PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

A workflow for annotating the knowledge gaps in metabolic reconstructions using known and hypothetical reactions

Evangelia Vayena[a,1], Anush Chiappino-Pepe[a,1,2], Homa MohammadiPeyhani[a,3], Yannick Francioli[a,4], Noushin Hadadi[a,5], Meriç Ataman[a,6], Jasmin Hafner[a,7], Stavros Pavlou[b,c], and Vassily Hatzimanikatis[a,8]

Edited by Jens Nielsen, BioInnovation Institute, Copenhagen, Denmark; received June 29, 2022; accepted September 30, 2022

Advances in medicine and biotechnology rely on a deep understanding of biological processes. Despite the increasingly available types and amounts of omics data, significant knowledge gaps remain, with current approaches to identify and curate missing annotations being limited to a set of already known reactions. Here, we introduce Network Integrated Computational Explorer for Gap Annotation of Metabolism (NICEgame), a workflow to identify and curate nonannotated metabolic functions in genomes using the ATLAS of Biochemistry and genome-scale metabolic models (GEMs). To resolve gaps in GEMs, NICEgame provides alternative sets of known and hypothetical reactions, assesses their thermodynamic feasibility, and suggests candidate genes to catalyze these reactions. We identified metabolic gaps and applied NICEgame in the latest GEM of *Escherichia coli*, iML1515, and enhanced the *E. coli* genome annotation by resolving 47% of these gaps. NICEgame, applicable to any GEM and functioning from open-source software, should thus enhance all GEM-based predictions and subsequent biotechnological and biomedical applications.

hypothetical biochemistry | gap-filling | missing annotation | metabolic model | genome annotation

The design of robust and effective medical therapies, drug targeting strategies, and bioengineering relies on a systems level understanding of biology. To this end, metabolic networks and annotated genomes are often used to gain a holistic picture of the cell functions. However, not all metabolic capabilities of cells are known, i.e., all known genomes are missing functional annotations for a relatively high portion of the open reading frames. For example, one of the best characterized organisms, *Escherichia coli*, lacks annotation for ∼1,600 genes, which represents 35% of its total number of genes (1). A limited knowledge of cell function is especially troublesome in infectious pathogens and organisms that could be used as a chassis in the industry to produce valuable compounds. Systematically identifying missing metabolic capabilities of the cell and accelerating the functional annotation of genomes can expedite and facilitate a wide range of medical and biotechnology applications.

The systematic analysis of metabolic functions and identification of knowledge gaps relies on computational models of metabolism. In fact, all known metabolic functions of different organisms are organized into databases termed genome-scale models (GEMs). These GEMs rely on the functional annotation of genes for their reconstruction, with better quality gene annotation leading to better predictions of cellular physiology. GEMs have been widely used to study the metabolism of model organisms, such as *E. coli* (2) and yeast (3), and pathogens such as *Salmonella* Typhimurium (4) and *Plasmodium falciparum* (5), and to identify host–pathogen interactions (6), drug targets (7), and metabolic engineering strategies (8), among others (9). Hence, the fact that all GEMs are currently missing knowledge and annotations can lead to false predictions that can affect both research as well as biomedical applications. Approaches to performing functional annotation of genomes involve both physical experiments (10) (e.g., in vitro assays) and bioinformatics (11) (e.g., sequence similarity). However, experiments require specific hypotheses and are time- and resource-consuming. Moreover, sequence similarity (12) and other computational approaches are so far limited to the space of known annotated proteins and biochemistry.

Exploring the space of unknown biochemistry is thus necessary to accelerate our understanding of cell function and include novel chemistry in our models of cells. The strategies to explore such unknown biochemical space are primarily based on machine learning (ML) or mechanistic approaches (13, 14). Recently, an ATLAS of Biochemistry was constructed based on a mechanistic understanding of enzyme function (15, 16) as a database of novel biochemistry, meaning not yet experimentally observed reactions, and the optimization-based exploration of metabolic models to identify missing biochemistry. The ATLAS of Biochemistry includes over 150,000 putative reactions

Significance

Bioengineering applications such as drug-targeting strategies and microbial product manufacturing rely heavily on a detailed understanding of cellular functions. However, the functionality of a considerable portion of each genome remains undefined. Experimental approaches to functionally annotate genomes require time and resources, and classic bioinformatics approaches are limited to the space of known annotated proteins and biochemistry. We introduce Network Integrated Computational Explorer for Gap Annotation of Metabolism (NICEgame), a computational workflow for characterizing and curating metabolic gaps at the reaction and enzyme level, using known and hypothetical reactions and computational enzyme annotation methods. The NICEgame workflow will help advance basic biology research, biotechnology, and biomedical engineering through postulating new hypotheses and accelerating the functional annotation of genomes.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2211197119/-/DCSupplemental.

Published November 7, 2022.

between known metabolites. Hence, it represents the upper limit of the possible biochemical space and allows an efficient exploration of the uncharacterized metabolic functions in cells. Furthermore, the tool BridgIT was recently developed as a method to map orphan biochemistry to enzymes (17), providing a tool for identifying uncharacterized genes. Together, these are separate tools for exploring the unknown biochemistry of GEMs at the reaction and the enzyme level, respectively, but are not currently integrated.

Therefore, we hypothesized that we use GEMs and leverage the potential of the ATLAS of Biochemistry (15, 16) coupled with BridgIT (17) to identify metabolic gaps and identify possible reactions with associated catalyzing enzymes and genes. This powerful combination of tools and methods came together to form our workflow, Network Integrated Computational Explorer for Gap Annotation of Metabolism (NICEgame). We applied NICEgame to suggest novel biochemistry in *E. coli* strain MG1655 and further enhance its genome annotation. From the most recently published *E. coli* GEM (2), NICEgame identified metabolic gaps that are responsible for 148 false gene essentiality predictions linked to 152 reactions in glucose minimal media. This refers to genes that the GEM considers essential for growth, but experimental data shows otherwise, meaning that there should be available biochemistry in the cell to perform these reactions in the case of gene knockout. We proposed 77 biochemical reactions linked to 35 candidate genes to fill 47% of these gaps. We integrated this information into a thermodynamically curated GEM of *E. coli,* iEcoMG1655, which has an increased gene essentiality prediction accuracy of 23.6% with respect to its predecessor iML1515 (2). Importantly, the NICEgame workflow is applicable to any organism or cell with a GEM and is available as a GitHub repository (https://github.com/EPFL-LCSB/NICEgame) with the combined use of available online resources, the ATLAS of Biochemistry (15, 16) and BridgIT (17). Overall, NICEgame is a workflow for the rapid and systematic identification of metabolic gaps, missing biochemistry, and candidate catalyzing genes. Hence it will accelerate the complete identification of metabolic functions and annotation of genomes and enable the design of robust bioengineering and drug targeting strategies.

Results

A Workflow to Identify and Curate Gaps in Cellular Metabolism. NICEgame involves seven main steps (Fig. 1), detailed in the *Materials and Methods* section. The first involves the harmonization of metabolite annotations with the ATLAS of Biochemistry, which is necessary for assuring the proper connectivity of metabolites between a GEM and the reaction database. The second step comprises a preprocessing of the GEM (e.g., by defining the media) and the identification of the metabolic gaps (e.g., by comparing in silico and in vitro gene knockout experiments. In the third step, NICEgame merges the GEM and ATLAS of Biochemistry, which is hereafter called ATLAS-merged GEM. The fourth step involves a comparative essentiality analysis with the isolated and ATLAS-merged GEM. At this point, we identify the reactions or genes, among the metabolic gaps, that are essential for in silico growth but are dispensable in the ATLAS-merged GEM. In other words, here, we look for reactions that were essential for growth in the original GEM, but the ATLAS-merged GEM has been able to overcome through alternative, currently unexplored pathways. We define such reactions or genes as rescued. The rescued reactions and genes will be the targets for gap-filling. Alternatively, if the wild-type model fails to simulate an observed phenotype, such as growth under given conditions, the comparative essentiality analysis can be omitted. In this case, the gap-filling algorithm seeks to reconcile the predictions of the wild-type model with the observed phenotype. In the fifth step, NICEgame systematically identifies alternative biochemistry to the rescued reactions or genes. In the sixth step, we evaluate and rank all alternative biochemistry. In the seventh and final step, NICEgame uses the BridgIT tool to identify a potential gene for catalyzing the top-ranked suggested biochemistry.

In all, NICEgame produces sets of alternative gap-filling reactions, which then must be evaluated based on their impact on the metabolic network and the performance of the model (*SI Appendix*, Fig. S1). The sets of reactions that are added to the network to reconcile a gap, termed solution sets, that result in a higher biomass yield or do not affect the yield are preferred to solutions that reduce the flexibility of the model; solutions that expand the metabolome or the enzymatic capabilities of the original model are ranked lower. The alternatives are also judged based on the number of reactions that are used to complement each rescued reaction. This mimics what happens in organisms where larger pathways are usually unfavored since they require more protein production, which is a highly energetically demanding process (18). Last, alternatives are ranked higher if they increase the ability of the model to correctly reproduce knockout phenotypes and do not add redundancy. Overall, these criteria are converted into scores (see *Materials and Methods*), with a positive value for any of the scores penalizing that alternative solution set in our ranking system.

