Reconstructing organisms in silico: genome-scale models and their emerging applications

Xin Fang[1], Colton J. Lloyd[1] and Bernhard O. Palsson[1,2,3] ✉

Abstract | *Escherichia coli* is considered to be the best-known microorganism given the large number of published studies detailing its genes, its genome and the biochemical functions of its molecular components. This vast literature has been systematically assembled into a reconstruction of the biochemical reaction networks that underlie *E. coli*'s functions, a process which is now being applied to an increasing number of microorganisms. Genome-scale reconstructed networks are organized and systematized knowledge bases that have multiple uses, including conversion into computational models that interpret and predict phenotypic states and the consequences of environmental and genetic perturbations. These genome-scale models (GEMs) now enable us to develop pan-genome analyses that provide mechanistic insights, detail the selection pressures on proteome allocation and address stress phenotypes. In this Review, we first discuss the overall development of GEMs and their applications. Next, we review the evolution of the most complete GEM that has been developed to date: the *E. coli* GEM. Finally, we explore three emerging areas in genome-scale modelling of microbial phenotypes: collections of strain-specific models, metabolic and macromolecular expression models, and simulation of stress responses.

Python
An interpreted, general-purpose programming language that is widely used in computational biology.

Genome-scale network reconstructions are built from curated and systematized knowledge[1,2] that enables them to quantitatively describe genotype–phenotype relationships. Genome-scale models (GEMs) are mathematical representations of reconstructed networks that facilitate computation and prediction of multiscale phenotypes through the optimization of an objective function of interest[3,4].

The development of a GEM requires curated metabolic knowledge bases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[5], and an annotated genome sequence of the organism of interest. By mapping the annotated genome sequence (FIG. 1a) to the knowledge base, one can reconstruct a metabolic network composed of all known metabolic reactions (FIG. 1b). This metabolic network can be converted into a mathematical format — a stoichiometric matrix (S matrix) — where the columns represent reactions, rows represent metabolites and each entry is the corresponding coefficient of a particular metabolite in a reaction (FIG. 1c). A cellular objective is needed to enable computation of a feasible metabolic flux that optimizes the model objective. A widely used objective function is to optimize growth rate, represented by a biomass function[6], composed of essential metabolites needed for growth. The detailed steps to reconstruct a GEM have been described in a formal protocol[1].

Flux balance analysis (FBA) is the most widely used[7] approach to characterize GEMs. GEMs can simulate metabolic flux states of the reconstructed network while incorporating multiple constraints to ensure the solution identified by FBA is physiologically relevant and compliant with governing constraints, such as the metabolic network topology represented by the S matrix, a steady-state assumption (that is, the internal metabolites must be produced and consumed in a flux-balanced manner), and other limits on nutrient uptake rates, enzyme capacities and protein/gene expression profiles. The S matrix and the objective function define a system of linear equations that can be solved given the imposed constraints, resulting in a solution space (that is, a space where all feasible phenotypic states exist) (FIG. 1d,e). FBA can identify a single optimal flux distribution or multiple optimal flux distributions that optimize the objective function in the solution space (FIG. 1f). FBA and many other GEM analysis methods are available through COBRApy[8] in Python or the COBRA Toolbox in MATLAB[9].

GEMs have been successfully implemented for a wide range of applications[10–17], including understanding microorganisms[16–22], metabolic engineering[23–27,23–26,28], drug development[29], prediction of enzyme functions[30], and understanding microbial community interactions[31–40]

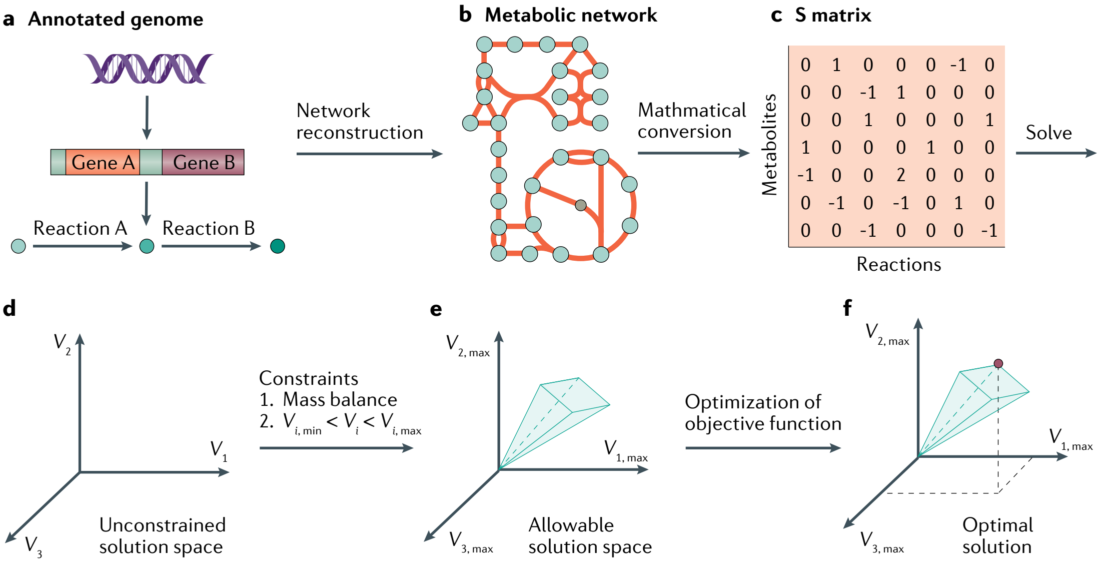
[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA.

[2]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA.

[3]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark.

✉e-mail: palsson@ucsd.edu

https://doi.org/10.1038/s41579-020-00440-4



0	1	0	0	0	-1	0
0	0	-1	1	0	0	0
0	0	1	0	0	0	1
1	0	0	0	1	0	0
-1	0	0	2	0	0	0
0	-1	0	-1	0	1	0
0	0	-1	0	0	0	-1

Fig. 1 | **Basic principles of constraint-based modelling of cellular functions. a** | Metabolic genes from annotated genomes of interest and metabolic knowledge lead to metabolic reactions. **b** | Integration of all the metabolic reactions through shared metabolites results in the construction of a metabolic network for the organism of interest. **c** | The metabolic network can be converted into a stoichiometric matrix (S matrix), where rows represent metabolites, columns represent reactions and each entry represents the reaction coefficient of a particular metabolite in a reaction. **d** | With the S matrix and the objective function of the model, one can solve for the flux distributions. The solution space is where all possible solutions of flux distribution reside, and each axis represents the metabolic flux of a reaction. **e** | Applying additional constraints will shrink the allowable solution space. Commonly used constraints include the steady-state assumption and feasible ranges of metabolic flux. **f** | One or multiple optimal solutions can be found in the allowable solution space that optimizes the objective function of the model (as represented by the red dot).

and human disease[41,42]. One metabolic engineering application focuses on suggesting gene deletion strategies to enable overproduction of a metabolite of interest[24]. The algorithm OptKnock uses GEMs to identify gene deletion combinations that ensure the metabolite of interest becomes an obligatory metabolic by-product of growth (known as growth-coupled production). This framework was applied to succinate, lactate and 1,3-propanediol production in *Escherichia coli*[24]. OptKnock combined with the *E. coli* GEM proposed gene knockout strategies that are in good agreement with published mutant strains in the literature, highlighting its potential in strain design[24]. GEMs have also been developed to study cancer metabolism. One study[29] used the GEM of cancer metabolism to predict potential drug targets by simulating gene knockdowns, evaluating the damage on ATP production and assigning cytostatic scores for genes. The model predicted 52 cytostatic drug targets, of which 40% are already targeted by known cancer treatments, leaving the rest as potential new drug targets.

Since the development of the first GEM for *Haemophilus influenzae*[43], the field has advanced substantially with a rapid rise in the number of GEMs built[14,44]. The number of tools and methods involved in network reconstruction and analysis has also bloomed, which accelerated the model-building process[45] and enabled numerous uses of GEMs[4] (BOX 1). As of 2019, GEMs had been generated for more than 6,000 sequenced genomes either manually or through automatic GEM reconstruction tools[45], covering bacteria, archaea and eukaryotes.

In addition to the well-developed uses of GEMs, recent explorations of new applications have emerged. In this Review, we describe the ongoing efforts in reconstruction to increase the coverage of the tree of life by GEMs, describe the expansion in the scope and applications of GEMs as illustrated by the example of *E. coli* and elaborate on three emerging areas where great potential exists: multistrain analysis using strain-specific GEMs; the incorporation of macromolecular expression pathways into existing models of metabolism to form metabolic and macromolecular expression (ME) models; and prediction of complex phenotypes, such as stress responses. We foresee the continual development and implementation of GEMs for many more organisms of interest, and their becoming an essential tool for synthetic genome engineering.

Growth of genome-scale reconstructions

Extensive effort has focused on reconstructing metabolic networks for a broad range of organisms. GEM development was initiated for bacteria and has gradually extended to archaea[46–48] and eukaryotes[49], including yeast[50], plants[51–53] and humans[54–56].

Exponentially growing numbers of genome sequences (FIG. 2a) enable the construction of a knowledge base of reactions and metabolites[57] and the generation of increasing strain-specific network reconstructions. As the manual reconstruction of genome-scale networks is laborious and time-consuming, many automated network reconstruction tools have been developed to accelerate the reconstruction process, including

Reactomes
All the reactions involved in genome-scale models (or a certain model of interest). Each base unit is a reaction, and the entities are metabolites involved in the reactions, such as proteins, nucleic acids and small molecules.

