA-predicted knockout phenotypes to determine 'cytostatic' drug targets that selectively block growth of the cancer model but that do not affect ATP generation and growth of the 'healthy' Recon 1 model. Interestingly, even though cancer cells have heterogeneous genotypes and phenotypes, it was found that approved or experimental cancer drugs exist for 40% of the determined cytostatic drug targets. Analyses of growth phenotypes using CBMs focus on the capacity for growth under single knockout conditions. The surprising agreement between computational predictions and experimental results for a generic model suggests that the metabolic capabilities of cancer cells are starkly different from those of healthy human cells, which allows drug combinations to be detected.

In a follow-up study[78], the researchers experimentally investigated fumarate hydratase deficiency that can cause hereditary leiomyomatosis and renal cell cancer. At the time, no mechanism for reduced NADH regeneration was known for fumarate hydratase-deficient cells. The researchers immortalized and constructed two cell lines, one of which expressed fumarate hydratase and one that was deficient in fumarate hydratase. Starting from Recon 1, transcription data were used to build two cell line-specific models. By comparing predicted knockout growth phenotypes, they identified a selectively essential pathway for haem biosynthesis and degradation in fumarate hydratase-deficient cells, which represented a potential mechanism for NADH regeneration. Haem oxygenase 1 (HMOX1) was experimentally inhibited in both cell lines, and fumarate hydratase-deficient cells were selectively killed, which shows that fumarate hydratase and HMOX1 are in fact synthetically lethal as predicted.

Auxotrophies
Metabolic limitations that impair the ability of a cell or organism to synthesize a particular metabolite that is essential for growth, which force the cell or organism to rely on an exogenous source of the nutrient.