Identification of Metabolic Gaps in *E. coli* with NICEgame. We applied NICEgame to the latest GEM of *E. coli*, iML1515 (2), which contains all available metabolic information on this bacterium to date. Our in silico essentiality analysis simulated a glucose minimal medium and identified 148 false-negative genes (Fig. 2*A*) corresponding to 152 false-negative essential reactions (Dataset S1), as compared with available experimental data (19). In other words, NICEgame identified 152 reactions that the GEM predicts are required for cell growth, but experimental data shows otherwise; these represent the gap that must be filled for a properly functioning GEM. Following the NICEgame workflow (Fig. 1), we next merged iML1515 with the ATLAS of Biochemistry. We performed two sets of analysis aiming to reduce (i) the number of metabolites added into the model and (ii) the uncertainty in the type of metabolites added into the model. For this purpose, we used two subsets of the ATLAS of Biochemistry to gap-fill the metabolic network of iML1515. The first subset, the *E. coli metabolites subset*, expands the reaction space of the model by adding reactions involving only metabolites from the iML1515 reconstruction. We thus examined whether the gaps in the model can be reconciled by expanding only the reaction space without increasing the metabolite space. The second subset, the *E. coli and yeast metabolites subset*, expands the reaction and metabolite space of the model by adding reactions involving only metabolites from *E. coli* (iML1515) and yeast [Yeast8 (3) metabolic reconstruction]. In this case, more information was extracted from ATLAS in a controlled way, expanding both the reaction and metabolite space of the original metabolic network. Since *E. coli* is often cultivated in yeast extract, the metabolic network of yeast was chosen, as it is likely that parts of the missing metabolome exist in the yeast metabolic reconstruction.

With these two approaches, we identified thermodynamically feasible biochemical reactions to resolve 93 out of the 152

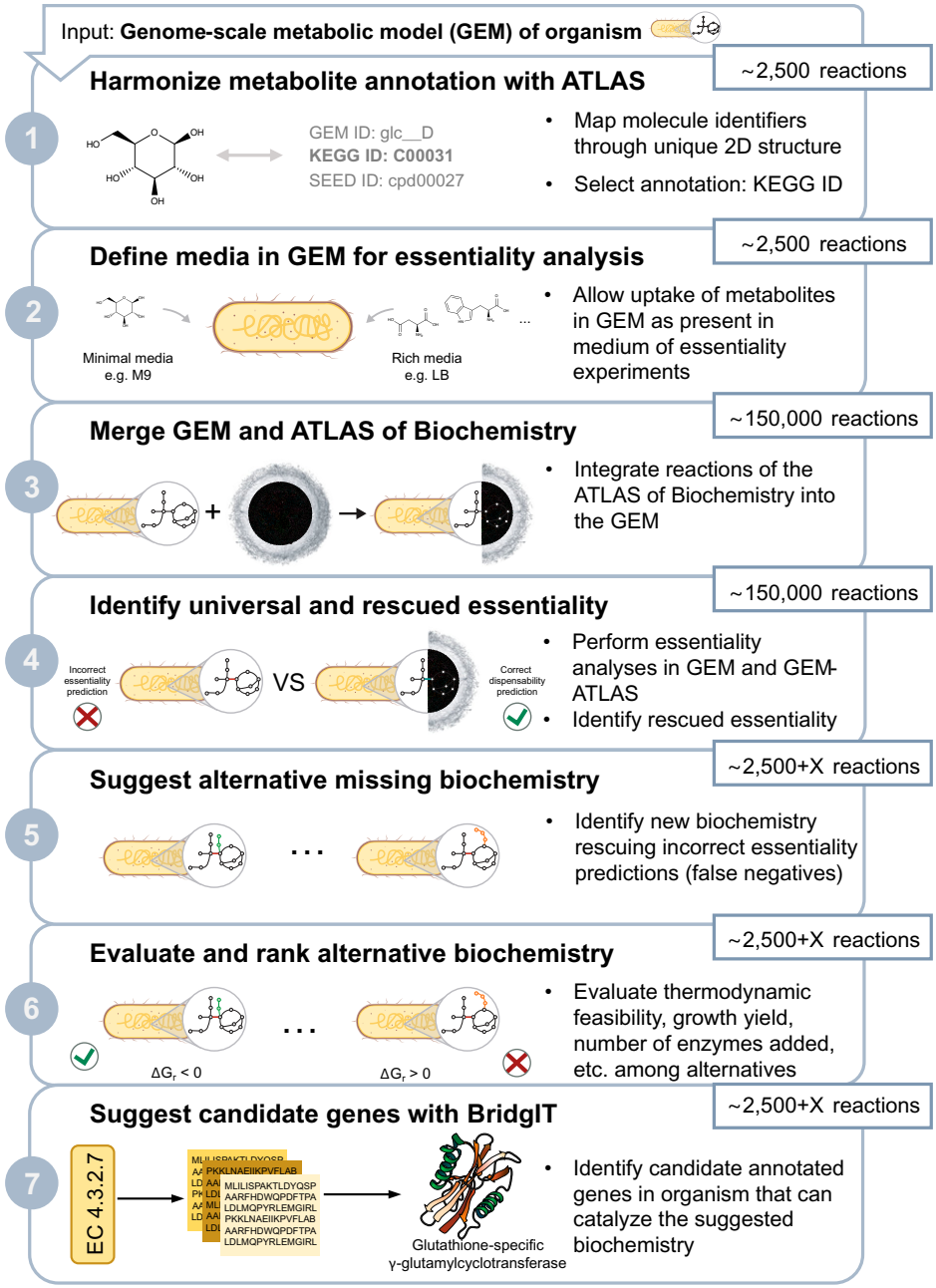


Fig. 1. Pipeline to construct and use the NICEgame workflow to annotate missing metabolic functions. The NICEgame workflow uses a GEM model as input. (*1*) The annotations of the GEM metabolites are harmonized to map them to compounds in the ATLAS of Biochemistry. (*2*) The conditions for subsequent essentiality analyses are defined, such as media composition. (*3*) The original GEM is merged with ATLAS and (*4*) an essentiality analysis is performed in the original and the expanded network to identify which gaps can be rescued. (*5*) Alternative reactions sets are generated to fill in the gaps and (6) are evaluated. (*7*) Finally, BridgIT identifies catalyzing genes for the suggested reactions.

false-negative reactions (Fig. 2*B*). Different hypotheses can be made for the remaining false-negative reactions: (i) they can be rescued if more information from ATLAS is used or (ii) chemical compounds are required as intermediates. An example of unresolved metabolic gap is the false-negative gene pabA (b3360), which catalyzes the synthesis of 4-amino-4-deoxychorismate, a precursor of folate (Fig. 2*C*). The reaction is rescued if the entire ATLAS is used as a reaction pool for the gap-filling.

Biotransformations among *E. coli* Metabolites Reconcile Model Predictions with Experimental Evidence. With the *E. coli metabolites subset* of the ATLAS of Biochemistry, we identified thermodynamically feasible gap-filling biotransformations for 86 out of 152 false-negative reactions (Dataset S2). We present here several examples of rescued reactions and alternative solutions that are ranked high or low according to our ranking system. One false-negative reaction that was rescued is AMAOTr, which describes the production of 7,8-diaminononanoate from 8-amino-7-oxononanoate in the biosynthesis pathway of biotin (Fig. 3*A*). This is catalyzed by the enzyme adenosylmethionine-8-amino-7-oxononanoate transaminase, with Enzyme Commission (20) (EC) number 2.6.1.62, which is encoded by the gene bioA. As alternatives to the bioA-linked reaction for the biosynthesis of biotin, NICEgame identified 116 thermodynamically feasible reaction sets of size one and size two. In one of the alternative solutions sets (Fig. 3 *A*, alternative solution 1), a single novel reaction can fill the gap. This reaction follows the same enzymatic mechanism (EC 2.6.1.-) as the original, although in