ModelSeed[58], CarveMe[59], RAVEN[60,61] and kbase[62]. According to a summary from 2019, around 5,897 bacteria, 127 archaea and 215 eukaryote metabolic network reconstructions have been reported[14]. Many of them can be found in GEM databases, including BiGG Models[63], BioModels[64], MetaNetX[65], MEMOSys[66] and Virtual Metabolic Human[67]. However, most of these reconstructions lacked manual refinement, which may result in an inaccurate description of the organism and unreliable predictions of the model[14]. Therefore, the community developed reconstruction and GEM quality standards (MEMOTE)[68] to provide an overall evaluation of the quality of a reconstruction and limitations on its use.

Of the published reconstruction models, we focus on 108 models deposited in BiGG Models[63], a widely used repository for high-quality GEMs, where all models have been benchmarked against MEMOTE[68]. The content of these curated GEMs is detailed in FIG. 2. As with the availability of genome sequences, the numbers of reactions and metabolites accounted for in curated models continue to grow (FIG. 2b). Particularly, we observe a rapid increase in both the number of reactions and the number of metabolites since 2015, due to the development of models for eukaryotes, cyanobacteria and species belonging to the phyla Firmicutes and Actinobacteria.

Despite the substantial growth in the number of network reconstructions, their coverage of the tree of life is still limited. A multiple correspondence analysis (MCA; the counterpart of principal component analysis for categorical data) of the reactomes of 108 GEMs (FIG. 2c) showed that the clustering of the models by their metabolic functions is strongly related to their phylogeny. MCA also suggested that the differences among prokaryotic models are relatively small. By overlaying the 108 models on the tree of life (FIG. 2d), we observed results similar to those obtained by MCA performed in 2014 (REF.[44]), namely that network reconstruction efforts have been mainly focused on Proteobacteria, leaving many other phylogenetic branches without any available reconstructions. Although this observation is only based on the 108 models in BiGG Models, it is clear that the development of GEMs for less studied organisms may greatly expand the coverage of metabolic pathways and the ‘reactome’ represented by curated GEMs (FIG. 2b). A large-scale effort is needed to establish a global metabolic atlas, with ‘global’ referring to the tree of life.

Evolution of the *E. coli* GEM

The serial development of *E. coli* metabolic reconstructions has led to the expansion in the scope and applications of GEMs. FIGURE 3 depicts the iterations[69–77] of the *E. coli* GEMs published since 2000 and the changes in the model content. In this section, we focus on the development of metabolic models (M models); ME models shown in FIG. 3 are discussed in later sections. The first two reconstructions (not shown in FIG. 3) were developed before the *E. coli* genome was sequenced and were based solely on biochemical knowledge. After the genomic sequence of *E. coli* K-12 MG1655 was established in 1997 (REF.[78]), its annotation and new discoveries of metabolic functions led to a series of genome-scale reconstructions of ever-increasing scope and content.

The latest *E. coli* model, iML1515, now includes 1,515 genes[76]. iML1515 has comprehensive coverage of metabolic functions integrated with protein structural information, enabling growth simulation on different nutrients for strains of interest as well as an evaluation of mutational impact across strains using structural biology methods[76,79,80]. iML1515 was used to simulate gene knockouts on 16 different carbon sources and predicted gene essentiality across conditions with an accuracy of 93.4% compared with experimental data, highlighting the potential to identify drug targets using GEMs of pathogenic organisms.

In addition, iML1515 was used to analyse transcriptomics data from 333 experiments with various conditions and provided valuable insight into transcriptional variation across conditions. For example, the three isozymes of aspartate kinase (*lysC*, *metL*, and *thrA*) have variable expression across conditions. iML1515 simulation suggests that when only *lysC* is expressed, *E. coli* is unable to synthesize L-threonine, L-methionine, L-isoleucine, biotin and adenosylmethionine biomass components, which explains why *lysC* is preferentially expressed in nutrient-rich conditions when these metabolites are available.

Of the 4,623 open reading frames annotated on the *E. coli* K-12 MG1655 genome sequence, 1,600 are of unknown function (the so-called *y* genes)[81], leaving 3,023 genes of known function on which to base a reconstruction. With the 1,515 genes in the latest metabolic

Box 1 | **Why build computational models?**

Computational models describe a system through a mathematical formalism enabling the study of its behaviour through simulation. Models are prevalent in the physical sciences, but are less common in biology. The motivation for building models can be broken down into five categories[133].

Organize disparate information into a coherent whole
Network reconstructions represent a formal organization of knowledge that can subsequently be converted into computational models. Genome-scale models (GEMs) enable systems-level understanding and analysis, and produce predictions based on the scope, coverage and quality of the underlying reconstruction[44,68].

Identify important components and interactions in a complex system
An early use of GEMs was to compute gene essentiality[72,134]. For a poorly characterized organism, *Geobacter sulfurreducens*, GEMs produced a deep understanding of acetate uptake, acetate activation and altered amino acid metabolism[135].

Make new discoveries
GEMs can be used to simulate perturbation to a metabolic system to identify essential metabolites and to find its structural analogues as candidate drugs that inhibit the enzymes that degrade the metabolite[136]. GEMs have enabled designs of growth-coupled methylation systems[137].

Fill in knowledge gaps
GEM prediction of ‘no growth’ under a condition where the organism experimentally grows is called a ‘false negative’ prediction, which is usually a result of a missing component in the GEM. Gap-filling procedures[138–140] and other methods[141] were developed to address this issue, driving discoveries and making important corrections to conventional wisdom.

Understand the essential and qualitative features
Qualitative features are important for complex systems. For example, global proteomics data and GEMs helped identify the fear–greed trade-off in *Escherichia coli* growth[125]. *E. coli* was shown to have nearly half of the proteome mass unused in certain environments. This ‘unused’ proteome is involved in nutrient- and stress-preparedness functions that may convey fitness benefits in changing environments.

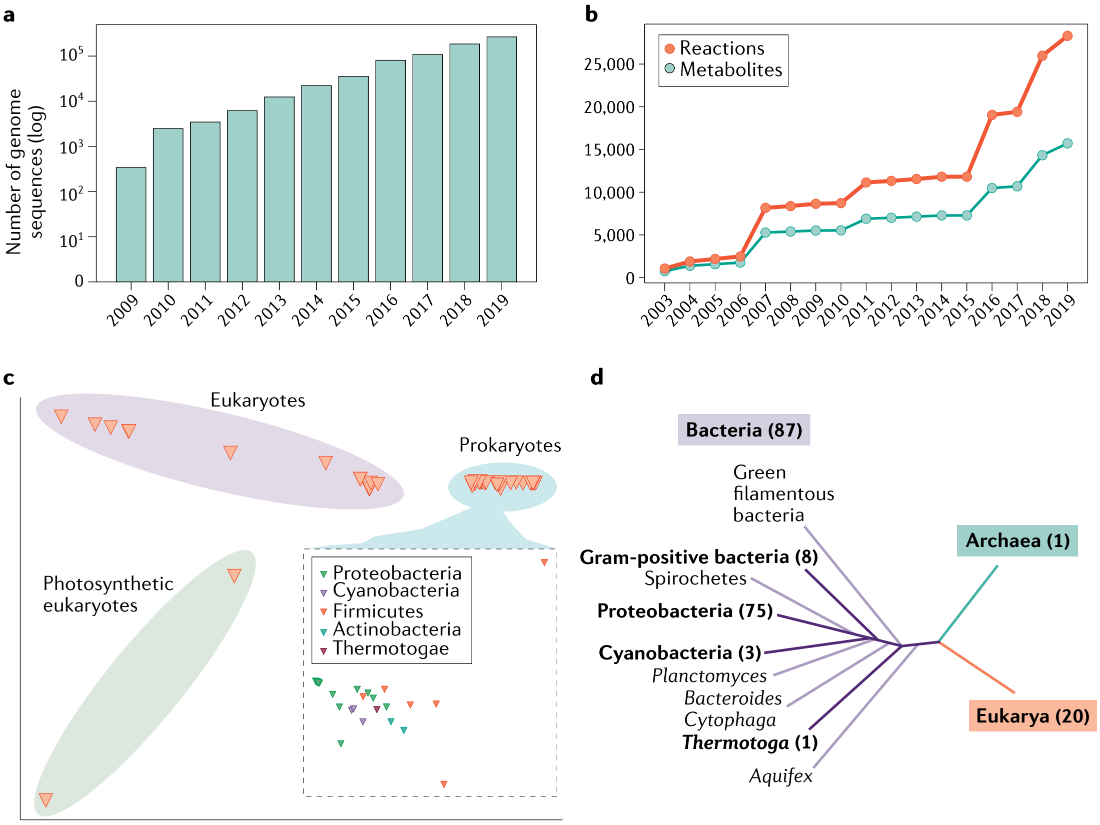


Fig. 2 | **The increasing number of genome sequences and the development of genome-scale models. a** | Number of public genome sequences in the PATRIC database[132]. **b** | Number of reactions and metabolites represented in 108 manually curated models in the BiGG Models database[63]. **c** | Multiple correspondence analysis of the reactomes of the 108 reconstructions. **d** | Coverage of the 108 reconstructions in the tree of life. The number in parentheses represents the number of reconstructions in each branch. Groups with at least one metabolic reconstruction are bolded.

network reconstruction, ~50% of the functionally annotated genes are accounted for. The known biochemical functions of the corresponding gene products can now be computationally assessed in the context of the function of all the other gene products. This coverage forms the genetic and biochemical basis for the metabolic systems biology of *E. coli*.