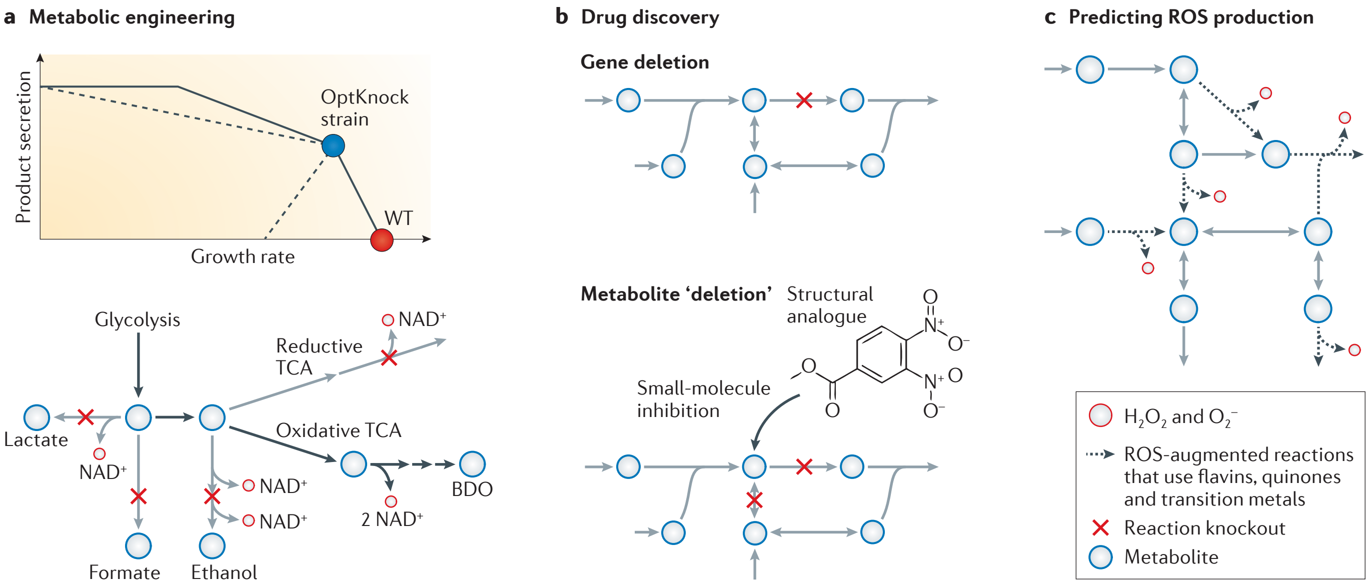


Figure 3 | **Predictive case studies in metabolic engineering and drug targeting.** Constraint-based models have been used for answering important questions in translational research. **a** | One study[70] used multiple computational and experimental tools to design an *Escherichia coli* strain that produces 1,4-butanediol (BDO). An unengineered wild-type (WT) strain trades off metabolite production with cellular growth (shown by the solid line in the solution space). Using the OptKnock algorithm, BDO production was 'coupled' with the growth objective of the cell by forcing the synthetic BDO pathway to be the sole route for *E. coli* to maintain redox balance (shown by black arrows). Thus, the solution space is modified such that BDO production is linked to cellular growth (shown by the dashed line in the solution space). **b** | In one study[81], researchers took an alternative, metabolite-centric approach to drug targeting, which computationally removes consuming reactions of a particular metabolite. The approach was experimentally confirmed for *Vibrio vulnificus* by a structural analogue of the endogenous metabolite, which also acts as a small-molecule inhibitor. **c** | Metabolic reactions in the *E. coli* model were augmented to capture the generation of reactive oxygen species (ROS), which allowed the use of flux-balance analysis to predict ROS production in one study[82]. In follow-up experiments, the authors show that it is possible to predict drug target strategies to enhance endogenous ROS production to increase the efficacy of other antibiotics. TCA, tricarboxylic acid cycle.

Interestingly, the haem pathway ranked only 39th in terms of overexpressed pathways in fumarate hydratase-deficient cells, which meant that the predictions were only possible through combining expression data with the CBM. These results are a step towards identifying effective anticancer drugs using genome-scale metabolic knowledge. As the predictions of the CBM focus on differential metabolic capacities, there is a potential for false-negative predictions, as additional layers of differences are not taken into account. In addition, it will be interesting to see whether these methods can be extended to other cancer types in which germline mutations are either unknown or absent. The identification of the haem pathway as synthetically lethal with fumarate hydratase represents a key success in using CBM predictions for prospective experimentation for studying human disease.

Essential metabolites guide antibiotic discovery. Gene knockout simulations in CBMs are accomplished by constraining the gene-associated reaction bounds to zero (BOX 1). Moving past a gene-centric approach, an alternative approach for drug targeting is metabolite essentiality analysis[80] (FIG. 3b). To 'remove' a metabolite in a CBM, the bounds of the reactions that consume the metabolite are constrained to zero, and the steady-state constraint for that particular metabolite is relaxed to allow internal metabolite accumulation.

One study[81] reconstructed a genome-scale metabolic network for *Vibrio vulnificus*, which is a Gram-negative pathogen. By applying metabolite essentiality analysis, the authors found 193 metabolites that are essential to cellular growth. They narrowed the list down to five essential metabolites that represented promising targets for drug development by removing metabolites that are found in humans to lower potential toxic adverse effects and by removing metabolites that have a single consuming reaction for a more robust effect on the pathogen. The identified metabolites typically affect a single gene, which means that traditional reaction knockouts could have been used.

However, using a metabolite-centric approach has its advantages. It allowed the authors to search for structural analogues of the essential metabolites to inhibit the enzymes that relied on them as substrates. They screened the inhibitory capability of 352 compounds that were structurally similar to the predicted essential metabolites, and the most potent compound was chosen for further evaluation as an antibiotic. The compound was confirmed to bind to the target enzyme in folate

Reaction bounds
User-defined constraints on the minimum and maximum allowable flux values for a particular metabolic reaction in a constraint-based model.

Metabolite essentiality analysis
A metabolite-centric approach to determine essential components for cellular growth. To computationally test the essentiality of a metabolite, the consuming reactions of the particular metabolites are constrained to zero.

biosynthesis, which validated the CBM and chemoinformatic predictions. Furthermore, they found that the compound was more effective than current antibacterials. By using a CBM to analyse metabolism, antibiotic discovery can be approached from multiple perspectives (for example, from the perspective of a gene, a reaction or a metabolite).

Increasing antibiotic efficacy through ROS production. ROS can weaken and kill pathogens, and modulation of ROS production could therefore be used as part of an antimicrobial strategy. As a proof of concept, one study[82] predicted genetic engineering strategies in *E. coli* to increase internal ROS generation in order to increase antibiotic efficacy. The current *E. coli* CBM[45] does not account for the major sources of ROS production. Thus, 133 metabolic reactions with potential for ROS generation were augmented in the *E. coli* CBM (FIG. 3c). With the updated CBM, computed flux distributions of single knockouts included a quantitative readout of ROS generation. Thus, gene knockouts that increase the endogenous ROS production were predicted, many of which increased inefficiencies in production or usage of ATP. For validation, the researchers experimentally knocked out genes that were predicted to increase endogenous ROS production, as well as genes that were predicted to have no effect as negative controls. There was high qualitative concordance of the predictions with experimental measurements of ROS production, which suggests that a CBM could be used to tune ROS production. These results are striking because little quantitative information was necessary in the coupling of flux with ROS production and because a statistical ensemble approach was used to account for unknown parameters. This study was able to predict genetic engineering strategies that were proven to increase ROS production and to potentiate oxidative attack from oxidants and antibacterials, which provides a novel approach for antibiotic discovery.