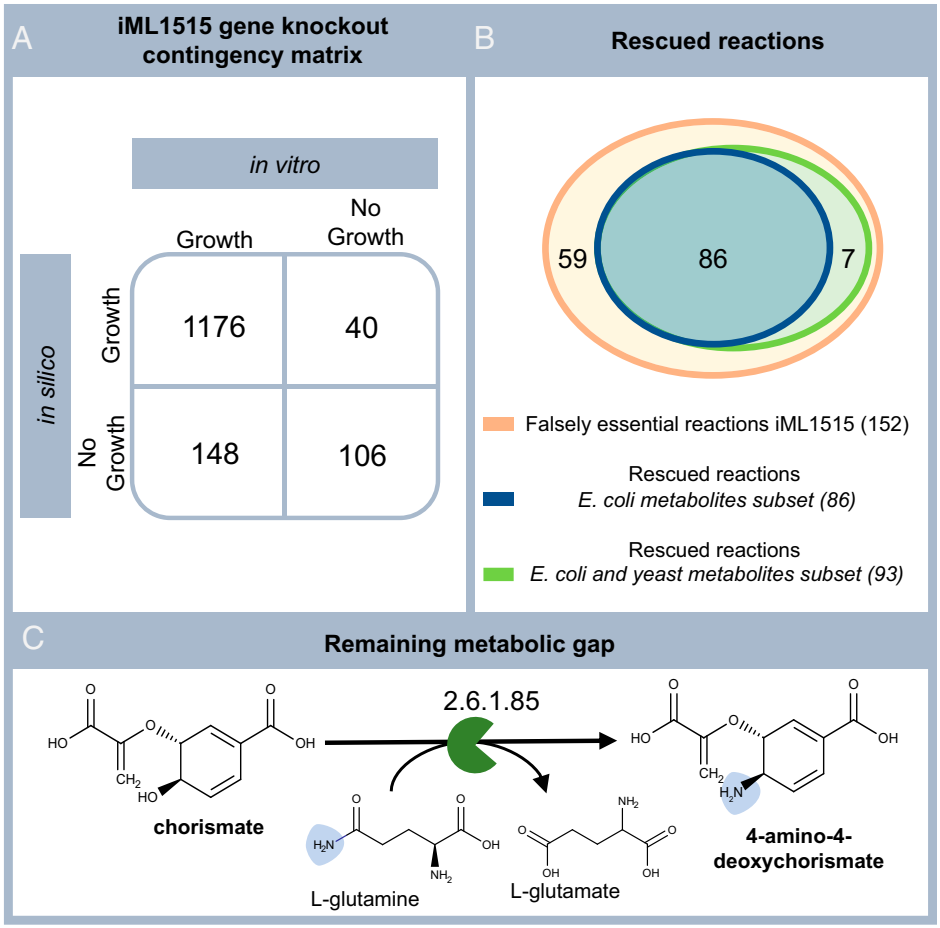


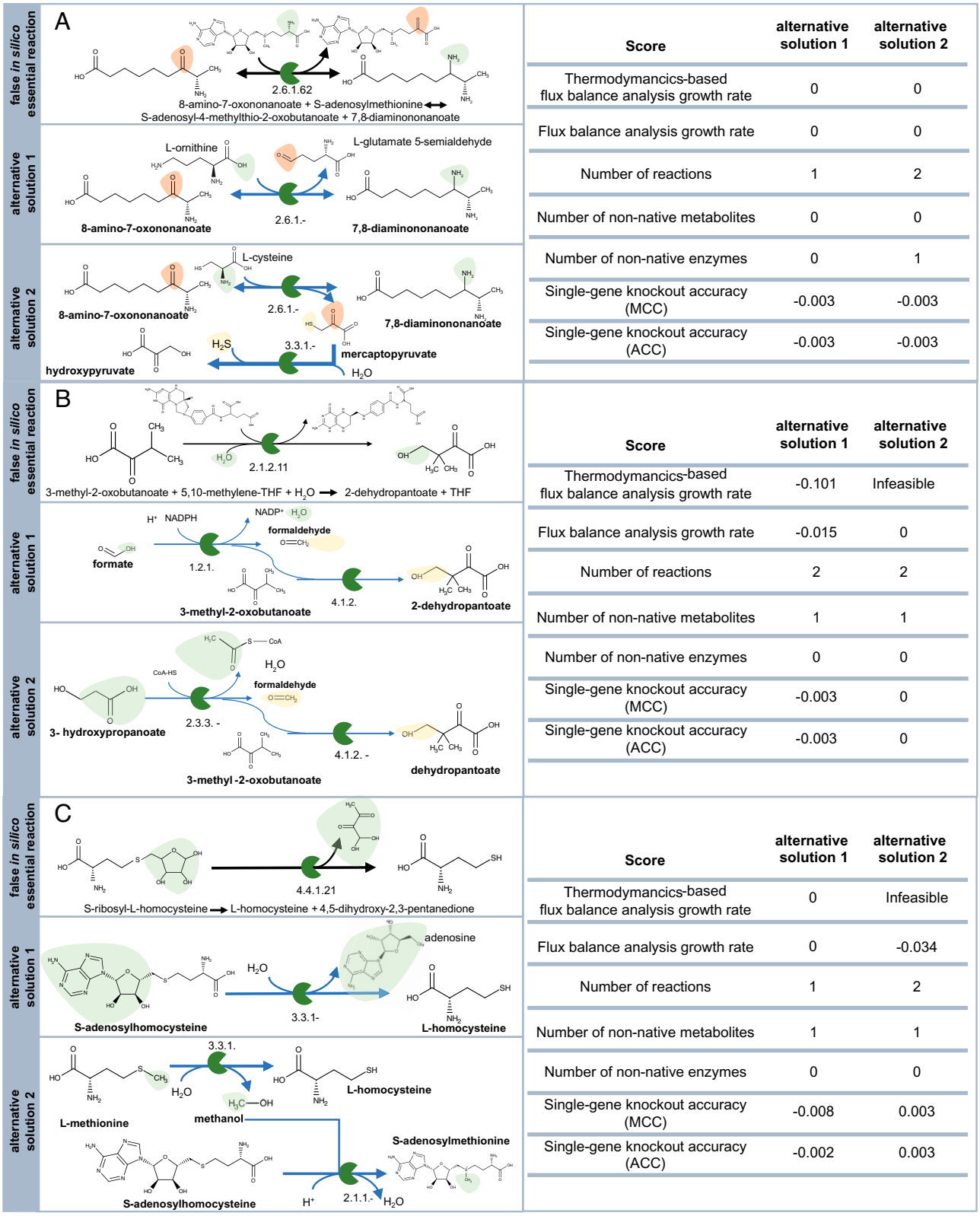
Fig. 2. (*A*) Comparison of essentiality between iML1515 and the extended networks. The 148 false-negative gene essentiality predictions are linked to 152 (orange) reactions in the model. Using the *E. coli metabolites subset* of ATLAS as a reaction pool for gap-filling, 86 of these gaps could be characterized (blue), while the *E. coli and yeast metabolites subset* of ATLAS rescued 93 reactions (green). (*B*) Contingency matrix for gene essentiality prediction accuracy of iML1515. The accuracy of the model is equal to 0.87 and the MCC equal is to 0.49. (*C*) Example of remaining gaps. The subsets of ATLAS used in this study could not rescue 59 false-negative reactions. The enzyme 4-amino-4-deoxychorismate synthase (EC 2.6.1.85) remains a false-negative.

this case, L-ornithine donates the amino-group and L-glutamate 5-semialdehyde is the byproduct. The reaction does not affect the predicted growth rate and does not require any additional enzymatic capabilities, and it improves the overall accuracy of the model in terms of predicting gene essentiality. To catalyze this reaction, BridgIT identified 12 candidate genes with an adequate BridgIT score (see *Materials and Methods*).

A second alternative for this step of the biotin biosynthesis pathway (Fig. 3 *A*, alternative solution 2) requires two novel reactions to fill the gap. The first reaction converts 8-amino-7-oxononanoate to 7,8-diaminononanoate by transferring the amino group from L-cysteine (EC 2.6.1.-), which produces mercaptopyruvate as a side-product. For this biotransformation, BridgIT suggests 19 candidate genes. The second reaction balances the production of mercaptopyruvate by converting it into hydroxypyruvate (EC 3.3.1.-) following a reaction mechanism that is not part of the original network, meaning the EC number is not part of the iML1515 GEM. This alternative solution is ranked lower than the previous one. Regardless, we identified one putative sequence to catalyze this reaction.

Another false-negative predicted reaction is the enzyme 3-methyl-2-oxobutanoate hydroxymethyl transferase that catalyzes the production of 2-dehydropantoate, a precursor of coenzyme A, from 3-methyl-2-oxobutanoate (EC 2.1.2.11). This enzyme is the product of the gene panB (b0134) that is predicted as a false-negative. NICEgame suggests 29 thermodynamically feasible solution sets, all involving the production of 2-dehydropantoate from 3-methyl-2-oxobutanoate and formaldehyde (EC 4.1.2.-), which is the orphan KEGG reaction R01216. To encode for this biotransformation, our method suggests 26 candidate genes. The first alternative (Fig. 3 *B*, alternative solution 1) contains a novel side reaction and describes the reduction of formate to formaldehyde (EC 1.2.1.-). BridgIT could not identify any candidate gene to encode this enzyme. The second alternative (Fig. 3 *B*, alternative solution 2) produces formaldehyde from 3-hydroxypropanoate through an acyltransferase (EC 2.3.3.-), but this novel reaction is not thermodynamically feasible in the desired direction. Thus, this alternative is discarded.

Another false-negative that can be reconciled by gap-filling with the *E. coli metabolites subset* is the gene luxS (b2687). Here, the S-ribosylhomocysteine cleavage enzyme, encoded by b2687, is responsible for the production of L-homocysteine, a precursor of L-methionine, from S-ribosyl-L-homocysteine (EC 4.4.1.21). Our workflow suggests 11 thermodynamically feasible solution sets. The first alternative produces L-homocysteine from an amylase acting on S-adenosylhomocysteine (EC 3.3.1.-), which is the KEGG reaction R00192 (Fig. 3 *C*, alternative solution 1). Though this enzymatic capability is not part of the original network, BridgIT could identify one candidate gene to encode for this enzyme. The second alternative (Fig. 3 *C*, alternative solution 2) uses the same reaction mechanism to produce L-homocysteine from L-methionine and a second reaction to balance the byproduct, methane. However, this solution set is thermodynamically infeasible and therefore rejected.



Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	0	0
Flux balance analysis growth rate	0	0
Number of reactions	1	2
Number of non-native metabolites	0	0
Number of non-native enzymes	0	1
Single-gene knockout accuracy (MCC)	-0.003	-0.003
Single-gene knockout accuracy (ACC)	-0.003	-0.003

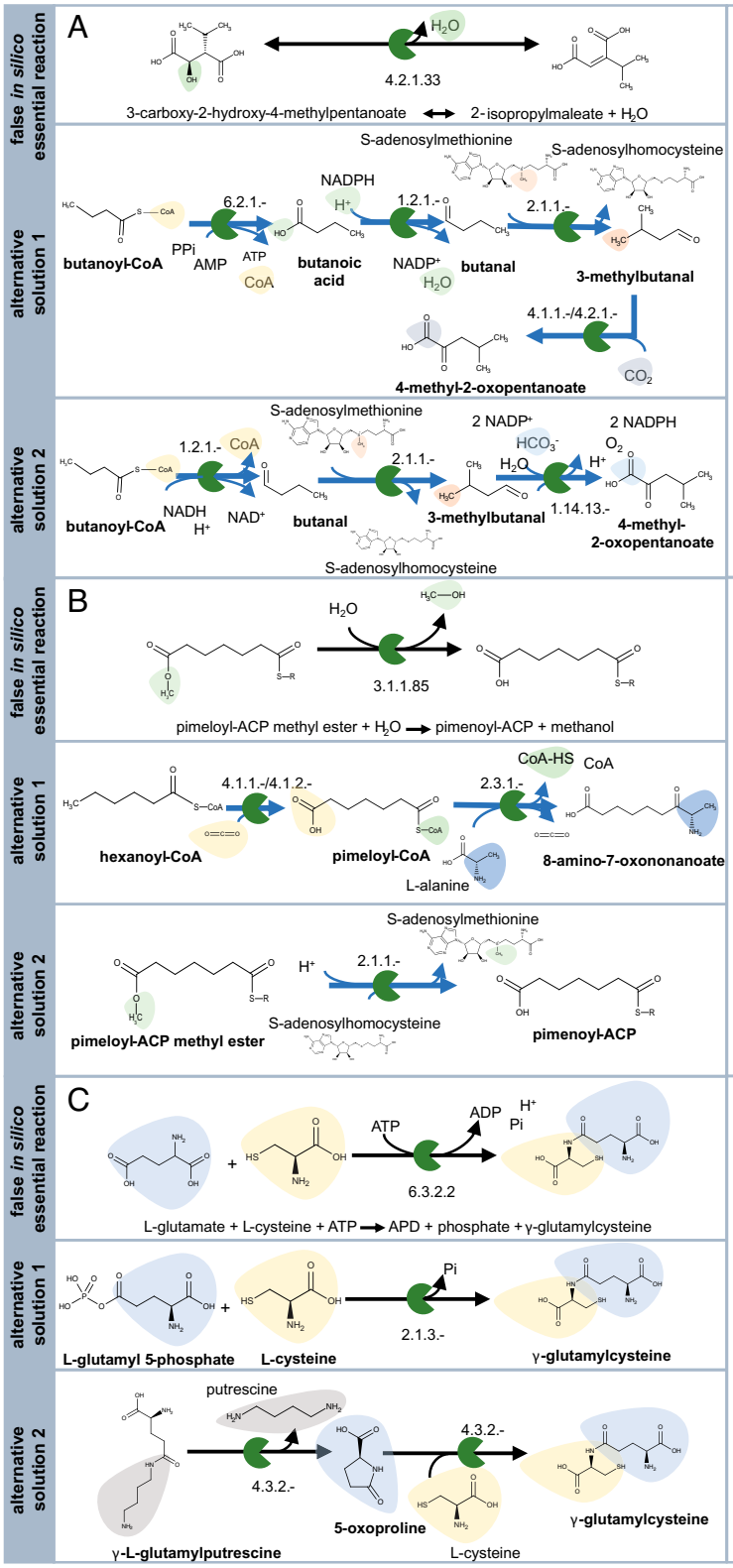
Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	-0.101	Infeasible
Flux balance analysis growth rate	-0.015	0
Number of reactions	2	2
Number of non-native metabolites	1	1
Number of non-native enzymes	0	0
Single-gene knockout accuracy (MCC)	-0.003	0
Single-gene knockout accuracy (ACC)	-0.003	0

Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	0	Infeasible
Flux balance analysis growth rate	0	-0.034
Number of reactions	1	2
Number of non-native metabolites	1	1
Number of non-native enzymes	0	0
Single-gene knockout accuracy (MCC)	-0.008	0.003
Single-gene knockout accuracy (ACC)	-0.002	0.003

Fig. 3. Cases incorrectly predicted as essential reactions (false-negatives) and alternative gap-filling reactions identified using the *E. coli metabolites subset*. (*A*) The reaction regulated by bioC in the original network, two gap-filling solutions, and their scores. (*B*) The reaction catalyzed by luxS in iML1515 and one thermodynamically favorable and one thermodynamically infeasible solution. (*C*) The reaction linked with the gene panB and two gap-filling solutions with their scores.

Biotransformations among *E. coli* and Yeast Metabolites Reconcile More Gaps. Including *E. coli* and yeast metabolites suggests more alternative solution sets for the already rescued false-negative reactions, and importantly, reactions for seven false-negative cases that had not yet been addressed (Dataset S2). One of the additionally rescued reactions is 3-isopropylmalate dehydratase, which interconverts 2-isopropylmaleate to 3-carboxy-2-hydroxy-4-methylpentanoate (EC 4.2.1.33). The reaction is

part of the leucine biosynthesis pathway, and it is encoded by two genes, leuD (b0071) and leuC (b0072). The first set of reactions (Fig. 4 *A*, alternative solution 1) produces 4-methyl-2-oxopentanoate, a precursor of leucine, from butanoyl-CoA in 4 steps. Three of the steps are novel reactions and the fourth is the KEGG reaction R01176, and BridgIT can identify candidate genes for two of them. The second set of reactions (Fig. 4 *A*, alternative solution 2) synthesizes 4-methyl-2-oxopentanoate again from butanoyl-CoA,



Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	-0.027	Infeasible
Flux balance analysis growth rate	-0.017	0.007
Number of reactions	4	3
Number of non-native metabolites	0	0
Number of non-native enzymes	1	1
Single-gene knockout accuracy (MCC)	-0.011	0
Single-gene knockout accuracy (ACC)	-0.003	0

Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	0	0
Flux balance analysis growth rate	0	0
Number of reactions	2	1
Number of non-native metabolites	0	0
Number of non-native enzymes	1	0
Single-gene knockout accuracy (MCC)	0.011	-0.003
Single-gene knockout accuracy (ACC)	-0.001	-0.003

Score	alternative solution 1	alternative solution 2
Thermodymancics-based flux balance analysis growth rate	0	0
Flux balance analysis growth rate	0	0
Number of reactions	1	2
Number of non-native metabolites	0	0
Number of non-native enzymes	0	1
Single-gene knockout accuracy (MCC)	-0.006	-0.003
Single-gene knockout accuracy (ACC)	-0.002	-0.001

Fig. 4. Cases incorrectly predicted as essential reactions (false-negatives) and alternative gap-filling reactions identified using the *E. coli and yeast metabolites subset*. (*A*) The reaction regulated by leuCD in the original network, one thermodynamically favorable and one thermodynamically infeasible gap-filling solution, and their scores. (*B*) The reactions catalyzed by bioH in iML1515, one low-ranking solution, and one high-ranking gap-filling solution. (*C*) The reaction linked with the gene gshA and two gap-filling solutions with their scores. The first alternative was generated by the *E. coli metabolites subset*, whereas the second one was generated by the *E. coli and yeast metabolites subset*.

via 3 novel reaction steps, but it is thermodynamically unfavorable. Both solution sets involve the metabolite 3-methylbutanal that is not part of the original iML1515 metabolic network. This compound has been characterized (21) as an alternative substrate of the enzyme 3-hydroxypropionaldehyde dehydrogenase (b1300), but to our knowledge it has not been detected as part of the metabolome of *E. coli*.

One reaction with that can be rescued via both subsets involves pimeloyl-ACP methyl ester esterase (EC 3.1.1.85), which is an enzyme encoded by the false-negative gene bioH (b3412). The enzyme is part of the biotin biosynthesis pathway, and it is responsible for the hydrolysis of pimeloyl-ACP methyl ester to pimeloyl-ACP, a precursor of biotin. To fill this gap, the *E. coli metabolites subset* provides three alternative solution sets of one reaction each. One solution (Fig. 4 *B*, alternative solution 2) produces pimeloyl-ACP by transferring the methyl group from pimeloyl-ACP methyl ester to S-adenosylhomocysteine to form S-adenosylmethionine (EC 2.1.1.-). BridgIT provides 11 genes to regulate this novel reaction. The *E. coli and yeast metabolites subset* provides one additional solution set (Fig. 4 *B*, alternative solution 1) of two steps: one novel reaction and the KEGG reaction R03210 describing the synthesis of 8-amino-7-oxononanoate, a successor metabolite of pimeloyl-ACP in the biotin biosynthesis pathway, itself made from hexanoyl-CoA with pimeloyl-CoA as an intermediate metabolite. Although pimeloyl-CoA is not part of the original reconstruction, it has been recently shown (22) to serve as the acyl chain donor of the 8-amino-7-oxononanoate synthase (bioF). BridgIT suggests 21 candidate genes to encode for this function, bioF among them. In the original reconstruction, bioF uses pimenoyl-ACP as substrate and is essential, but the gene is not essential in vitro, flagging it a false-negative gene. Interestingly, this solution set provides an alternative precursor for biotin, thus resolving the false-negative case of bioC (b0777) that is responsible for the synthesis of malonyl-CoA methyl ester, a precursor of pimenoyl-ACP and thus biotin. However, this solution is rejected since it adds redundancy to the model, having a Matthews Correlation Coefficient (23) (MCC) score equal to 0.0108, with 0 being uncorrelated and 1 being correlated, since the genes fabZ (b0180) and fabH (b1091) become false-positives after adding this solution set to the network.