Thus, the scope of GEM applications has increased with the expansion of metabolic coverage. The early models were used to compute basic phenotypes such as growth rate, by-product secretion and yield of cofactors. Other applications of *E. coli* GEMs have been reviewed elsewhere[10,11]. The most recent GEMs now enable applications such as pan-genome analysis, computation of proteome allocation and the simulation of various stress responses, which we discuss in detail next.

Emerging applications of GEMs

The availability of genome-scale multi-omics data sets is growing rapidly, including whole genome sequences, transcriptomics, proteomics and metabolomics data. This calls for the development of tools to interpret and contextualize such data sets. Therefore, to enable direct integration of such data with GEMs, recent model development introduced macromolecular expression into the M models to produce ME models, which allow direct comparison between the simulation and experimental data. Additionally, earlier GEMs were usually developed on the basis of a representative strain from a species, but the availability of multiple genome sequences within a species allows us to develop strain-specific GEMs to explore variation across strains.

In this section, we discuss three new directions in the development of GEMs and their emerging applications; multistrain analyses that enable investigations into strain-specific variation; ME models that can compute proteome allocation; and simulation of stress responses that facilitate an understanding of complex phenotypes. Other directions in GEM development that have been addressed in other reviews include but are not limited to the integration of GEMs with structural biology[82], modelling of complex communities such as the microbiota[83], and tissue- or cell-specific models constrained by multi-omics data[84].

Multistrain analysis

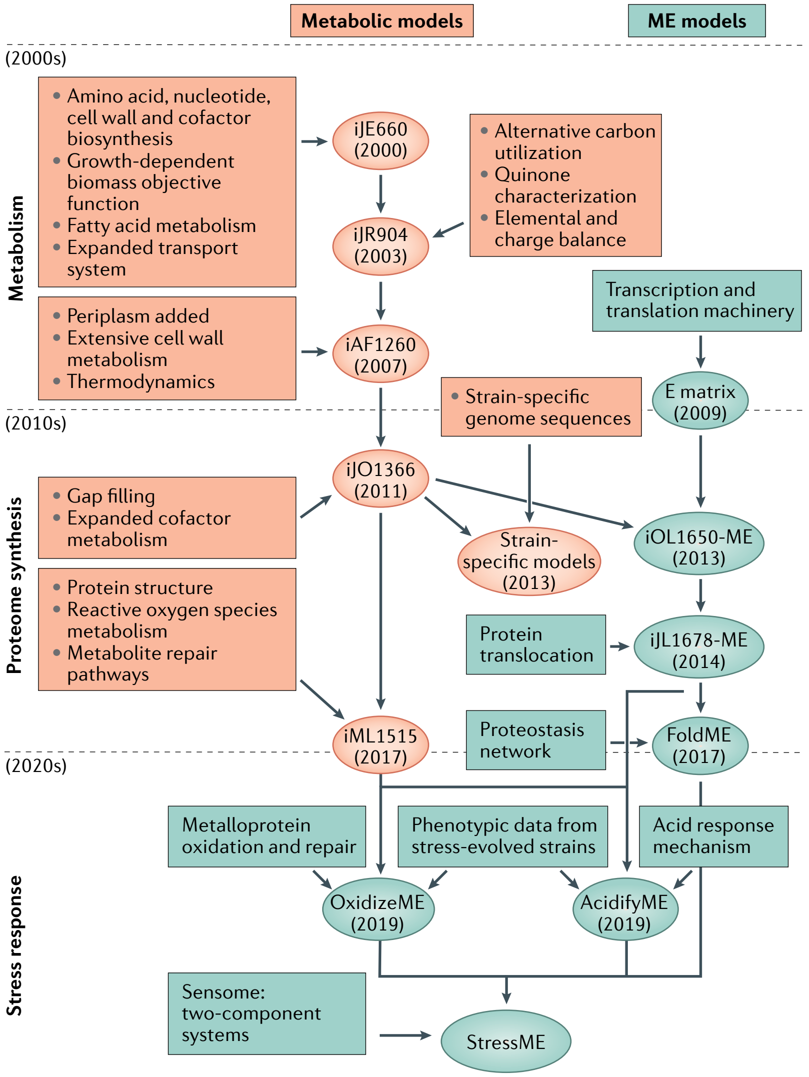
With the ever-increasing number of genome sequences, it has become clear that large variations exist in the gene portfolio across strains of a species. In 2005, the concept of the pan-genome — the total list of genes found in all sequenced genomes of strains belonging to a species — was introduced. The pan-genome is composed of a core genome (that is, genes shared by all strains within a species) and an accessory genome (that is, genes present

Proteome allocation
The partition of proteomics resources into different functions to fulfil the organism's need under the given condition.

Sensome
The components (such as genes and proteins) in an organism or cell that are involved in sensing the changes in the environment.

in only a subset of strains)[85]. Although some species have relatively conserved gene portfolios (known as a closed pan-genome), other species have substantial variability in strain-specific gene portfolios (known as an open pan-genome).

E. coli was shown to have substantial differences in gene portfolios across strains, with as little as ~20% of the total number of genes annotated being shared across the sequenced strains[86]. The diversity in gene portfolios is thought to be a reflection of adaptation to different microenvironments. Many other microorganisms share this characteristic, including *Salmonella* spp.[87], *Staphylococcus aureus*[88] and *Klebsiella pneumoniae*[89]. It has become clear that it is important to understand the broad range of metabolic capabilities encoded by accessory genes as they could potentially contribute to the pathogenicity and interactions with a human host[90].



Metabolic models
ME models
(2000s)
(2010s)
(2020s)
Metabolism
Proteome synthesis
Stress response
• Amino acid, nucleotide, cell wall and cofactor biosynthesis
• Growth-dependent biomass objective function
• Fatty acid metabolism
• Expanded transport system
iJE660 (2000)
• Alternative carbon utilization
• Quinone characterization
• Elemental and charge balance
iJR904 (2003)
• Periplasm added
• Extensive cell wall metabolism
• Thermodynamics
iAF1260 (2007)
Transcription and translation machinery
E matrix (2009)
• Strain-specific genome sequences
iJO1366 (2011)
• Gap filling
• Expanded cofactor metabolism
Strain-specific models (2013)
iOL1650-ME (2013)
• Protein structure
• Reactive oxygen species metabolism
• Metabolite repair pathways
Protein translocation
iJL1678-ME (2014)
iML1515 (2017)
Proteostasis network
FoldME (2017)
Metalloprotein oxidation and repair
Phenotypic data from stress-evolved strains
Acid response mechanism
OxidizeME (2019)
AcidifyME (2019)
Sensome: two-component systems
StressME

Fig. 3 | **Historical development of *E. coli* genome-scale models.** Development of existing and potential future genome-scale models (both metabolic models, shown in orange, and metabolic and macromolecular expression (ME) models, shown in blue) of *Escherichia coli*. The genome-scale metabolic model of *E. coli* first appeared in the early years of the first decade of the twenty-first century (2000s). An increasing scope of biological functions has been incorporated into the model, leading to various generations of the metabolic models as new discoveries were made. In the early 2010s, ME models that incorporate transcription and translation mechanisms emerged. Multiple efforts followed to improve and expand the ME model. Going into the 2020s, extensions of stress response modules have been added to ME models. Future directions involve incorporation of the sensome to form the StressMe model, and the inclusion of toxins, biosynthetic gene clusters and cell cycle. Ovals indicate models, and boxes represent data incorporated to generate the models. According to the naming convention for network reconstructions, model names consist of an *i* for 'in silico' followed by the initials of the person(s) who built the model, and the number of open reading frames accounted for in the reconstruction. E matrix, expression matrix.

'Pan-genome analysis' typically refers to comparative analysis of genes across strains. Building GEMs for many strains offers a much deeper analysis based on all the mechanisms that GEMs contain for metabolic processes. The workflow to generate strain-specific models is illustrated in the left panel in FIG. 4. Genomes of strains of interest are mapped to a curated reference reconstruction to generate a homology matrix, which is used to guide the deletion of genes and reactions from the reference model to create draft models. Manual curations are needed to finalize strain-specific GEMs (FIG. 4). The first multistrain GEM studies from 2013 established GEMs for a set of 55 *E. coli* and *Shigella* spp. strains[91]. By simulating growth capabilities on different nutrient resources, the study predicted strain-specific auxotrophies and unique metabolic capabilities that correspond to their pathotypes and colonization sites. The simulated growth phenotypes separated the strains on the basis of their pathotypes, as most commensal strains were unable to grow on a set of nutrients, such as *N*-acetyl-D-galatosamine, which supports growth for 100% of extraintestinal pathogenic strains. In addition, 12 of the 55 strains were predicted to be unable to produce at least one essential biomass component, including folate, thiamin and amino acids from glucose M9 minimal medium, some of which are confirmed in the literature.

More recent pan-genome studies of *E. coli* explored the linkage between metabolism and health outcomes. A study of metabolic capabilities of clinical isolates of *E. coli* strains from individuals with inflammatory bowel disease (IBD)[92] (FIG. 4) compared growth simulation of strain-specific models of clinical isolates and commensal strains, and identified a pathway specific to strains from the B2 phylogroup that are prevalent in individuals with IBD. This pathway is involved in metabolizing the mucus glycan through the action of tagatose bisphosphate aldolase, which potentially aids *E. coli* strains in the colonization of intestinal mucosa[92].