Coupling with other cellular processes

Constraint-based modelling has been mainly applied to metabolism. However, researchers have recently extended the scope of CBMs and combined them with different modelling methods to address questions beyond metabolism. Two approaches that have emerged are extending CBMs of metabolism to include additional cellular processes and connecting different modelling methods.

Modelling transcription, translation and metabolism. CBMs have been reconstructed for cellular processes other than metabolism[9–12]. However, until recently, CBMs of different cellular processes had not been integrated. One study[83] integrated a CBM of *Thermotoga maritima* metabolism[84] with a CBM for transcription and translation[83] (FIG. 4a). By adding information about the transcription and translation machinery, the CBM accounts for mRNA transcription, protein translation, necessary post-translational modifications of proteins and use of the protein complex to catalyse metabolic reactions in *T. maritima* (FIG. 4a). To couple the necessary machinery for a particular metabolic reaction, the authors used coupling constraints[85] that mathematically link a metabolic reaction flux with its required molecular and enzymatic machinery in formulating the linear programming problem. The result is an integrated network reconstruction that contains the molecular biology and metabolism of *T. maritima* at the genome scale and that allows the computation of the functional proteome that is needed to express a given phenotype. The incorporation of new cellular processes in the constraint-based modelling framework is exciting but requires additional parameterization of enzyme efficiencies under different biological conditions. A key challenge for the improvement and the use of these new models is the development of parameterization techniques that are driven by proteomic and transcriptomic data.

The integrated model hopes to address some of the crucial challenges that have limited metabolic CBMs. First, the integrated model takes into account the variability of cellular composition at different growth rates, while metabolic CBMs only use one biomass function for growth rate optimization. Cellular composition is dependent on growth rate, and metabolic CBMs have previously accounted for variations in growth rate with different cellular compositions[86]. However, an integrated model explicitly represents nucleotide and protein demands as a function of growth rate, so that a traditional biomass function is no longer necessary. Second, by coupling transcript and protein synthesis with active metabolic reactions, the authors quantitatively predicted differential experimental transcriptome and proteome levels across varying conditions. They used upstream genomic sequences of the differentially expressed genes to determine putative consensus sequences for transcription factor binding. The newly derived sequences helped to identify a candidate metabolite transporter, which was subsequently verified experimentally[87]. Finally, by incorporating the required demands for the machinery for metabolism, the integrated CBM unifies the three-objective Pareto surface that was discussed above into a single objective[88]. Thus, as the content of these models increases, the ability of CBMs to explain and predict biological functions grows in scope.

Merging statistics with mechanistic networks. Statistical approaches are useful when there is limited knowledge of the underlying biological networks. Unlike metabolic networks, the functional states of transcriptional regulatory networks (TRNs) are harder to define mechanistically because the underlying biochemistry and biophysics are often unknown. One study modelled the cellular processes of metabolism and transcriptional regulation using two different modelling formulations, which included a CBM for metabolism and a probability metric that is based on omic data for the TRN[89] (FIG. 4b). For *E. coli* and *Mycobacterium tuberculosis*, the authors amassed the available transcriptional profiling data sets and the existing

Coupling constraints
Constraints that enforce strict relationships between model biochemical transformations, thereby connecting the fluxes for different cellular processes (such as transcription, translation, and tRNA and protein use for a metabolic reaction).

Linear programming
A mathematical optimization technique that calculates the maximum or minimum value of a particular variable (that is, the objective function) on the basis of a set of linear constraints; an example of this is flux-balance analysis.

Consensus sequences
Conserved sequences of nucleotides or amino acids that represent the target for a biomolecular event, often for proteins binding to the genome.