Another false-negative gene that can be resolved using both metabolite subsets is the gene gshA (b2688), which regulates the synthesis of γ-glutamylcysteine from L-cysteine and L-glutamate (EC 6.3.2.2). With the *E. coli metabolites subset*, NICEgame provides 12 alternative thermodynamically feasible solution sets to fill in this gap. In the highest-ranking solution (Fig. 4 *C*, alternative solution 1), γ-glutamylcysteine is produced by L-cysteine and L-glutamyl 5-phosphate, releasing orthophosphate (EC 2.1.3.-). The *E. coli and yeast metabolites subset* provides three additional thermodynamically favorable solution sets, all involving the metabolite 5-oxoproline as an intermediate. This intermediate is not part of the metabolome of the original reconstruction, but it has been detected in *E. coli* (24). In the best-performing solution (Fig. 4 *C*, alternative solution 2), the first reaction describes the degradation of γ-L-glutamylputrescine to putrescine and 5-oxoproline and is novel, whereas the second reaction is a KEGG reaction (R02743) reconstructed in ATLAS. BridgIT identifies the gene chaC as a candidate to encode for this metabolic function.

Gene Annotation of Metabolic Gaps Identifies New Functions in *E. coli*. When gap-filling the metabolic network of *E. coli*, NICEgame suggested over 7,000 known and novel reactions, with over 6,600 among them part of thermodynamically feasible solution sets. To catalyze these reactions, BridgIT adequately identified candidate sequences in the genome of *E. coli* to catalyze 6,319 of these reactions (see *Materials and Methods* for scoring), which it assigned to 2,165 EC numbers (Dataset S3). Finally, we suggest 590 candidate promiscuous genes in the genome of *E. coli* to catalyze 6,118 reactions. In an example shown in Fig. 5 wherein γ-L-glutamylputrescine is degraded to 5-oxoproline and putrescine, we highlight how BridgIT found adequate similarity scores between an ATLAS novel reaction and five KEGG reference reactions. Here, though, only one reaction had a corresponding gene in *E. coli*, so BridgIT identifies the gene chaC as a promising candidate to catalyze this novel reaction.

Updated Genome-Scale Model of *E. coli* Shows Increased Essentiality Prediction Accuracy. Our approach expanded the original *E. coli* metabolic network by 77 reactions and 9 metabolites. We suggest 35 genes, only the top-rated BridgIT predictions, associated with these 77 reactions, of which 2 genes were not part of the original reconstruction (Dataset S4). Using our criteria and ranking method (see *Materials and Methods*), we extracted an updated version of the metabolism of *E. coli* strain MG1655, termed iEcoMG1655 (Fig. 6). The updated reconstruction includes 2,450 network reactions, 1,176 metabolites, and 1,517 genes, while it has an enhanced accuracy in gene essentiality prediction. iEcoMG1655 achieves a MCC equal to 0.60 and an accuracy measurement (ACC) (25) equal to 0.92, as compared to the performance of the previous best GEM iML1515 at 0.49 and 0.8, respectively, for the conditions examined in this study.

The added biochemistry reconciles metabolic gaps linked to the amino acid metabolism, cofactor metabolism and biosynthesis of cell membrane peptidoglycans (*SI Appendix*, Fig. S2). This result hints the contribution of underground metabolism in the biosynthetic pathways. It also highlights the aforementioned subsystems as targets for further research, in contrast to the central metabolism, where no gaps were identified based on our analysis (Dataset S1).

The added genes not part of the original reconstruction are ArcA and LacA. The first of these, ArcA (b4401), is part of the ArcAB (aerobic respiratory control) regulatory system (26), where ArcA and ArcB have been shown to regulate the expression of oxygen-requiring pathways (26). ArcAB has also been known to participate in the proper expression of catabolic genes for pyruvate utilization and sugar fermentation pathways (26). In our expanded reconstruction, this gene regulates the hydrolysis of N2-succinyl-L-arginine to urea and N2-succinyl-L-ornithine (EC 3.5.3.-), which provided an alternative pathway to compensate for the knockout of argG (b3172).

The second added gene, LacA (b0342), encodes the enzyme galactoside O-acetyltransferase, which catalyzes the transfer of an acetyl group from acetyl-CoA to the 6-hydroxyl of some galactopyranosides (27). This enzyme is known to act on a broad range of substrates and can acetylate galactosides, thiogalactosides, glucosides, and lactosides (27). In our expanded reconstruction, it participates in lipopolysaccharide biosynthesis/recycling and catalyzes the degradation of dodecanoyl-KDO2-lipid IV(A) to KDO2-lipid A. Altogether, its addition compensates for the knockout of lpxM (b1855).

The 33 genes that already are part of the model show substrate or mechanism promiscuity. For example, in the original reconstruction, galK (b0757) encodes for the enzyme galactokinase, which is responsible for the phosphorylation of D-galactose (EC 2.7.1.6). This enzyme shows substrate promiscuity, with BridgIT suggesting it can phosphorylate ADP-D-glycero-D-manno-heptose, 5-methylthio-D-ribose, D-glycero-beta-D-manno-heptose-7-phosphate, 6-hydroxymethyl-7,8-dihydropterin, and

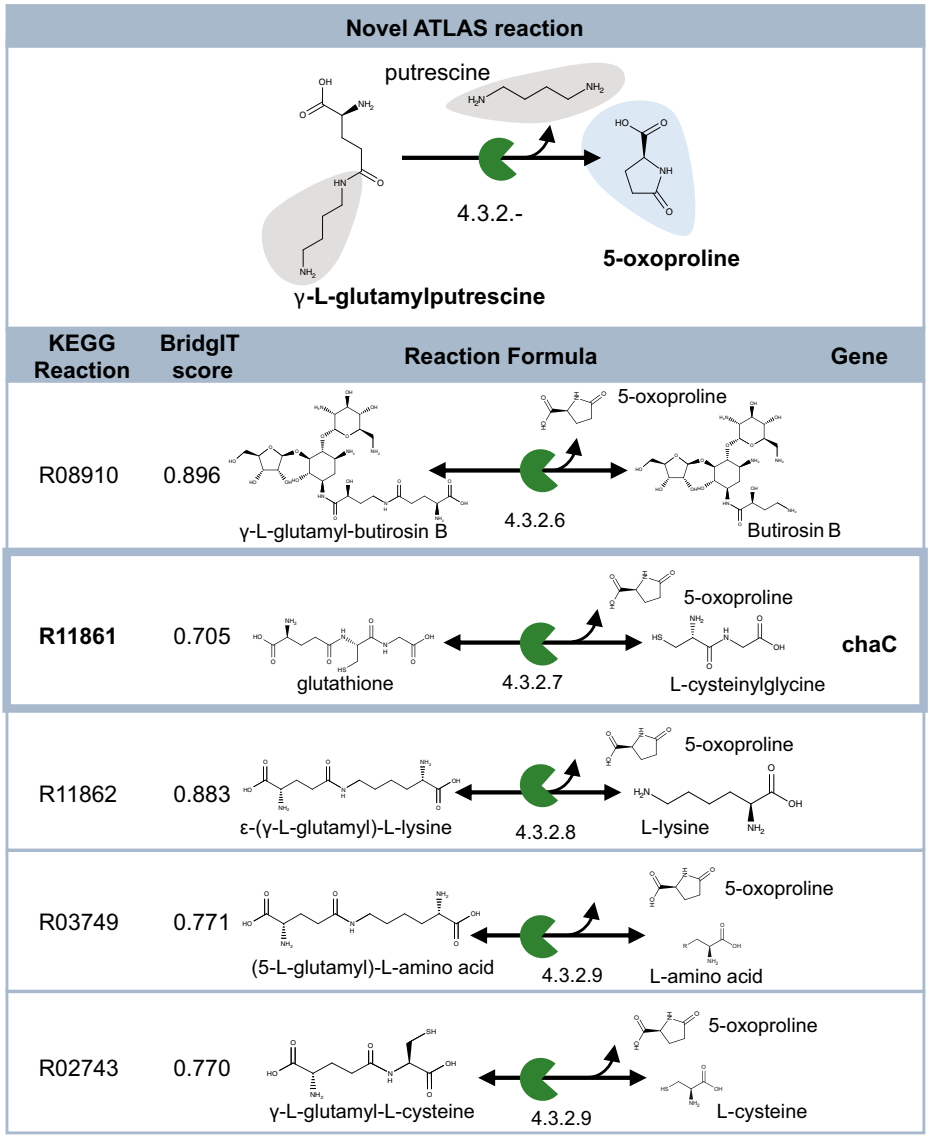


Fig. 5. Example of BridgIT suggestions for genes to catalyze the gap-filled reactions. BridgIT identified five reactions similar to the novel reaction that degrades γ-L-glutamylputrescine to 5-oxoproline and putrescine. Out of these five reactions, only R11861 is linked to a sequence in the genome of *E. coli*.

D-ribose-5-phosphate. Similarly, whereas xapA (b2407) acts as a glycosyltransferase (EC 2.4.2.-) in the iML1515 network, BridgIT suggests it can also encode for phosphotransferases (EC 2.7.4.-), and though ydfG (b1539) acts as dehydrogenase (EC 1.1.1.-) in the iML1515 network, BridgIT suggests it can also act as a carbon-carbon lyase (EC 4.1.1.-, 4.1.2.-). Nine of these 33 genes already show substrate promiscuity in the iML1515 network (Dataset S4).