In a separate study using time series metagenomics data from an individual with IBD[93] (FIG. 4), we found multiple *E. coli* strains dominating the microbiota at different time points as the inflammation level varied. Strain-specific GEMs were reconstructed for each strain, and the metabolic capabilities delineated by strain-specific GEMs were vastly different across these dominant strains. The models suggest that the strain extracted during the peak inflammation is the most similar to known representative pathogenic strains in IBD, whereas dominant strains extracted from low-inflammation time points were more similar to commensal strains. Specifically, the dominant strain present during peak inflammation and known

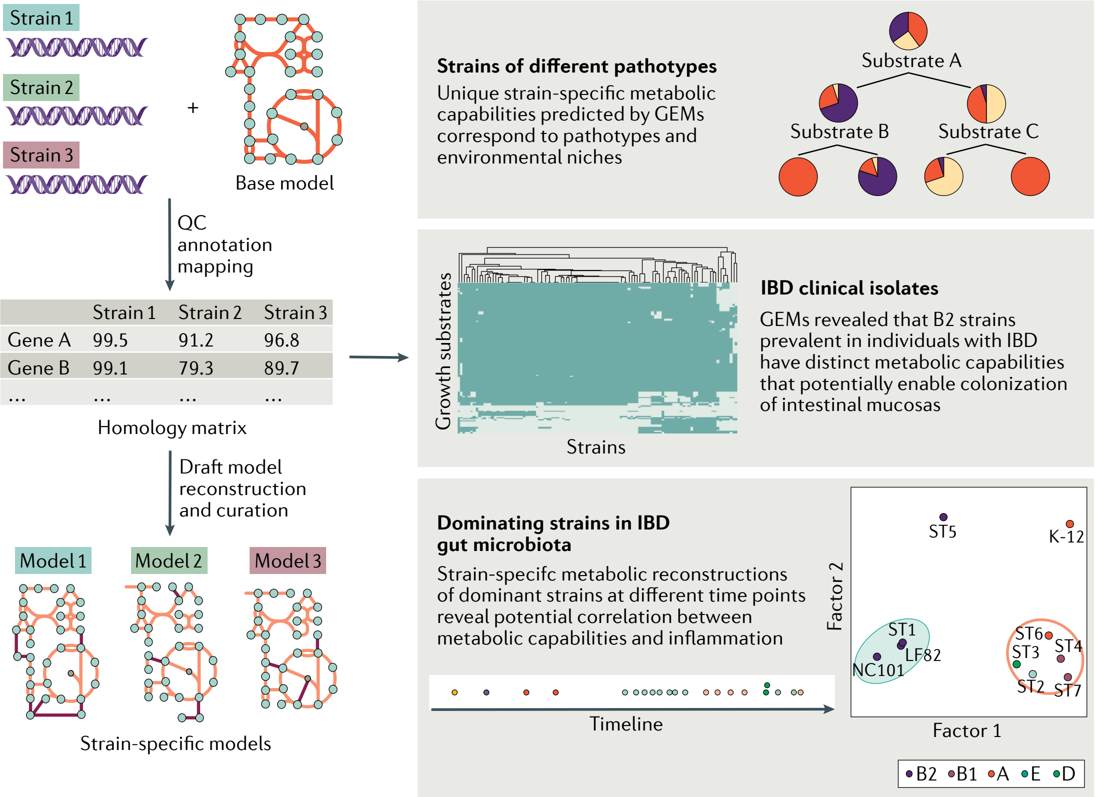

Strain 1
Strain 2
Strain 3
+
Base model
QC
annotation
mapping
Strain 1 Strain 2 Strain 3
Gene A 99.5 91.2 96.8
Gene B 99.1 79.3 89.7
… … … …
Homology matrix
Draft model
reconstruction
and curation
Model 1
Model 2
Model 3
Strain-specific models
Strains of different pathotypes
Unique strain-specific metabolic capabilities predicted by GEMs correspond to pathotypes and environmental niches
Substrate A
Substrate B
Substrate C
Growth substrates
Strains
IBD clinical isolates
GEMs revealed that B2 strains prevalent in individuals with IBD have distinct metabolic capabilities that potentially enable colonization of intestinal mucosas
Dominating strains in IBD gut microbiota
Strain-specifc metabolic reconstructions of dominant strains at different time points reveal potential correlation between metabolic capabilities and inflammation
Timeline
Factor 2
Factor 1
ST5
K-12
ST1
LF82
NC101
ST6
ST3
ST4
ST2
ST7
B2
B1
A
E
D

Fig. 4 | **Generation of strain-specific *E. coli* genome-scale models and their application in multistrain studies.** Strain-specific models were generated from genome sequences of strains of interest and a curated reference model. The annotated genome sequences of target strains are mapped to the reference genome sequence to generate the homology matrix that delineates the gene sequence similarity across strains. The homology matrix can be used to create draft models of target strains. These models can then be finalized by manual curation. Strain-specific models were used to reveal variation in metabolic capabilities across different pathotypes, as illustrated in three studies shown on the right. The first multistrain study of *Escherichia coli* genome-scale models (GEMs) found metabolic capabilities predicted by GEMs correspond to the pathotype and environment. In the second study, comparison of GEMs constructed for inflammatory bowel disease (IBD) clinical isolates suggested a possible link between metabolic functions of B2 strains and their prevalence in individuals with IBD. Lastly, GEMs of dominant strains in an individual with IBD revealed the potential correlation between metabolism and inflammation[91–93]. QC, quality control. Panel 'Strains of different pathotypes' adapted with permission from REF.[91], PNAS. Panel 'IBD clinical isolates' adapted from REF.[92], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Panel 'Dominating strains in IBD gut microbiota' adapted from REF.[93], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

pathogenic strains were predicted to share the capability to grow in a set of substrates, including cellobiose, deoxyribose and monosaccharides derived from intestinal mucosa, suggesting that strain-specific features are potentially linked to pathogenicity and disease progression.

The application of GEMs for pan-genome analysis is not limited to *E. coli*. Great potential exists for use of GEMs to study pathogens to understand strain-specific features and their association with colonization sites, pathogenicity and antibiotic resistance and their impact on human health. Several published studies have already used strain-specific GEMs to further understand strain-specific characteristics of various microorganisms.

Salmonella spp. were shown to have serovar-specific metabolic traits, including auxotrophies and catabolic pathways that may be associated with adaptations to their colonization sites[87]. The metabolic capabilities of *S. aureus* were found to link to pathogenic traits and virulence acquisitions, which can then be used to classify mild versus severe infections[94]. For example, two *S. aureus* USA300 isolates were predicted to be the only strains capable of using spermidine as a sole source of carbon and nitrogen[94]. Spermidine is produced in areas of inflammation and wound healing[95], which give these strains the opportunity to cause skin infection. A study of *K. pneumoniae* strains with antibiotic resistance phenotypes suggested differential utilization of nitrogen sources may help discriminate between antibiotic resistance phenotypes[96]. Similar studies have also been performed for other species: strain-specific *Acinetobacter baumannii* GEMs[97] revealed the significant variation in lipopolysaccharide across strains; GEMs of *Leptospira* spp. delineated the differences in lysine metabolism between pathogenic and commensal *Leptospira* spp.[98]; and *Pseudomonas putida* strain-specific models reflected the diverse metabolic capabilities across strains due to variations in environmental niches[99].

For a large number of sequenced genomes (more than 1,000 strains), it has been shown that the gene portfolio of individual strains can be characterized not only in terms of the presence or absence of a gene but also in terms of the particular allele of the gene. Thus, a field of 'alleleomics' may have emerged. Alleleomic analysis was shown to be valuable for studying organisms with closed pan-genomes. Using a GEM-based machine learning classifier, one study[100] was able to predict antimicrobial resistance in *Mycobacterium tuberculosis*, while enabling a biochemical interpretation of the genotype–phenotype map. Specifically, through investigation of key flux states discriminating between *M. tuberculosis* strains that are resistant to pyrazinamide and *M. tuberculosis* strains that are susceptible to pyrazinamide, the study authors correctly identified *pncA* and *ppsA* alleles as major genetic determinants, which had been reported in the literature, and proposed a new hypothesis that *ansP2* mutants may potentially contribute to resistance through L-aspartate-based modulation of the coenzyme A pool.

A semi-automated protocol for generating strain-specific models from a collection of strain-specific genome sequences has been made available[2] to aid researchers in reconstructing and utilizing the strain-specific GEMs. This protocol details the major stages involved in strain-specific model generation and curation, accompanied by easy-to-follow tutorials in Python notebooks to ensure strain-specific GEMs are accessible to researchers interested in applying them to different organisms. The protocol does not require advanced coding skills.

The ME model

The demonstrated predictive ability and broad applications of GEMs of metabolism (M models) challenged their boundaries and drove further development. M models can be improved by increasing the number of constraints or by expanding their scope in terms of cellular processes represented. For instance, a framework that incorporates enzyme abundances as constraints in M models substantially reduces the solution space, but requires enzyme turnover numbers (so-called k_{cat} values)[101]. Researchers have also developed models that integrate multiple layers (metabolism, transcription and signal transduction) of the bacterial organism using multi-omics data[102].

Expanding the scope of GEMs to include proteome allocation. A major effort focused on expanding M models to include a genome-scale account of translation and transcription, leading to so-called ME models (for 'metabolism and expression'). ME models are more fundamental than proteome- or enzyme-constrained models, as they explicitly incorporate a full reconstruction of the pathways that constitute transcription and translation in addition to metabolism, enabling the simulation of proteome composition. Thus, the constraints on the proteome are generated by the ME model itself as a part of computing a particular phenotypic state. The general formulation of ME models is depicted in FIG. 5. Like M models, ME models are solved using flux balances. ME models can thus be used to compute the proteome allocation between growth conditions of a strain (proximal causation) or evolutionary adaptation to a new condition (distal causation), which greatly expand the range of biological functions and behaviours over an M model.