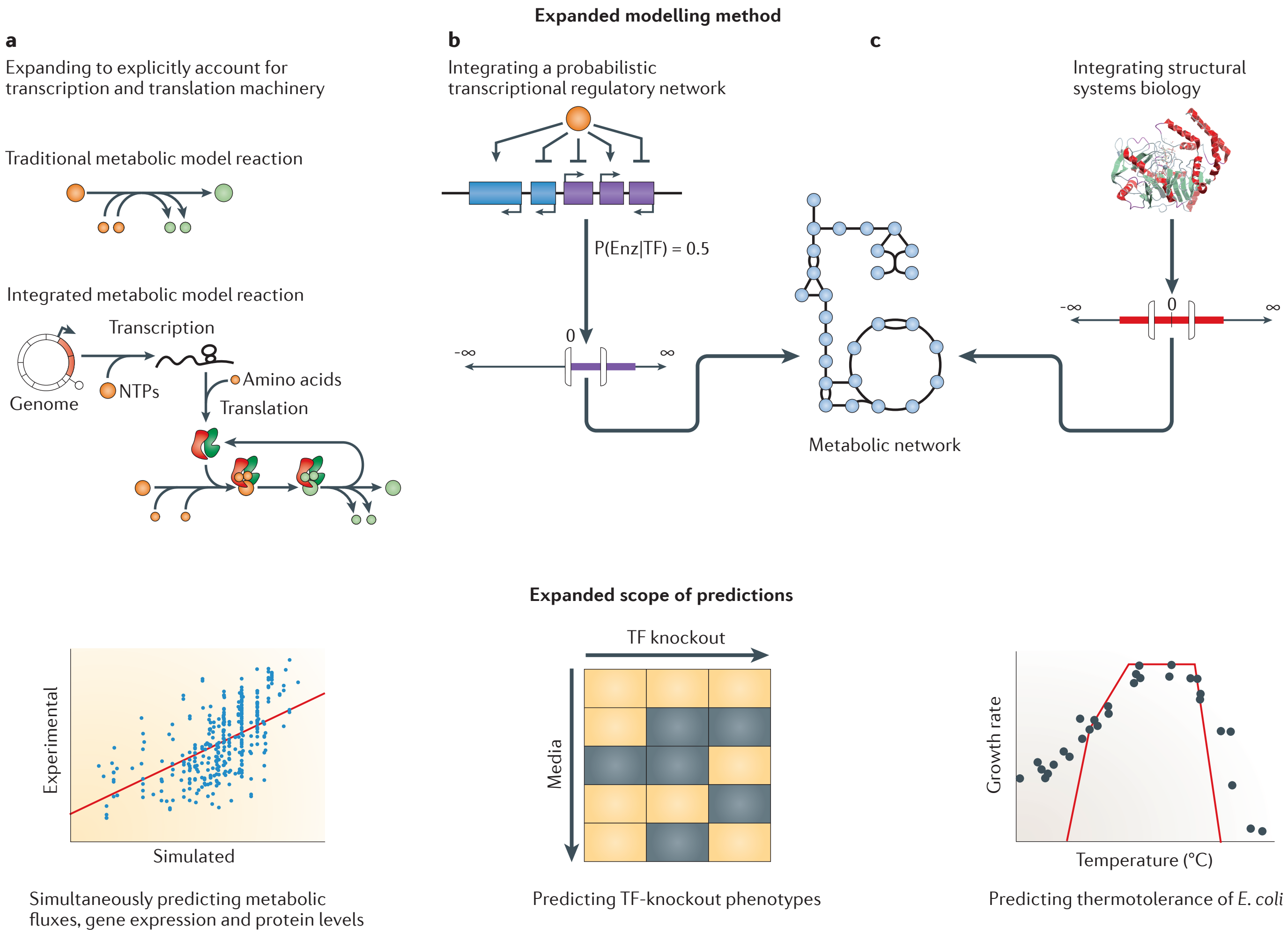

Expanded modelling method
a
Expanding to explicitly account for transcription and translation machinery
Traditional metabolic model reaction
Integrated metabolic model reaction
Transcription
Genome
NTPs
Amino acids
Translation
b
Integrating a probabilistic transcriptional regulatory network
P(Enz|TF) = 0.5
0
-∞
∞
Metabolic network
c
Integrating structural systems biology
0
-∞
∞
Expanded scope of predictions
Experimental
Simulated
Simultaneously predicting metabolic fluxes, gene expression and protein levels
TF knockout
Media
Predicting TF-knockout phenotypes
Growth rate
Temperature (°C)
Predicting thermotolerance of *E. coli*