Although NICEgame suggests thermodynamically feasible alternatives to rescue 93 reactions, solutions for 36 reactions were not added to the updated reconstruction for two reasons. In the first situation, we choose to not include a gap-filling solution when all the identified solution sets added redundancy in the metabolic network, resulting in an increase of the false-positive gene essentiality predictions. The second situation was one in which BridgIT identified an essential or a false-negative gene to catalyze the suggested reactions. The addition of such solution sets to the network would not resolve the gap.

NICEgame Offers Improved Gap-Filling Performance over Published Approaches. To evaluate the performance of NICEgame against existing gap-filling approaches, we performed three comparative studies. In the first, we repeated the generation of gap-filling alternative solutions using our in-house algorithm, but this time used only known biochemical reactions, meaning only the *E. coli and yeast metabolites subset* of the KEGG database (28), as a pool for the gap-filling. The second study compares our gap-filling algorithm to published algorithms, as in the algorithms included in the RAVEN (29) and COBRA (30) toolboxes, using the *E. coli and yeast metabolites subset*. In the third, we compared NICEgame against the CarveMe (31) gap-filling approach.

In the first study, the *E. coli and yeast metabolites subset* of KEGG identified thermodynamically feasible gap-filling solutions for 53 out of the 152 target reactions (Dataset S5), as compared to the 93 rescued reactions managed when including the *E. coli and yeast metabolites subset* of ATLAS. The average number of solutions per rescued reaction is 2.3 for the *E. coli and yeast metabolites subset* of KEGG versus 252.2 for that subset of ATLAS. However, this subset of KEGG did suggest solutions for 8 reactions that the ATLAS equivalent cannot rescue. A further analysis to understand why the ATLAS database could not capture these 8 solutions (Dataset S6) found that these KEGG reactions could not be reconstructed in ATLAS because of the complex structure of the metabolites.

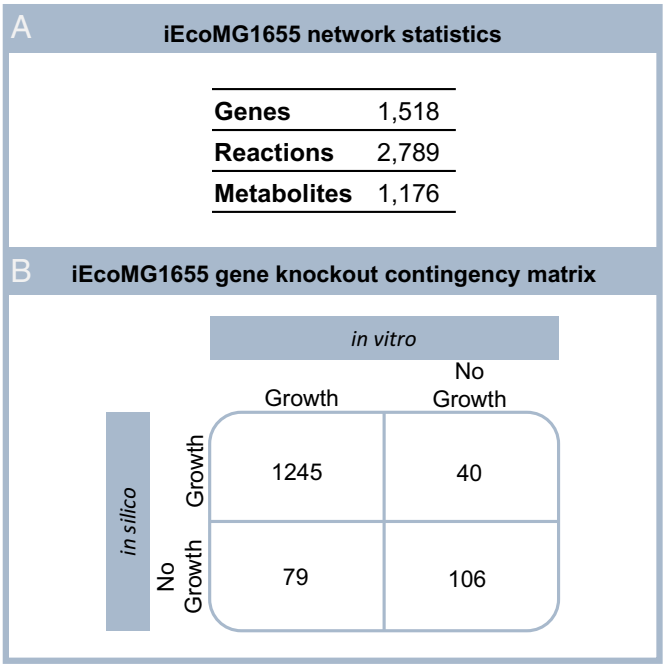


Fig. 6. (*A*) The iEcoMG1655 network statistics. The gap-filled network contains 77 novel reactions, two additional genes, and nine new metabolites. (*B*) Contingency matrix for gene essentiality prediction accuracy of iEcoMG1655. NICEgame could reconcile 69 out of the 148 false-negative gene essentiality predictions leading to an increased accuracy, MCC = 0.60.

We next compared NICEgame to the gap-filling approaches RAVEN and COBRA, which both also use a mixed-integer linear programming (MILP) algorithm. With the RAVEN algorithm, we adjusted the input parameters to include as few reactions as possible from the database that would still satisfy the model constraints, such as the mass balances and the basal growth rate under the defined media. The mathematical formulation in RAVEN is very similar to the one defined by the NICEgame, however RAVEN does not generate alternative solutions, not allowing to choose the most biologically relevant solution and to account for problems in the first solution. This resulted in thermodynamically feasible solutions for 67 out of the 152 target reactions (Dataset S7). Similarly, COBRA also minimizes the number of reactions added from the database to the model to achieve a given metabolic task. The COBRA algorithm also gives different weights to the reactions of the database, penalizing uptakes and transporters as compared to metabolic reactions. A difference with COBRA is that it accounts for alternative solutions by assigning a bigger weight to reactions that have already appeared in previous solutions. To test COBRA against NICEgame, we demanded 10 alternatives per target reaction. Here, the COBRA gap-filling approach rescued 68 reactions. However, the same solution often reappeared, and there is no systematic enumeration of all minimal solution sets (Dataset S8). As an example of the reaction ANPRT, the algorithm identifies the ATLAS reaction rat45874 as a solution four times. The size of the remaining six solutions ranged between two and three reactions per solution set.

For the final comparison, the CarveMe gap-filling approach is also based on a MILP formulation that aims to add the minimal number of reactions from the universal bacterial model (31) to the GEM under curation. The method does not generate alternative solutions sets. Comparing NICEgame to CarveMe, we identified thermodynamically feasible solutions for 33 target reactions and reconciled 24 gaps not curated by NICEgame (Dataset S9). These additional gap-filling alternatives above those of NICEgame came from the use of a different database, which could be considered in the future for NICEgame. For example, the *E. coli and yeast metabolites subset* of ATLAS can gap-fill the target reaction AICART with solution sets of one reaction, whereas the universal bacterial model provides a solution of 20 reactions. However, the method suggests transporters, such as CAt6, CITt13, and Cuabc, and pseudoreactions, such as sink_4hba_c, as part of the solutions. These pseudoreactions are mathematical inventions to represent parts of the metabolism that are unknown or not important under certain scope, and thus not described in a mechanistic way. This type of solution is not desired since they do not explicitly describe the metabolism and are trivial.

Validation of the NICEgame Workflow. To validate our workflow, a series of tests were conducted to compare the performance of the iEcoMG1655 network against the iML1515 network. Single gene knock-out *in silico* deletions were performed on 15 carbon sources (2) to assess the effect of the added biochemistry (gene-reaction pairs) on other phenotypes (Dataset S10). Even though the choice of the gap-filling solutions did not consider the gene essentiality data for these 15 carbon sources, the iEcoMG1655 network showed an increased ACC compared to iML1515 in all cases. Furthermore, the MCC remains at the same levels for the two networks. Overall, the iEcoMG1655 network shows a better performance regarding the FN and, consequently, the TP predictions. Additionally, double gene knock-out simulations with glucose as the carbon source for the iML1515 and the iEcoMG1655 networks. Driven by data availability, we generated in silico knockouts strains of the gdhA mutant and contrasted our results against experimental data (32) (Dataset S11). The two networks show similar performance. We further examined the feasible flux ranges in the two networks. The iEcoMG1655 network shows more flexibility compared to the iML1515 network, for all the 16 carbon sources examined (Dataset S12). The average flux range increase varies from about 13.1% to 37.7%. Furthermore, more reactions can carry flux in the iEcoMG1655 network. The unblocked reactions mainly are assigned to the metabolic subsystems of transport and exchange of metabolites, the cofactor biosynthesis, and the alternate carbon metabolism (Dataset S12). Finally, a study was performed to evaluate the sensitivity of the NICEgame workflow in retrieving known gene-reaction pairs (Dataset S13). To this end, we removed essential genes in the iML1515, one at a time, and all corresponding reactions, generating *knock-out* GEMs. NICEgame could correctly retrieve the correct gene-reaction pairs for 98% of the cases.

Discussion

To address the lack of tools available for the systematic identification and reconciliation of metabolic gaps at a genome scale, we have herein presented the NICEgame workflow. We applied our workflow to the most recently published *E. coli* GEM, iML1515, which contained 148 false-negative essential genes for the media conditions used in this study. The newly included biochemistry, meaning reactions and catalyzing genes, reconciled 69 false-negative genes. However, there are still gaps in the metabolism of *E. coli* that remain unresolved. This is likely due to our selected subset from the ATLAS of Biochemistry. In the future, more gaps could be filled by using a bigger subset to integrate more information from the ATLAS of Biochemistry with the *E. coli* metabolic network.

A benefit here of our approach is that it was designed to be a living tool that adapts with new scientific discoveries; users can revisit and reevaluate the suggested solution sets for each rescued reaction when new quantitative and qualitative data are released. The results generated by NICEgame are experimentally testable, and future experiments designed to test these findings would provide interesting information about our understanding of *E. coli* metabolism. For instance, the validity of the gap-filling solutions suggested in this study can be examined by double-gene knockout studies, meaning the simultaneous deletion of the gene predicted as false negative and the sequence predicted by BridgIT to rescue it. A further indication of the validity of our predictions would be the in vitro identification of the newly added metabolites in the iEcoMG1655 network, which could be done through liquid chromatography-mass spectrometry (LC-MS) or NMR spectroscopy techniques.