Building ME models. The first large-scale network reconstructed to describe the transcriptional and translational machinery in *E. coli* appeared in 2009 (REF.[103]). The reconstruction was mathematically described by the expression matrix (E matrix) representing 13,694 biochemical reactions that delineate the expression of genes and protein synthesis in *E. coli*. The E matrix incorporated all the functional components (proteins, nucleotides and so on) and pathways known at the time underlying translation and transcription, including biosynthesis, modification and degradation of RNA and protein complexes. This reconstruction was also converted to a computational model to enable quantitative integration of omics data and simulation of phenotypic states; for example, the model predicted the ribosome production accurately under different conditions without any parameterization.

An ME model is an integration of the E matrix with an M model (FIG. 5). The M model describes the metabolic function and the E matrix delineates the macromolecular expression pathways. The M model and E matrix are combined through their shared metabolites and coupling constraints; that is, macromolecules are produced at a rate proportional to the rate of enzyme dilution to daughter cells (growth rate), proportional to the activity of the metabolic reaction and inversely proportional to the enzyme turnover rate (k_{cat}). By incorporating the E matrix into an M model, ME models enable the calculation of the cellular cost of enzyme synthesis, which is coupled to the reaction they catalyse. The maximum growth rate in ME models is thus solved by iteratively plugging in increasing growth rates until the maximum value that produces a solvable model is found.

Towards 'proteometrics'. The ME model's formulation essentially produces an econometric model of cellular functions. Each cell has a limited space for protein to perform its metabolic and growth functions (the size of the *E. coli* proteome is estimated to be about 2.5 million protein molecules per cell)[104]. By assigning a 'capital expense' (that is, investment in proteome synthesis — the hardware of the cell) to each metabolic function, the ME model provides a framework to determine the most protein-cost-effective way for the cell to perform its required functions. A consequence of this ME model characteristic is that the substrate uptake rates do not need to be defined a priori, as is the case for M models. Optimal substrate uptake rates are determined by the optimal protein composition. As ME models are econometric in the sense that they compute the best 'capital expenditures' (that is, proteome allocation) and 'operating expenses' (that is, best metabolic state) to achieve a particular phenotypic state, one might think of them as being 'proteometric' models.

Whereas M model solutions fall within a multidimensional solution space (that is, there are alternative solutions for any optimal objective value), ME model solutions at their maximum feasible growth rate are effectively unique. Furthermore, the ME model not only

Proximal causation
Explains traits/events (such as change in proteome allocation) in terms of immediate physiological or environmental factors.

Distal causation
Explains traits/events (such as change in proteome allocation) in terms of evolutionary forces acting on them.

Expression matrix
(E matrix). A matrix that describes all components (including DNA, mRNA, proteins and metabolites) and reactions that are involved in the transcriptional and translational machinery in the organism of interest.

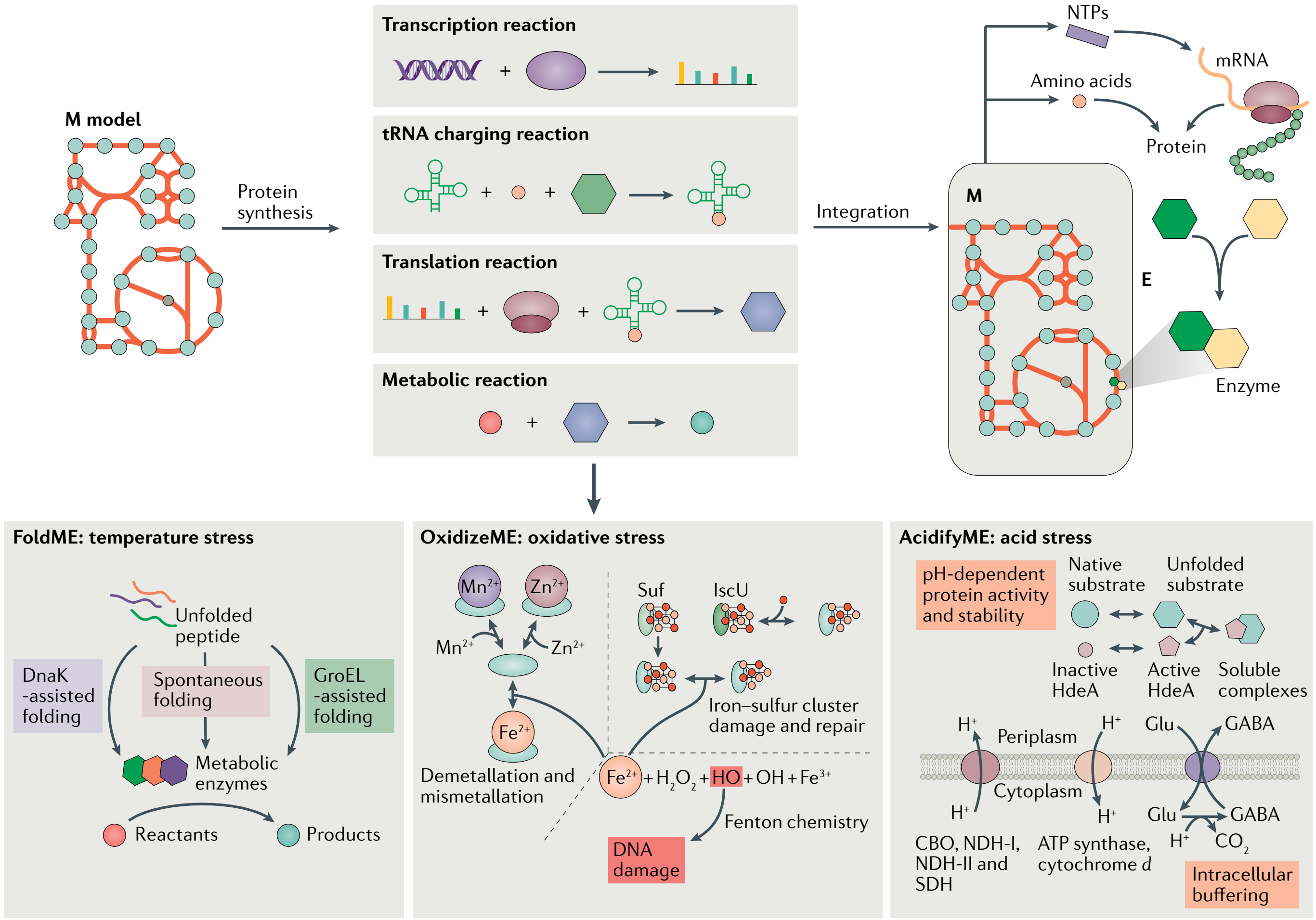


Fig. 5 | **General formulation of a ME model and its application to the study of stress response.** Metabolic and macromolecular expression (ME) models are generated through the integration of metabolic models (M models) and protein synthesis pathways, including transcription, tRNA charging and translation. Therefore, the ME model describes the biosynthesis of proteins and their roles in catalysing the metabolic reactions. Stress-specific response mechanisms are integrated with the *Escherichia coli* ME model to produce stress-specific ME models: FoldME, OxidizeME and AcidifyME. FoldME models respond to temperature stress through the incorporation of chaperone-mediated (GroEL or DnaK) or spontaneous folding pathways. OxidizeME simulates the response to oxidative stress through the inclusion of oxidation and demetallation in metalloproteins, iron–sulfur cluster cofactor damage and repair, and DNA damage. AcidifyME models the mechanisms related to acid stress, including pH-dependent protein activity and stability, membrane composition, and intracellular buffering. CBO, cytochrome *bo* terminal oxidase; GABA, γ-aminobutyric acid; Glu, glutamate; NDH-I, NADH dehydrogenase I; NDH-II, NADH dehydrogenase II; NTPs, nucleoside triphosphates; SDH, succinate dehydrogenase. FoldME panel adapted from REF.[127], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). OxidizeME panel adapted with permission from REF.[129], PNAS.

predicts a cell's maximal growth rate and corresponding metabolic fluxes but also computes the optimal proteome allocation and gene product expression level. The ME model basically represents molecular biology and biochemistry on a genome scale, and through its mathematical representation allows the computation of its fully balanced operation. However, it is worth noting that although the ME model covers both transcription and translation, it does not model the regulatory processes.

ME models are based on optimality principles with the implicit assumption that regulation will produce the computed phenotypic state. This characteristic opens up the ability to address a fundamental question: namely, do the evolved transcriptional regulatory processes reflect optimality principles that can be represented in an ME model? In other words, can evolution- and adaptation-produced outcomes be represented by the appropriate statement of an optimal function?

Experience with specific ME models and their applications. An ME model was first reconstructed for *Thermotoga maritima*, which has a genome with 1,877 annotated genes. The ME model for *T. maritima* was developed as a prototype, returned accurate predictions of cellular composition and gene expression, and showed potential for aiding in the discovery of new regulons and genome annotation[105]. Growth simulation identified a set of genes with strong differential expression when *T. maritima* grows in minimal medium with L-arabinose or cellobiose as the carbon source, suggesting the presence of transcriptional regulation. The predicted differentially expressed genes led the study authors to discover potential transcription factor-binding motifs that are similar to known motifs in other organisms, highlighting how ME models can guide discovery of new regulons.