Figure 4 | **Expanding predictive scope through integrative modelling.** The predictive scope of constraint-based modelling has been extended beyond metabolism either by explicitly accounting for non-metabolic components in the constraint-based modelling approach or by coupling with other modelling frameworks. Metabolites are represented by circles. **a** | The transcription and translation of the necessary mRNA, proteins and cofactors have been explicitly represented in a constraint-based modelling framework alongside the metabolism of *Thermotoga maritima*[83] (upper panel). This allows simultaneous computation of metabolic fluxes, mRNA transcript expression and proteome levels (lower panel). **b** | Metabolic models have also been coupled with other modelling frameworks. The probability of metabolic gene activation and repression by transcription factors (TFs) can be computed using a probabilistic transcriptional regulatory network that is based on high-throughput data sets (upper panel). The calculated probabilities are then relayed into the constraints of the metabolic reaction fluxes in the constraint-based model[89], which allow prediction of TF-knockout phenotypes (lower panel). **c** | Structural systems biology can predict biophysical properties of proteins. One study[91] calculated the individual activity changes of each metabolic enzyme during temperature shift. The combined effect of all the metabolic enzymes on the cell was computed by integrating the individual enzyme changes into the flux constraints of the *Escherichia coli* constraint-based model (upper panel), which allowed growth rate to be predicted as a function of temperature (lower panel). Enz, enzyme; NTP, nucleoside 5′-triphosphate; P, probability.

transcriptional regulatory interaction networks. Rather than using a Boolean formulation for the TRN[90], they calculated the probabilities of activation and repression on the basis of the collected expression data sets for each pair of transcription factor and target gene.

Similarly to how basic constraints are added (BOX 2), the TRN was combined with metabolism by adjusting upper and lower bounds of individual metabolic reactions in the CBM on the basis of both the calculated probabilities of activation of associated target metabolic genes and the allowable flux states (which are determined by FVA) (FIG. 4b). The integrated *E. coli* metabolic regulatory model was more accurate in predicting transcription factor-knockout phenotypes than previous attempts that used integrated models[90]. The newly developed integrated *M. tuberculosis* network predicted drug targets and aided the identification of novel regulatory roles of transcription factors. Although the TRN modifies the CBM of metabolism, the calculated flux distributions and the metabolites that are present do not feedback to parameterize the TRN. A further improvement of integrated modelling between transcriptional and metabolic networks will be to include feedback mechanisms from metabolism.

Structural systems biology. One study expanded the *E. coli* metabolic network by including the experimentally derived protein structures (where available) and the computationally predicted protein structures for the metabolic enzymes in the network[91] (FIG. 4c). Using structural bioinformatic[92] techniques, changes in enzyme activity were predicted and were used as constraints on the activity of individual metabolic reactions. The researchers focused on thermostability of *E. coli* enzymes to study growth rate as a function of temperature. With this approach, they were able to computationally predict growth rates at varying temperatures, which were consistent with experimental data. The growth-limiting enzymes were then determined on the basis of temperature-dependent flux constraints. Although other temperature-dependent parameters, such as cellular composition, were not considered[93], the predicted growth-limiting enzymes significantly overlapped with mutated genes from a previous study[94] on adaptive evolution of *E. coli* to higher temperatures. For direct experimental validation, the growth-constraining enzymes were bypassed by supplementing growth media with the metabolic product of the enzyme or of the pathway to which it belongs. Such supplementation was beneficial for *E. coli* that was grown at superoptimal temperatures, which supported the predictive capability of CBMs to account for enzyme thermosensitivity. These promising results raise the prospect of the substantial effects that structural modelling might have on improving CBM predictions in the future.

Conclusions

Gregor Mendel described discrete quanta of information travelling from one generation to the next, which determines the form and the function of an organism. Subsequently, Wilhelm Johannsen formulated the concept of a gene as the quanta of information, which led him to the definition of genotype and phenotype. Since then, a major goal of biology has been the quantitative description of the fundamental genotype–phenotype relationship.

The push in the quantitative biological sciences to understand macroscopic properties from microscopic measurements has parallels to the elucidation of fundamental principles in physics several hundred years ago. For example, the Einstein–Smoluchowski relation is a model for Brownian motion that quantitatively predicts properties of diffusion. Although the theory was an approximation of the physical processes[95], it has been applied and has helped to develop more sophisticated models. This Review suggests that the life sciences have now reached a point at which many aspects of the genotype–phenotype relationship for metabolism can be quantified and used to build mechanistic models that allow meaningful biological predictions to be made. The formulation of high-dimensional models that are required to compute full molecular phenotypes are enabled by genome sequencing technology, which allows the generation of a cellular parts list; by various omic data types, which allows a functional readout of these parts; and by mechanistic modelling frameworks that are amenable to reconciling omic data, network structure and knowledge from primary literature. The successes of the 14 studies discussed here demonstrate that constraint-based modelling is an approach that enables the genome-scale study of metabolism.