In the future, the NICEgame workflow can be extended to additionally (i) account for transport reactions (33), apart from biotransformations, for the reconciliation of metabolic gaps, (ii) handle higher order gene deletions (34) and (iii) aim to the restoration of flux through blocked reactions (33).

Overall, NICEgame should advance studies and applications of any organismal metabolism, as it is applicable to any existing GEM. Filling metabolic gaps will increase the predictive capability of GEMs and will thus help advance the fields of biotechnology and biomedicine, by suggesting more efficient pathways and favorable conditions for bio-manufacturing, as well as novel drug targets and therapeutic strategies. Finally, our method and the library of alternative solution sets can also be used as a resource of novelty and discovery in metabolic engineering to design strains with improved performance, such as higher biomass or product yield.

Materials and Methods

Reconciliation of Annotation. Since the ATLAS of Biochemistry is KEGG (28)-based, the metabolite identifiers of the iML1515 network first needed to be translated to KEGG notation. For this purpose, we extracted and compared manually information from the KEGG and the BiGG (35) databases. Overall, 909 out of the 1,169 metabolites were mapped to a KEGG ID. Apart from the KEGG database, the metabolites of the iML1515 were also mapped similarly to the SEED (36) database to impose thermodynamic constraints on the model. Here, 1,106 out of the 1,169 metabolites were mapped to a unique SEED ID.

Databases Used for Gap-Filling. To reduce ATLAS in the two subdatabases used in this study, i.e., the *E. coli metabolites subset* and the *E. coli and yeast metabolites subset* (Dataset S14), we extracted all ATLAS reactions that integrate intra- and extracellular compounds that are already part of the iML1515 GEM and reactions that integrate compounds that are already part of the iML1515 and the Yeast8 (3) GEMs, respectively. Likewise, the *E. coli and yeast metabolites subset* of KEGG contains only KEGG (2018 version) reactions among metabolites included in the iML1515 and the Yeast8 (3) GEMs. For gap-filling using CarveMe, we used the universal bacterial model (31), a compartmentalized model that contains transporters and pseudoreactions. The size of the databases is shown in Table 1.

Table 1. Size of the databases used for gap-filling

	Reactions	Metabolites
E. coli metabolites subset of ATLAS	9,810	778
E. coli and yeast metabolites subset of ATLAS	13,298	1,050
E. coli and yeast metabolites subset of KEGG	1,756	1,128
Universal bacterial model	5,532	2,861

Gap-Filling Formulation. The gap-filling algorithm uses a MILP formulation and generates binary use variables for each reaction in the database. These variables indicate whether flux is allowed through a reaction or not. The gap-filling algorithm is a parsimonious algorithm whose objective is to minimize the number of active reactions in the metabolic network, demanding at the same time a basal flux through the biomass reaction in the wild-type model. The mathematical formulation of the MILP problem is:

$$\begin{aligned} &\max_{z} \sum z_i \\ &s.t. \\ &1 * F_i + 1 * R_i + M * z_i \leq M, \ (i) \\ &1 * F_i + 1 * R_i + 1 * z_i \geq m, \ (ii) \\ &F_{biomass} \geq threshold * (WT\ growth\ rate), \end{aligned} \quad [1]$$

where F_i represents the flux variables of the irreversible forward reactions, R_i are the flux variables of the irreversible backward component reactions of the reversible reaction i, $F_{biomass}$ is the flux variable of the irreversible forward biomass reaction, *WT growth rate* is the growth rate predicted by the wild-type model for the given media, *threshold* is the parameter that defines the minimally required growth rate as a percentage of the *WT growth rate*, M is a big-M value, m is a small value, and z_i are the binary use variables. For this study, the gap-filling was performed for M9 glucose minimal media and aerobic conditions and the wild-type biomass reaction as an objective function i.e., BIOMASS_Ec_iML1515_WT_75p37M. Neither the carbon nor oxygen uptake were constrained, i.e., the maximum allowed uptake rate was set to 50 $\frac{mmol}{gDW\ h}$.

Every time the solver identifies a solution, the solution is integrated as a cut constraint to the MILP problem, so the solver cannot identify the same solution more than once:

$$\sum z_k > 0. \quad [2]$$

To avoid generating long pathways, we demanded that the minimum solution size be less than ten reactions and the subsequent solutions can be at most five reactions bigger than the minimum size solution.

Identification of Metabolic Gaps. Gene essentiality data (19) were used to identify putative false-negative reactions. We considered M9 glucose minimal media and aerobic conditions and the wild-type biomass reaction as an objective function. We performed a single gene deletion analysis, where a gene was considered essential in silico if the growth rate of the knockout mutant was less than 10% of the growth rate of the wild-type. This analysis revealed 258 genes essential in silico, with 105 of them also being essential in vitro, while 7 of them were not represented in the experimental data. We identified all reactions associated with the 148 remaining genes, which numbered 200 in total, and after a single reaction deletion analysis, we concluded that 152 of them are essential in silico. We consider that these 152 are falsely essential and thus constitute the target reactions for gap-filling.

Scoring the Alternatives. For each gap-filled model, the output of the framework is a set of ranked alternatives for each rescued reaction. The main criteria for ranking the different alternatives are the thermodynamic feasibility of the solution and minimum impact on the model, which means that a solution is ranked lower the more it alters the biochemistry and the predictive capability of the model.

To examine the maximum biomass yield under thermodynamic constraints, thermodynamics-based flux balance analysis (TFA) (37) was carried out for each alternative. To examine the maximum biomass yield for each alternative solution, the rescued reaction was blocked, and flux was allowed through the set of alternative reactions. To define the score, the ratios of the optimal growth rate of the wild-type GEM to the gap-filled GEM were calculated, and one point is subtracted. If the result of the subtraction was greater than zero, the addition of the alternative reduced the performance compared to the original GEM, whereas when the result of the subtraction was less than zero, the addition of the alternative led to higher performance compared to the original GEM. If a gap-filled GEM did not predict growth when it was thermodynamically restricted, the alternative was rejected. The performance of the gap-filled models without thermodynamic constraints was also tested. Here to examine the optimal growth rate, flux balance analysis (FBA) was carried out for each alternative and the results were analyzed similarly to the TFA test.

The number of reactions of each alternative solution was also considered. Since the set of reactions of each alternative replaced one reaction in the model, one point was subtracted from the number of reactions in the solution set. Organisms tend to favor shorter paths (18), so the alternatives that integrated fewer reactions were ranked higher than those that integrated more reactions. An extra point was added for every reaction that was linked to a third level EC number that was not included in the original GEM, as the integration of such reactions also entailed the integration of new enzymatic capabilities into the model. At the metabolite level, for every unique nonnative metabolite, one point was added.

Lastly, we tested the ability of the models to properly predict gene essentiality. To this end, the overall accuracy (ACC) (25) and MCC (23) were calculated for each gap-filled model and were compared to the wild-type.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}, \quad [3]$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad [4]$$

where TP stands for true-positive, TN for true-negative, FP for false-positive, and FN for false-negative gene essentiality predictions.

The values of all the scores were added, and the alternatives were ranked. The closer the absolute value of the score was to zero, the more similar the performance and the biochemistry of the gap-filled model was to the original model.

Enzyme Annotation Using BridgIT. For annotating ATLAS reactions, we used the online version of the BridgIT method with default parameters as discussed in the original paper (17). The BridgIT method is inspired by the theory of a lock and key, which assumes that two similar reactions will be catalyzed by the same enzyme. BridgIT compares the similarity of each input reaction with all the nonorphan characterized metabolic reactions cataloged in the KEGG database (reference reactions) and proposes enzymes associated with the most similar reference reactions as the best candidate for the input reaction. Therefore, BridgIT systematically screens for the best promiscuous candidate enzymes that might be able to catalyze the input reaction. The degree of similarity or probability of catalyzation is quantified in the BridgIT score, ranging from zero (no similarity) to one (identical). The optimal threshold value for the BridgIT score is 0.3, meaning predictions with a score higher than 0.3 are considered promising (17). For each input reaction, BridgIT outputs a list of enzymes with their EC number ranked in descending order based on BridgIT score. Then, the EC number is used to query the Uniport (38) database for the corresponding protein sequences in the organism of interest, which in this study was *E. coli* K12.

Choosing Gap-Filling Solutions: The iEcoMG1655 Network. The NICEgame workflow generated over 47,000 solutions for the rescued reactions. However, a minimal set out of them was used to generate the iEcoMG1655 network. Thermodynamically infeasible solutions and solutions that decreased the accuracy of predictions (solutions that reduce ACC and/or MCC) were rejected (Stage 1 of solution evaluation). From the remaining solutions, those with BridgIT predictions and those increasing the accuracy of predictions (solutions that improve ACC and/or MCC) were prioritized (Stage 2 of solution evaluation). Then, the remaining scores were considered (Stage 3 of solution evaluation). In the case of solutions set that made it to Stage 3 and had identical scores, all alternatives were added. To assign genes to the 77 reactions, we chose the BridgIT predictions with the highest BridgIT scores. Predictions involving an essential or a false-negative gene were rejected.

Validation of the NICEgame Workflow. Single gene essentiality analyses with thermodynamic constraints were performed on the iML1515 and iEcoMG1655 networks on minimal media and aerobic conditions for 15 carbon sources (Dataset S10). Neither the carbon nor oxygen uptake was constrained. The double gene deletion was performed for both networks with thermodynamic constraints and for the aerobic glucose minimal media condition. The allowable flux ranges were calculated for both networks by performing flux variability analysis with thermodynamic constraints on glucose and the 15 other carbon sources. Finally, we removed the 257 essential genes on glucose minimal media and aerobic conditions in the iML1515, one at a time, and all corresponding reactions, generating 257 *knock-out* GEMs. We then used the NICEgame workflow to gap-fill each one of the *knock-out* GEMs by allowing the generation of 1,000 alternatives of minimal size, or bigger size if required, and the *E. coli metabolites subset* as a database for gap-filling. In the case that the gap was filled, we filtered the solutions to identify if they contain the original essential reactions removed from the iML1515 network to generate the *knock-out* GEM (Datasets S13 and S15). We then used the BridgIT algorithm to associate genes to the reactions (Datasets S13 and S16). We finally calculated the percentage of retrieved gene-reaction pairs. For this analysis, we first examined whether the removed reaction sets are included in ATLAS (Dataset S13). Out of the 257 *knock-out* GEMs 146 were linked to reaction sets that were not part of ATLAS. We thus consider the remaining 111 *knock-out* GEMs.

Software. This work was supported by EPFL through the use of the facilities of its Scientific IT and Application Support Center. We performed the gap-filling using the defined MILP formulation in MATLAB (2016a and 2018a) and IBM ILOG Cplex 12.7.1 as a solver. The simulations were run on a high-performance computing cluster of 408 nodes. We used two CPUs per simulation and 3,875 MB per CPU. One simulation was defined for a unique combination of parameters. The analysis of the gap-filling solutions was performed on Mac Pro-32 GB in MATLAB 2017a and IBM ILOG Cplex 12.7.1 as a solver. The gap-filling with RAVEN was performed with Gurobi Optimizer Version 9.3 as a solver. The gap-filling with COBRA and CarveMe was performed in python 3.6 and IBM ILOG Cplex 12.8.0 as a solver.

Data, Materials, and Software Availability. All study data are included in the article and/or supporting information. The code and model have been deposited in Github (https://github.com/EPFL-LCSB/NICEgame) (39).

ACKNOWLEDGMENTS. The authors would like to thank Dr. Alan R. Pacheco for the valuable discussions and feedback, Anastasia Sveshnikova for her valuable contribution and Dr. Kaycie Butler for her valuable input on the wording and structure of this paper. Funding for this work was provided by the Swiss National Science Foundation (SNSF): grant 200021_188623, NCCR Microbiomes, a National Centre of Competence in Research (grant number 180575), SystemsX.ch MicroScapeX grant 2013/158, and SystemsX.ch MalarX grant 2013/155, European Union's Horizon 2020 research and innovation programme grants: PacMen, under the Marie Sklodowska-Curie grant agreement No 72228, and ShikiFactory100, under grant agreement 814408, Swedish Research Council Vetenskapsradet (grant no. 2016-06160), and the École Polytechnique Fédérale de Lausanne.

Author affiliations: [a]Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne, EPFL, 1015 Lausanne, Switzerland; [b]Department of Chemical Engineering, University of Patras, 26504 Patras, Greece; and [c]Institute of Chemical Engineering Sciences (FORTH/ICE-HT), 26504 Patras, Greece

Author contributions: E.V., A.C.P. and V.H. conceptualized the study. E.V., A.C.P., H.M., Y.F., N.H., M.A., J.H. performed the data curation. E.V., A.C.P. and Y.F. developed the software. E.V., A.C.P., M.A. and V.H. developed the methodology. E.V., A.C.P., and H.M. carried out the formal analysis. Investigation was conducted by E.V., A.C.P., H.M., Y.F. and V.H.. E.V. and A.C.P. wrote the manuscript. E.V., A.C.P., and V.H. developed the visualizations. All authors edited and reviewed the manuscript. A.C.P., S.P. and V.H. managed and supervised the project.

[1]E.V. and A.C.P. contributed equally to this work.

[2]Present address: Department of Genetics, Harvard Medical School, Boston, MA 02115.

[3]Present address: Pharmaceutical Research and Early Development, Roche Glycart AG, 8952 Schlieren, Switzerland.

[4]Present address: Department of Biology, University of Texas at Arlington, Arlington, TX 76019.

[5]Present address: Novigenix, 1066 Epalinges, Switzerland.

[6]Present address: Computational and Systems Biology, Biozentrum, University of Basel, 4056 Basel, Switzerland.

[7]Present address: Department of Environmental Chemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, CH-8600 Dübendorf, Switzerland.

[8]To whom correspondence may be addressed. Email: vassily.hatzimanikatis@epfl.ch.

1. S. Ghatak, Z. A. King, A. Sastry, B. O. Palsson, The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Res.* **47**, 2446–2454 (2019).
2. J. M. Monk *et al.*, iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* **35**, 904–908 (2017).
3. H. Lu *et al.*, A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).
4. I. Thiele *et al.*, A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst. Biol.* **5**, 8 (2011).
5. A. Chiappino-Pepe, S. Tymoshenko, M. Ataman, D. Soldati-Favre, V. Hatzimanikatis, Bioenergetics-based modeling of Plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks. *PLOS Comput. Biol.* **13**, e1005397 (2017).
6. J. Jansma, S. El Aidy, Understanding the host-microbe interactions using metabolic modeling. *Microbiome* **9**, 16 (2021).
7. R. R. Stanway *et al.*, Genome-scale identification of essential metabolic processes for targeting the plasmodium liver stage. *Cell* **179**, 1112–1128.e26 (2019).
8. B. Kim, W. J. Kim, D. I. Kim, S. Y. Lee, Applications of genome-scale metabolic network model in metabolic engineering. *J. Ind. Microbiol. Biotechnol.* **42**, 339–348 (2015).
9. C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, S. Y. Lee, Current status and applications of genome-scale metabolic models. *Genome Biol.* **20**, 121 (2019).
10. E. Gasperskaja, V. Kučinskas, The most common technologies and tools for functional genome analysis. *Acta Med. Litu.* **24**, 1–11 (2017).
11. G. F. Ejigu, J. Jung, Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology (Basel)* **9**, 1–27 (2020).
12. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
13. N. Hadadi, V. Hatzimanikatis, Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
14. C. E. Lawson *et al.*, Machine learning for metabolic engineering: A review. *Metab. Eng.* **63**, 34–60 (2020).
15. N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, V. Hatzimanikatis, ATLAS of biochemistry: A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
16. J. Hafner, H. MohammadiPeyhani, A. Sveshnikova, A. Scheidegger, V. Hatzimanikatis, Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *ACS Synth. Biol.* **9**, 1479–1482 (2020).
17. N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, V. Hatzimanikatis, Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7298–7307 (2019).
18. Z. Abd Algfoor, M. Shahrizal Sunar, A. Abdullah, H. Kolivand, Identification of metabolic pathways using pathfinding approaches: A systematic review. *Brief. Funct. Genomics* **16**, 87–98 (2017).
19. E. C. A. Goodall *et al.*, The essential genome of *Escherichia coli* K-12. *MBio* **9**, e02096-17 (2018).
20. Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes: International Union of Biochemistry and Molecular Biology. https://archive.org/details/enzymenomenclatu0000inte_d6c2.
21. P. Falkenberg, A. R. Strøm, Purification and characterization of osmoregulatory betaine aldehyde dehydrogenase of *Escherichia coli*. *Biochim. Biophys. Acta* **1034**, 253–259 (1990). Accessed 19 October 2022.
22. M. Manandhar, J. E. Cronan, A canonical biotin synthesis enzyme, 8-amino-7- oxononanoate synthase (BioF), utilizes different acyl chain donors in *Bacillus subtilis* and *Escherichia coli*. *Appl. Environ. Microbiol.* **84**, e02084-17 (2017).
23. B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
24. N. Ishii *et al.*, Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science (80-.)* **316**, 593–597 (2007).
25. C. E. Metz, Basic Principles of ROC Analysis. https://www.sciencedirect.com/science/article/pii/S0001299878800142. Accessed 28 February 2022.
26. K. A. Salmon *et al.*, Global gene expression profiling in *Escherichia coli* K12: Effects of oxygen availability and ArcA. *J. Biol. Chem.* **280**, 15084–15096 (2005).
27. K. J. Andrews, E. C. C. Lin, Thiogalactoside transacetylase of the lactose operon as an enzyme for detoxification. *J. Bacteriol.* **128**, 510–513 (1976).
28. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
29. H. Wang *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).
30. L. Heirendt *et al.*, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
31. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
32. J. P. Côté *et al.*, The genome-wide interaction network of nutrient stress genes in *Escherichia coli*. *MBio* **7**, e01714-16 (2016).
33. V. Satish Kumar, M. S. Dasika, C. D. Maranas, Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).
34. A. R. Zomorrodi, C. D. Maranas, Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC Syst. Biol.* **4**, 178 (2010).
35. C. J. Norsigian *et al.*, BiGG Models 2020: Multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **48**, D402–D406 (2020).
36. C. S. Henry *et al.*, High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–982 (2010).
37. P. Salvy *et al.*, pyTFA and matTFA: A Python package and a Matlab toolbox for thermodynamics-based flux analysis. *Bioinformatics* **35**, 167–169 (2019).
38. T. U. Consortium *et al.*, UniProt Consortium, UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
39. E. Vayena *et al.*, A workflow for annotating the knowledge gaps in metabolic reconstructions using known and hypothetical reactions. Github. https://github.com/EPFL-LCSB/NICEgame. Deposited 23 August 2022.