A year later, an ME model was built for *E. coli* through the integration of the E matrix with the most

Overflow metabolism
When cells incompletely oxidize their substrate (which yields less energy) instead of using the more energetically efficient respiratory pathways to completely oxidize their substrates even when oxygen is available.

recent M model available[75] (FIGS 3,5). This ME model was able to better predict some phenotypes than M models due to its expanded scope and additional constraints. For example, unlike previous GEMs, the growth rate predicted by the *E. coli* ME model has a non-linear relationship with the substrate uptake rate, which is consistent with the long-standing empirical models of microbial growth. ME model simulation suggests that under nutrient-limited conditions, growth is constrained by substrate availability, whereas under nutrient-excess conditions, growth is limited by internal constraints on protein synthesis and catalysis. ME models can also predict the maximum batch growth rate and optimal substrate uptake rate, which closely match experimental data from laboratory-evolved strains.

Subsequent efforts focused on the improvement of several aspects of the *E. coli* ME models (FIG. 3). One study added the protein translocation pathways across the inner membrane, leading to four cellular compartments and membrane constraints that reflect the cell morphology[106]. Efforts to refine the enzyme turnover rate (k_{cat}) through machine learning methods[107] were undertaken, and a reformulation of the *E. coli* ME model computations by grouping major cellular processes and implementing explicit coupling constraints drastically reduced the size of the stoichiometric matrix and computational solving time[108].

The expanded predictive capability of ME models motivated their construction for other microorganisms of interest. The development of an ME model for *Clostridium ljungdahlii* enabled the prediction of overflow metabolism that shed light onto medium optimization strategies for bioproduction[109]. The Wood–Ljungdahl pathway is the only known CO_2-fixing pathway coupled to energy conservation in *C. ljungdahlii*, and trace metals are crucial in this pathway. The ME model was able to evaluate the impact of trace metals on metabolite secretion as the model incorporated protein modifications accounting for these metals. Specifically, simulation results suggested that removal of nickel from the medium may reduce acetate production, leading to ethanol production as the main fermentation product, providing valuable insights into bioproduction design strategies. A summary of published ME models and their characteristics can be found in TABLE 1. Additional species-specific models are under development. It is worth noting that another ME model formulation has been developed[110] to model metabolism, gene expression and thermodynamic constraints, enabling new insights into the diauxic behaviour in bacteria[111].

Thus far, ME models have been developed for only a few microorganisms besides *E. coli* due to the challenges regarding computational resources and model development. However, the reconstruction process of ME models has now been made easier with the development of the software framework COBRAme[108], a Python package that simplifies the process of reconstructing and analysing ME models. With the use of COBRAme, draft ME models can be constructed from a high-quality M model, a standard GenBank genome annotation file, curated enzyme subunit stoichiometries, mappings of enzyme complexes to a metabolic reaction and enzyme turnover rates. The ME model can also be made more sophisticated by incorporating enzyme prosthetic group information, post-transcriptional or post-translational modifications, protein translocation information, transcription unit information and other cellular processes. Once they have been reconstructed, researchers are able to edit and simulate ME models using COBRAme, which uses a software architecture mirrored after popular GEM analysis tools such as COBRApy. The streamlined computational and analysis pipelines in COBRAme have

Table 1 | **Summary of published ME models**

Model	Organism	Coverage	Key findings
T. maritima ME[105]	*Thermotoga maritima*	Metabolism, macromolecular synthesis, post-transcriptional modification and dilution to daughter cells	Accurately predicted cellular composition and gene expression; enabled new regulon discovery and genome annotation
iOL1650-ME[124]	*Escherichia coli*	1,650 genes and 1,295 protein complexes accounting for metabolism, gene expression and macromolecular synthesis	Accurate prediction of multiscale phenotypes: revealed the importance of proteomic constraints on growth, by-product secretion, metabolic flux and optimal uptake rates
iJL1678-ME[106]	*E. coli*	Incorporated 4 compartments, (cytoplasm, periplasm and inner and outer membranes), translocation pathways and membrane constraints in iOL1650-ME	Enabled prediction of enzyme abundances and their cellular location; predicted impact of perturbations such as membrane crowding and enzymatic efficiency
iJL1678b-ME[108]	*E. coli*	Compared with iJL1678-ME, reformulated coupling constraints, grouped major cellular processes and included non-equivalent changes	Significantly reduced free variables and solution time; increased accuracy in model prediction
iJL965-ME[109]	*Clostridium ljungdahlii*	965 genes and 735 protein complexes accounting for central metabolism, transcription, translation, macromolecule modifications and translocation	Produced accurate prediction of fermentation profiles, yielding deep interpretation of overflow metabolism products, gene expression and use of cofactors and metals

ME, metabolic and macromolecular expression.

enabled a substantially expanded range of computational predictions, as we discuss in the following section.

It is worth mentioning that assembling the two data types mentioned above — enzyme complex stoichiometry and enzymatic turnover rates — can present bottlenecks when one is constructing ME models. For well-studied microorganisms, assembling enzyme complex compositions can be aided with the use of expertly curated organism knowledge bases such as MetaCyc[112]. Alternatively, elucidating enzyme turnover rates on a systems level is an area of ongoing research[113] that has recently been facilitated by the use of machine learning and omics data sets[107,114]. Further work is necessary to determine the sensitivity of ME models to these parameters and the degree to which these quantities may be conserved across microorganisms.

Beyond ME models lie many additional cellular processes. Whole-cell models of the human pathogen *Mycoplasma genitalium*[115], *Saccharomyces cerevisiae*[116] and *E. coli*[117] have been developed. The whole-cell model is composed of independent modules describing particular cellular processes, such as cell replication, transcriptional regulation and DNA maintenance. Differently from ME models, whole-cell models are computed by simulating each module over a short time increment. Additionally, the enzyme abundances in whole-cell models are variables determined by the previous simulation increment, whereas enzymes in ME models are directly imposed as metabolites in their metabolic reactions, which ensures that protein limitation has a dominant role in defining the metabolic flux state.

From growth to stress responses

The scope and range of prediction continue to grow as the coverage of cellular processes in GEMs expands[118–123]. Whereas M models enable predictions of growth on different nutrients, metabolite secretion and auxotrophy, ME models have added capabilities to simulate the proteome allocation and RNA-to-protein mass ratio for a given phenotype[124], and differential gene expression levels across environmental shifts[105].

Combined with genome-scale multi-omics data, ME models have become useful tools that provide a mechanistic and systems-level understanding of *E. coli*. The integration of ME models and global proteomics data was used to characterize the unused proteome, that is, protein molecules that are not utilized or are underutilized for cellular growth (although they might be synthesized), and protein molecules that are present in excess in *E. coli*. By comparing the number of protein molecules needed for growth predicted by the ME model with quantitative proteomics data, O'Brien and colleagues identified proteins that were not used for growth. The unused proteins were shown to decrease with increasing growth rate, suggesting that there exists a fitness trade-off between growth rate and the unused proteins encoding stress- and nutrient-preparedness functions. This trade-off possibly conveys fitness benefits in changing environments while taking resources away from growth[125].

The ME model formulation can demand that translated proteins are folded, equipped with the proper prosthetic groups and assembled into protein complexes to perform their enzymatic function. Modelling the proteome with this level of detail inherently provides a robust link between metabolism and the biosynthesis of functional enzyme complexes[126]. ME models therefore enable genome-scale investigations into the cellular response to any dysfunction in protein synthesis or maintenance, such as those that can occur when cells experience stress conditions. Thus, several extensions of ME models have recently been developed to describe stress response and mitigation functions in mechanistic detail[127–129]. Taking *E. coli* as an example, reconstructions of known stress response mechanisms have been integrated with ME models to form a new generation of models: FoldME[128], OxidizeME[129] and AcidifyME[127], which simulate the response to thermal, oxidative, and low-pH stress, respectively (FIGS 3,5). Each of these environmental stresses are relevant to the lifestyle of *E. coli*, particularly when existing in a host organism.

The FoldME model extension expands the ME model to include peptide folding (chaperone-mediated or spontaneous folding) while taking into account basic biochemical properties such as protein folding rates and thermostability. By detailing these proteostatic mechanisms, FoldME is capable of describing protein folding, denaturing and catalytic activity as a function of temperature on a genome scale. Use of FoldME produced multiscale predictions for cellular adaptations under high temperature by introducing the unfolded state of the proteins and in vivo protein folding as a competition between spontaneous folding and DnaK- or GroEL-assisted folding (FIG. 5). FoldME faithfully recapitulated the temperature-dependent growth rate and changes in protein abundances[128] — as the optimal growth temperature for *E. coli* is exceeded, more proteome denatures, forcing more chaperones to be expressed, and therefore less of the total proteome is available for growth functions.

Another universal stress that may hinder cell growth is reactive oxygen species (ROS). Oxidative damage in a cell can manifest itself in multiple ways, including oxidation and demetallation of the mononuclear iron cofactors in metalloproteins, iron–sulfur cluster cofactor damage and DNA damage. The OxidizeME extension was constructed by incorporating pathways involved in these ROS-based damage and repair processes (FIG. 5). Furthermore, structural biology was applied to determine which proteins, based on the position of metallic cofactors in the 3D structure of the enzyme, were most susceptible to ROS damage. As ME models explicitly require the presence of the proper unimpaired cofactors in order for an enzyme to possess any catalytic function, the model could assess the systems-level effects of oxidative damage and repair in *E. coli*. OxidizeME correctly predicted the phenotypes under oxidative stress, such as aromatic amino acid auxotrophy, carbon source-dependent ROS sensitivity and stress-specific differential gene expression, and traced the possible mechanisms involved in iron–sulfur cluster biosynthesis[129].

Such effort was also extended to pH stress to elucidate the changes in cellular responses under acidic conditions. AcidifyME simulates pH-dependent membrane lipid fatty acid composition, periplasmic protein stability and periplasmic chaperone protection, and membrane protein activity at low pH (FIG. 5). It recapitulated differential gene expression under acid conditions, enabled a systematic and mechanistic understanding of acid stress response and most importantly suggested potential intervention strategies[127]. For example, model simulation suggests that knocking out *hdeB*, the gene encoding the only known periplasmic chaperone in *E. coli*, would result in no growth under acidic conditions. If such predictions can be verified by experimental studies, HdeB could become a promising antimicrobial target to inhibit *E. coli* growth in acidic environments such as the human digestive tract.

Conclusions

The first annotated genome sequences appeared in the mid to late 1990s. With metabolism being a well-characterized cellular process, a comprehensive list of metabolic genes was identified in these newly sequenced genomes. The recognition that the biochemical functions of enzymes could be defined led to the formulation of a process for network reconstruction at the genome scale. In other words, one could, in principle, reconstruct the entire metabolic network from an annotated genome sequence. In practice, reconstruction technology has advanced over the past 20 years to include protocols to deal with issues arising from incomplete genome annotation and the development of quality control standards.

Reconstructions are knowledge bases that have many uses. One use detailed here is the conversion of knowledge into computational models that represent the functions of an ‘in silico’ cell whose properties can be computationally simulated. These models open up the comparison between characterization of what is known about an organism (that is, the GEM) and how the organism actually functions. As we do not have complete knowledge of any organism, the difference between the two (observed and simulated functions) has proved to be a guide to the discovery of missing parts and an understanding of integrated cellular functions.

The computation of biological functions needs to represent proximal and distal causation. GEMs formulated through a constraint-based formalism can represent both, and thus simulate dual causation[130]. Proximal causation can be comprehensively detailed through the inclusion of increasingly accurate biophysical representations of cellular processes. This approach has led to the formulation of whole-cell models of *M. genitalium*[115], *S. cerevisiae*[116] and *E. coli*[117] to describe in increasing biophysical detail their molecular components and interactions[115]. These models become increasingly specific to a particular strain functioning in particular environments.

Distal causation can be pursued through adaptive laboratory evolution and through pan-genomics. Here, the differences between strains and species are considered, and the question of interest is how natural selection leads to adaptation and longer-term evolution. Reconstruction and GEMs are used as tools to understand how selection pressures have shaped the gene portfolios with the corresponding phenotypic potential. The most comprehensive description of the formulation, underlying philosophy and use of constraint-based models is found in a recent textbook[131].

As reviewed here, GEMs have developed over 20 years, starting with metabolism then expanding in scope to include transcription and translation and stress functions. They will continue to grow in their scope and accuracy in the representation of known cellular functions. Comprehensive representations of two-component systems and the structural proteome[76,79] are now possible, as are cell division mechanisms, whose inclusion will refine the models from representations of populations to individual cells. This process will continually improve our understanding of how microbial cells function and evolve and will likely one day assist with the design of synthetic genomes.

Published online: 21 September 2020

1. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
A detailed protocol to reconstruct a GEM from a genome sequence.
2. Norsigian, C. J., Fang, X., Seif, Y., Monk, J. M. & Palsson, B. O. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. *Nat. Protoc.* **15**, 1–14 (2020).
This article provides a semi-automated workflow to generate strain-specific GEMs from a curated reference model.
3. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).
4. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305 (2012).
5. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
6. Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–349 (2010).
7. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245 (2010).
This is primer that thoroughly explains FBA and its applications.
8. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst. Biol.* **7**, 74 (2013).
9. Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
10. Feist, A. M. & Palsson, B. Ø. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* **26**, 659–667 (2008).
11. McCloskey, D., Palsson, B. Ø. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9**, 661 (2013).
12. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15**, 107–120 (2014).
13. O’Brien, E. J., Monk, J. M. & Palsson, B. O. Using genome-scale models to predict biological capabilities. *Cell* **161**, 971–987 (2015).
14. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol.* **20**, 121 (2019).
15. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).
16. Mazumdar, V., Snitkin, E. S., Amar, S. & Segrè, D. Metabolic network model of a human oral pathogen. *J. Bacteriol.* **191**, 74–90 (2009).
17. Henry, C. S., Jankowski, M. D., Broadbelt, L. J. & Hatzimanikatis, V. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys. J.* **90**, 1453–1461 (2006).
18. Navid, A. & Almaas, E. Genome-level transcription data of Yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC Syst. Biol.* **6**, 150 (2012).
19. Motter, A. E., Gulbahce, N., Almaas, E. & Barabási, A.-L. Predicting synthetic rescues in metabolic networks. *Mol. Syst. Biol.* **4**, 168 (2008).
20. Montagud, A., Navarro, E., Fernández de Córdoba, P., Urchueguía, J. F. & Patil, K. R. Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst. Biol.* **4**, 156 (2010).
21. Mahadevan, R., Edwards, J. S. & Doyle, F. J. III. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys. J.* **83**, 1331–1340 (2002).
22. Hastings, J. et al. Multi-omics and genome-scale modeling reveal a metabolic shift during C. elegans aging. *Front. Mol. Biosci.* **6**, 2 (2019).
23. Guan, N. et al. Comparative genomics and transcriptomics analysis-guided metabolic engineering of Propionibacterium acidipropionici for improved propionic acid production. *Biotechnol. Bioeng.* **115**, 483–494 (2018).
24. 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).
25. Hefzi, H. et al. A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism. *Cell Syst.* **3**, 434–443.e8 (2016).

26. Cautha, S. C. et al. Model-driven design of a Saccharomyces cerevisiae platform strain with improved tyrosine production capabilities. *IFAC Proc. Vol.* **46**, 221–226 (2013).

27. McAnulty, M. J., Yen, J. Y., Freedman, B. G. & Senger, R. S. Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC Syst. Biol.* **6**, 42 (2012).

28. Cardoso, J. G. R., Andersen, M. R., Herrgård, M. J. & Sonnenschein, N. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. *Front. Bioeng. Biotechnol.* **3**, 13 (2015).

29. Folger, O. et al. Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, 501 (2011).

30. Guzmán, G. I. et al. Model-driven discovery of underground metabolic functions in Escherichia coli. *Proc. Natl Acad. Sci. USA* **112**, 929–934 (2015).

31. Kumar, M., Ji, B., Zengler, K. & Nielsen, J. Modelling approaches for studying the microbiome. *Nat. Microbiol.* **4**, 1253–1267 (2019).

32. Zuñiga, C. et al. Environmental stimuli drive a transition from cooperation to competition in synthetic phototrophic communities. *Nat. Microbiol.* **4**, 2184–2191 (2019).

33. Zomorrodi, A. R. & Maranas, C. D. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* **8**, e1002363 (2012).

34. Stolyar, S. et al. Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* **3**, 92 (2007).

35. 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).

36. Harcombe, W. R. et al. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* **7**, 1104–1115 (2014).

37. Zomorrodi, A. R. & Segrè, D. Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nat. Commun.* **8**, 1563 (2017).

38. Zelezniak, A. et al. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl Acad. Sci. USA* **112**, 6449–6454 (2015).

39. Chowdhury, S. & Fong, S. S. Computational modeling of the human microbiome. *Microorganisms* **8**, 197 (2020).

40. 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).

41. Mardinoglu, A. et al. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014).

42. Mardinoglu, A. & Nielsen, J. Systems medicine and metabolic modelling. *J. Intern. Med.* **271**, 142–154 (2012).

43. Edwards, J. S. & Palsson, B. O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416 (1999).

44. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. *Nat. Biotechnol.* **32**, 447–452 (2014).

45. Mendoza, S. N., Olivier, B. G., Molenaar, D. & Teusink, B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* **20**, 158 (2019).

46. Satish Kumar, V., Ferry, J. G. & Maranas, C. D. Metabolic reconstruction of the archaeon methanogen Methanosarcina acetivorans. *BMC Syst. Biol.* **5**, 28 (2011).

47. Benedict, M. N., Gonnerman, M. C., Metcalf, W. W. & Price, N. D. Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon Methanosarcina acetivorans C2A. *J. Bacteriol.* **194**, 855–865 (2012).

48. Peterson, J. R. et al. Genome-wide gene expression and RNA half-life measurements allow predictions of regulation and metabolic behavior in Methanosarcina acetivorans. *BMC Genomics* **17**, 924 (2016).

49. Sheikh, K., Förster, J. & Nielsen, L. K. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol. Prog.* **21**, 112–121 (2005).

50. Lu, H. et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).

51. de Oliveira Dal'Molin, C. G. & Nielsen, L. K. Plant genome-scale metabolic reconstruction and modelling. *Curr. Opin. Biotechnol.* **24**, 271–277 (2013).

52. Mintz-Oron, S. et al. Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Natl Acad. Sci. USA* **109**, 339–344 (2012).

53. Cheung, C. Y. M. et al. A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant. J.* **75**, 1050–1061 (2013).

54. Thiele, I. et al. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425 (2013).

55. Swainston, N. et al. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12**, 109 (2016).

56. Brunk, E. et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.* **36**, 272–281 (2018).

57. Kumar, A., Suthers, P. F. & Maranas, C. D. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics* **13**, 6 (2012).

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

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

60. Agren, R. et al. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput. Biol.* **9**, e1002980 (2013).

61. Wang, H. et al. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLoS Comput. Biol.* **14**, e1006541 (2018).

62. Arkin, A. P. et al. KBase: the United States department of energy systems biology knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).

63. Norsigian, C. J. et al. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **48**, D402–D406 (2019).

64. Glont, M. et al. BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res.* **46**, D1248–D1253 (2018).

65. Moretti, S. et al. MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).

66. Pabinger, S. et al. MEMOSys 2.0: an update of the bioinformatics database for genome-scale models and genomic data. *Database* **2014**, bau004 (2014).

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

68. Lieven, C. et al. MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* **38**, 272–276 (2020).

69. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli II. Optimal growth patterns. *J. Theor. Biol.* **165**, 503–522 (1993).

70. Pramanik, J. & Keasling, J. D. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* **56**, 398–421 (1997).

71. 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).

72. Edwards, J. S. & Palsson, B. O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).

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

74. 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).

75. Orth, J. D. et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism — 2011. *Mol. Syst. Biol.* **7**, 535 (2011).

76. Monk, J. M. et al. iML1515, a knowledgebase that computes Escherichia coli traits. *Nat. Biotechnol.* **35**, 904–908 (2017).
This article provides the most up-to-date genome-scale metabolic model of *E. coli*.

77. Archer, C. T. et al. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics* **12**, 9 (2011).

78. Blattner, F. R. et al. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).

79. Brunk, E. et al. Systems biology of the structural proteome. *BMC Syst. Biol.* **10**, 26 (2016).

80. Mih, N. et al. ssbio: a Python framework for structural systems biology. *Bioinformatics* **34**, 2155–2157 (2018).

81. Ghatak, S., King, Z. A., Sastry, A. & Palsson, B. O. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res.* **47**, 2446–2454 (2019).

82. Mih, N. & Palsson, B. O. Expanding the uses of genome-scale models with protein structures. *Mol. Syst. Biol.* **15**, e8601 (2019).

83. Magnúsdóttir, S. & Thiele, I. Modeling metabolism of the human gut microbiome. *Curr. Opin. Biotechnol.* **51**, 90–96 (2018).

84. Opdam, S. et al. A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell Syst.* **4**, 318–329.e6 (2017).

85. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7**, 72–85 (2015).

86. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. *Microb. Ecol.* **60**, 708–720 (2010).

87. Seif, Y. et al. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. *Nat. Commun.* **9**, 3771 (2018).

88. John, J., George, S., Nori, S. R. C. & Nelson-Sathi, S. Phylogenomic analysis reveals the evolutionary route of resistant genes in Staphylococcus aureus. *Genome Biol. Evol.* **11**, 2917–2926 (2019).

89. Holt, K. E. et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl Acad. Sci. USA* **112**, E3574–E3581 (2015).

90. Yang, Z.-K., Luo, H., Zhang, Y., Wang, B. & Gao, F. Pan-genomic analysis provides novel insights into the association of E.coli with human host and its minimal genome. *Bioinformatics* **35**, 1987–1991 (2019).

91. Monk, J. M. et al. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl Acad. Sci. USA* **110**, 20338–20343 (2013).
This is the first study to focus on a strain-level study of *E. coli* using GEMs.

92. Fang, X. et al. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intefstinal mucosa. *BMC Syst. Biol.* **12**, 66 (2018).

93. Fang, X. et al. Metagenomics-based, strain-level analysis of Escherichia coli from a time-series of microbiome samples from a Crohn's disease patient. *Front. Microbiol.* **9**, 2559 (2018).

94. Bosi, E. et al. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl Acad. Sci. USA* **113**, E3801–E3809 (2016).

95. Seiler, N. & Atanassov, C. L. The natural polyamines and the immune system. *Prog. Drug Res.* **43**, 87–141 (1994).

96. Norsigian, C. J. et al. Comparative genome-scale metabolic modeling of metallo-beta-lactamase–producing multidrug-resistant *Klebsiella pneumoniae* clinical isolates. *Front. Cell. Infect. Microbiol.* **9**, 161 (2019).

97. Norsigian, C. J., Kavvas, E., Seif, Y., Palsson, B. O. & Monk, J. M. iCN718, an updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE. *Front. Genet.* **9**, 121 (2018).

98. Fouts, D. E. et al. What makes a bacterial species pathogenic?: Comparative genomic analysis of the genus *Leptospira*. *PLoS Negl. Trop. Dis.* **10**, e0004403 (2016).

99. Nogales, J. et al. High-quality genome-scale metabolic modelling of Pseudomonas putida highlights its broad metabolic capabilities. *Environ. Microbiol.* **22**, 255–269 (2020).

100. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* **11**, 2580 (2020).
101. Sánchez, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
102. Carrera, J. et al. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Mol. Syst. Biol.* **10**, 735 (2014).
103. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. Ø. 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).
104. Milo, R. & Phillips, R. *Cell Biology by the Numbers* (Garland Science, 2015).
105. Lerman, J. A. et al. In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3**, 929 (2012).
106. Liu, J. K., O'Brien, E. J. & Lerman, J. A. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Syst. Biol.* **8**, 110 (2014).
107. Heckmann, D. et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat. Commun.* **9**, 5252 (2018).
108. Lloyd, C. J. et al. COBRAme: a computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput. Biol.* **14**, e1006302 (2018).
This article provides a framework that streamlines ME model reconstruction and analysis.
109. Liu, J. K. et al. Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen. *PLoS Comput. Biol.* **15**, e1006848 (2019).
110. Salvy, P. & Hatzimanikatis, V. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. *Nat. Commun.* **11**, 30 (2020).
This article describes an implementation of the ME models called 'ETFL', for 'expression and thermodynamics flux models'.
111. Salvy, P. & Hatzimanikatis, V. Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.15.204420 (2020).
112. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
113. Nilsson, A., Nielsen, J. & Palsson, B. O. Metabolic models of protein allocation call for the kinetome. *Cell Syst.* **5**, 538–541 (2017).
114. Heckmann, D. et al. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc. Natl Acad. Sci. USA* **117**, 23182–23190 (2020).
115. Karr, J. R. et al. A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
116. Ye, C. et al. Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C. *Biotechnol. Bioeng.* **117**, 1562–1574 (2020).
117. Macklin, D. N. et al. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science* **369**, eaav3751 (2020).
118. Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D. & Koffas, M. A. G. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.* **13**, 578–587 (2011).
119. Akesson, M., Förster, J. & Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6**, 285–293 (2004).
120. Yim, H. et al. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat. Chem. Biol.* **7**, 445–452 (2011).
121. Wang, Q., Chen, X., Yang, Y. & Zhao, X. Genome-scale in silico aided metabolic analysis and flux comparisons of Escherichia coli to improve succinate production. *Appl. Microbiol. Biotechnol.* **73**, 887–894 (2006).
122. Imielinski, M., Belta, C., Rubin, H. & Halász, A. Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. *Biophys. J.* **90**, 2659–2672 (2006).
123. Yang, L. et al. Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proc. Natl Acad. Sci. USA* **112**, 10810–10815 (2015).
124. O'brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).
This study outlines the development of the first *E. coli* ME model.
125. O'Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and classification of E. coli proteome utilization and unused protein costs across environments. *PLoS Comput. Biol.* **12**, e1004998 (2016).
126. Lloyd, C. J., Monk, J., Yang, L., Ebrahim, A. & Palsson, B. O. Computation of condition-dependent proteome allocation reveals variability in the macro and micro nutrient requirements for growth. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.23.003236 (2020).
127. Du, B., Yang, L., Lloyd, C. J., Fang, X. & Palsson, B. O. Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in Escherichia coli. *PLoS Comput. Biol.* **15**, e1007525 (2019).
As one of the three stress response extensions of ME model, this study focuses on modelling response to acid stress.
128. Chen, K. et al. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proc. Natl Acad. Sci. USA* **114**, 11548–11553 (2017).
As one of the three stress response extensions of the ME model, FoldME focuses on simulation of response to temperature stress.
129. Yang, L. et al. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. *Proc. Natl Acad. Sci. USA* **116**, 14368–14373 (2019).
As one of the three stress response extensions of the ME model, OxidizeME incorporates mechanisms involved in response to oxidative stress.
130. Mayr, E. *This is Biology: The Science of the Living World* (Harvard Univ. Press, 1998).
131. Palsson, B. *Systems Biology* (Cambridge Univ. Press, 2015).
132. Davis, J. J. et al. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
133. Bailey, J. E. Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnol. Prog.* **14**, 8–20 (1998).
134. 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 (2004).
135. Mahadevan, R., Palsson, B. Ø. & Lovley, D. R. In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. *Nat. Rev. Microbiol.* **9**, 39–50 (2011).
136. Kim, H. U. et al. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Mol. Syst. Biol.* **7**, 460 (2011).
137. Luo, H. et al. Coupling S-adenosylmethionine-dependent methylation to growth: design and uses. *PLoS Biol.* **17**, e2007050 (2019).
138. Reed, J. L. et al. Systems approach to refining genome annotation. *Proc. Natl Acad. Sci. USA* **103**, 17480–17484 (2006).
139. Kumar, V. S. & Maranas, C. D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput. Biol.* **5**, e1000308 (2009).
140. Orth, J. D. & Palsson, B. Ø. Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* **107**, 403–412 (2010).
141. Medlock, G. L. & Papin, J. A. Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. *Cell Syst.* **10**, 109–119.e3 (2020).

Acknowledgements

This work was supported by the US National Institutes of General Medical Sciences of the National Institutes of Health (grant R01GM057089) and the Novo Nordisk Foundation (grant NNF10CC1016517). Some figures were adapted from previously published figures, and the authors acknowledge the original authors of these figures and thank them for their contributions: J. M. Monk (Fig. 3), L. Yang (Fig. 5), K. Chen (Fig. 5), B. Du (Fig. 5) and A. Bordbar (Fig. 1).

Author contributions

X.F. researched data for the article. X.F. and B.O.P. substantially contributed to discussion of the content. All authors wrote the article, and reviewed or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Microbiology thanks María Suárez Diez, Sara Moreno Paz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.