As with any model, the mathematical theory and the applications of constraint-based modelling will continue to be challenged and refined, thus improving our interpretation of biological phenomena. We foresee progress to unfold in several major directions. First, constraint-based modelling has mainly focused on metabolism, and more integrative modelling approaches must be explored. Trends in current literature[89,91,96] indicate that other cellular processes may be modelled using alternative frameworks that are better suited for a particular biological phenomena. Statistical approaches are also powerful for modelling biological processes that are poorly understood. Integrating other approaches with CBMs of metabolism can expand the scope of quantitative prediction. Second, the majority of applications of CBMs have been for single-cell organisms. We see two areas of application into which CBMs are likely to expand: human disease and the microbiome. Although the human reconstruction (that is, Recon 1) is far from complete, the cancer drug target studies showed that quantitative predictions are still possible. With the availability of the second build (that is, Recon 2)[97], we foresee greater applied uses in human disease. There has also been a steady increase in the amount of omic data of the human microbiome, and CBMs will have an important role in analysing these complex data sets[98,99]. Third, the underlying assumptions and methods for constraint-based modelling analyses will continue to evolve as more data types become available. Similarly to the testing of optimality assumptions of FBA, other key assumptions of CBMs will be tested in the next few years. For example, with the increasing availability of time-course metabolomics, the steady-state assumption can be bypassed and concentration changes can be explicitly modelled. Rather than assuming constant internal metabolite levels, these concentrations can be directly measured over a time course in an experiment and the rate of change can be integrated explicitly. In addition, the increasing availability of genomic data and sophisticated models for the interpretation of these data will allow explicit description and integration of the dependence of genomic sequence on gene expression, protein synthesis and protein structures for metabolic reactions in CBMs. We anticipate that these developments will enable even greater growth in the diversity of predictions and in the biological discoveries that are achievable by using constraint-based modelling.

1. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. Reconstruction of biochemical networks in microorganisms. *Nature Rev. Microbiol.* **7**, 129–143 (2009).
   This is a review on constructing and validating a genome-scale metabolic network.
2. Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protoc.* **5**, 93–121 (2010).
3. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. *Nature Rev. Microbiol.* **10**, 291–305 (2012).
   This is a thorough review of the various constraint-based modelling methodologies.
4. Zhuang, K. *et al.* Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *ISME J.* **5**, 305–316 (2011).
5. Klitgord, N. & Segre, D. Environments that induce synthetic microbial ecosystems. *PLoS Comput. Biol.* **6**, e1001002 (2010).
6. Bordbar, A. *et al.* A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst. Biol.* **5**, 180 (2011).
7. Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. O. & Jamshidi, N. Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Mol. Syst. Biol.* **6**, 422 (2010).
8. Lewis, N. E. *et al.* Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nature Biotech.* **28**, 1279–1285 (2010).
9. Papin, J. A. & Palsson, B. O. The JAK–STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87**, 37–46 (2004).
10. Li, F., Thiele, I., Jamshidi, N. & Palsson, B. O. Identification of potential pathway mediation targets in Toll-like receptor signaling. *PLoS Comput. Biol.* **5**, e1000292 (2009).
11. Gianchandani, E. P., Joyce, A. R., Palsson, B. O. & Papin, J. A. Functional states of the genome-scale *Escherichia coli* transcriptional regulatory system. *PLoS Comput. Biol.* **5**, e1000403 (2009).
12. Thiele, I., Jamshidi, N., Fleming, R. M. & Palsson, B. O. Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5**, e1000312 (2009).
13. Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* **238**, 781–786 (1986).
14. Majewski, R. A. & Domach, M. M. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnol. Bioeng.* **35**, 732–738 (1990).
15. Savinell, J. M. & Palsson, B. O. Optimal selection of metabolic fluxes for *in vivo* measurement. II. Application to *Escherichia coli* and hybridoma cell metabolism. *J. Theor. Biol.* **155**, 215–242 (1992).
16. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731 (1994).
17. Schuster, S. & Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Systems* **2**, 165–182 (1994).
18. Schilling, C. H., Letscher, D. & Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).
19. Clarke, B. L. in *Advances in Chemical Physics* Vol. 43 (eds. Prigogine, I. & Rice, S. A.) 1–215 (Wiley, 1980).
20. Dandekar, T., Schuster, S., Snel, B., Huynen, M. & Bork, P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**, 115–124 (1999).
21. Liao, J. C., Hou, S. Y. & Chao, Y. P. Pathway analysis, engineering and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.* **52**, 129–140 (1996).
22. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
23. Edwards, J. S. & Palsson, B. O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416 (1999).
24. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotech.* **19**, 125–130 (2001).
25. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
26. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193 (2002).
27. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**, 186–189 (2002).
28. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N. & Barabasi, A. L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
29. Papp, B., Pal, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661–664 (2004).
30. Pal, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genet.* **37**, 1372–1375 (2005).
31. Hyduke, D. R., Lewis, N. E. & Palsson, B. O. Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst* **9**, 167–174 (2013).
   This is a review of techniques to integrate omic data with CBMs.
32. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA* **102**, 2685–2689 (2005).
33. Kharchenko, P., Church, G. M. & Vitkup, D. Expression dynamics of a cellular metabolic network. *Mol Syst Biol* **1**, 2005.0016 (2005).
34. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nature Biotech.* **26**, 1003–1010 (2008).
35. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4**, e1000082 (2008).
36. Carlson, R. & Srienc, F. Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: creation of overall flux states. *Biotechnol. Bioeng.* **86**, 149–162 (2004).
37. Carlson, R. & Srienc, F. Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: identification of reactions. *Biotechnol. Bioeng.* **85**, 1–19 (2004).
38. Harcombe, W. R., Delaney, N. F., Leiby, N., Klitgord, N. & Marx, C. J. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLoS Comput. Biol.* **9**, e1003091 (2013).
39. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol.* **3**, 119 (2007).
40. Molenaar, D., van Berlo, R., de Ridder, D. & Teusink, B. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**, 323 (2009).
41. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336**, 601–604 (2012).
42. Lewis, N. E. *et al.* Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
43. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
44. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337**, 1101–1104 (2012).
45. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3**, 121 (2007).
46. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* **2**, 2006.0008 (2006).
47. Scheer, M. *et al.* BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **39**, D670–D676 (2011).
48. Lobel, L., Sigal, N., Borovok, I., Ruppin, E. & Herskovits, A. A. Integrative genomic analysis identifies isoleucine and CodY as regulators of *Listeria monocytogenes* virulence. *PLoS Genet.* **8**, e1002887 (2012).
49. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
50. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
51. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* **39**, D98–D105 (2011).
52. Segre, D., DeLuna, A., Church, G. M. & Kishnoy, R. Modular epistasis in yeast metabolism. *Nature Genet.* **37**, 77–83 (2005).
53. Harrison, R., Papp, B., Pal, C., Oliver, S. G. & Delneri, D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl Acad. Sci. USA* **104**, 2307–2312 (2007).
54. He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nature Genet.* **42**, 272–276 (2010).
55. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genet.* **43**, 656–662 (2011).
56. Mo, M. L., Palsson, B. O. & Herrgard, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).
57. Wessely, F. *et al.* Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol. Syst. Biol.* **7**, 515 (2011).
58. Notebaart, R. A., Teusink, B., Siezen, R. J. & Papp, B. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput. Biol.* **4**, e26 (2008).
59. Kaleta, C., de Figueiredo, L. F. & Schuster, S. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.* **19**, 1872–1883 (2009).
60. Faith, J. J. *et al.* Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* **36**, D866–D870 (2008).
61. Orth, J. D. & Palsson, B. O. Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* **107**, 403–412 (2010).
   This is a review on techniques and applications of CBMs for a targeted expansion of biochemical knowledge.
62. Reed, J. L. *et al.* Systems approach to refining genome annotation. *Proc. Natl Acad. Sci. USA* **103**, 17480–17484 (2006).
63. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* **104**, 1777–1782 (2007).
64. Rolfsson, O., Paglia, G., Magnusdottir, M., Palsson, B. O. & Thiele, I. Inferring the metabolism of human orphan metabolites from their metabolic network context affirms human gluconokinase activity. *Biochem. J.* **449**, 427–435 (2013).
65. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
66. Nakahigashi, K. *et al.* Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.* **5**, 306 (2009).
67. Lee, S. Y., Lee, D. Y. & Kim, T. Y. Systems biotechnology for strain improvement. *Trends Biotechnol.* **23**, 349–358 (2005).
68. Park, J. H. & Lee, S. Y. Towards systems metabolic engineering of microorganisms for amino acid production. *Curr. Opin. Biotechnol.* **19**, 454–460 (2008).
   This is a review of using systems biology methodologies for metabolic engineering applications.
69. Caspeta, L. & Nielsen, J. Economic and environmental impacts of microbial biodiesel. *Nature Biotech.* **31**, 789–793 (2013).
70. Yim, H. *et al.* Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chem. Biol.* **7**, 445–452 (2011).

71. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
72. Constantinou, L. & Gani, R. New group-contribution method for estimating properties of pure compounds. *AIChE J.* **40**, 1697–1710 (1994).
73. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
74. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657 (2003).
75. Oberhardt, M. A., Yizhak, K. & Ruppin, E. Metabolically re-modeling the drug pipeline. *Curr. Opin. Pharmacol.* **13**, 778–785 (2013). **This is a review on using constraint-based modelling for drug discovery.**
76. Hsu, P. P. & Sabatini, D. M. Cancer cell metabolism: Warburg and beyond. *Cell* **134**, 703–707 (2008).
77. Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, 501 (2011).
78. Frezza, C. *et al.* Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* **477**, 225–228 (2011).
79. Jerby, L., Shlomi, T. & Ruppin, E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.* **6**, 401 (2010).
80. Kim, P. J. *et al.* Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc. Natl Acad. Sci. USA* **104**, 13638–13642 (2007).
81. Kim, H. U. *et al.* Integrative genome-scale metabolic analysis of *Vibrio vulnificus* for drug targeting and discovery. *Mol. Syst. Biol.* **7**, 460 (2011).
82. Brynildsen, M. P., Winkler, J. A., Spina, C. S., MacDonald, I. C. & Collins, J. J. Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. *Nature Biotech.* **31**, 160–165 (2013).
83. Lerman, J. A. *et al. In silico* method for modelling metabolism and gene product expression at genome scale. *Nature Commun.* **3**, 929 (2012).
84. Zhang, Y. *et al.* Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* **325**, 1544–1549 (2009).
85. Thiele, I., Fleming, R. M., Bordbar, A., Schellenberger, J. & Palsson, B. O. Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional and translational machinery. *Biophys. J.* **98**, 2072–2081 (2010).
86. Pramanik, J. & Keasling, J. D. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol. Bioeng.* **60**, 230–238 (1998).
87. Rodionova, I. A. *et al.* Diversity and versatility of the *Thermotoga maritima* sugar kinome. *J. Bacteriol.* **194**, 5552–5563 (2012).
88. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).
89. Chandrasekaran, S. & Price, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis. Proc. Natl Acad. Sci. USA* **107**, 17845–17850 (2010).
90. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
91. Chang, R. L. *et al.* Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli. Science* **340**, 1220–1223 (2013).
92. Gu, J. & Bourne, P. E. *Structural bioinformatics* (Wiley-Blackwell, 2009).
93. Marr, A. G. & Ingraham, J. L. Effect of temperature on the composition of fatty acids in *Escherichia coli*. *J. Bacteriol.* **84**, 1260–1267 (1962).
94. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
95. Mörters, P., Peres, Y., Schramm, O. & Werner, W. *Brownian motion* (Cambridge Univ. Press, 2010).
96. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
97. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nature Biotech.* **31**, 419–425 (2013).
98. Borenstein, E. Computational systems biology and *in silico* modeling of the human microbiome. *Brief Bioinform.* **13**, 769–780 (2012).
99. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl Acad. Sci. USA* **110**, 12804–12809 (2013).
100. Atkinson, D. E. The energy charge of the adenylate pool as a regulatory parameter. Interaction with feedback modifiers. *Biochemistry* **7**, 4030–4034 (1968).
101. Weisz, P. B. Diffusion and chemical transformation. *Science* **179**, 433–440 (1973).
102. Reed, J. L. Shrinking the metabolic solution space using experimental datasets. *PLoS Comput. Biol.* **8**, e1002662 (2012). **This is a review of the potential constraints that have been placed on CBMs.**
103. Colijn, C. *et al.* Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput. Biol.* **5**, e1000489 (2009).
104. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nature Biotech.* **28**, 245–248 (2010). **This paper presents a primer on the theory, applications and software toolboxes for FBA.**
105. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
106. Wilkinson, D. J. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Rev. Genet.* **10**, 122–133 (2009).
107. Steuer, R. Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry* **68**, 2139–2151 (2007).
108. de Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103 (2002).
109. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Computat. Biol.* **7**, 601–620 (2000).
110. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–690 (2009).
111. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
112. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genet.* **44**, 841–847 (2012).

Acknowledgements

The authors thank D. Zielinski, J. Lerman, N. E. Lewis and H. Nagarajan for their criticisms and comments on the manuscript. This work was supported by the US National Institutes of Health grants GM068837 and GM057089, and by the Novo Nordisk Foundation. Z.A.K. is supported through the US National Science Foundation Graduate Research Fellowship (DGE-1144086).

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

BiGG database: http://bigg.ucsd.edu/

Literature on COBRA (constraint-based reconstruction and analysis) methods: http://sbrg.ucsd.edu/cobra-methods

Literature on model-driven analysis: http://sbrg.ucsd.edu/cobra-predictions

OpenCOBRA project: http://opencobra.sourceforge.net/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF