Teasing out Missing Reactions in Genome-scale Metabolic Networks through Deep Learning

Can Chen[1,†], Chen Liao[2,†] & Yang-Yu Liu[1,*]

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA*

[2]*Program for Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA*

† *These authors contributed equally to this work.*

* *To whom correspondence should be addressed (yyl@channing.harvard.edu).*

GEnome-scale Metabolic models (GEMs) are powerful tools to predict cellular metabolism and physiological states in living organisms. However, due to our imperfect knowledge of metabolic processes, even highly curated GEMs have knowledge gaps, e.g., reactions may be missing, and unknown gene products may catalyze known reactions. An algorithmic approach to predicting missing reactions in a GEM is a crucial step of gap-filling. Despite existing efforts of predicting missing reactions in GEMs, we still lack a method that is both efficient and accurate. The majority of existing methods require phenotypic or taxonomic information to predict missing reactions. Here we present a deep learning-based method — CHEbyshev Spctral HyperlInk pREdictor (CHESHIRE) — to predict missing reactions of GEMs purely from the metabolic network topology. We demonstrate compelling evidence that CHESHIRE outperforms other topology-based methods in 108 high-quality GEMs.

Furthermore, CHESHIRE is able to improve the prediction of fermentation products over 24 draft GEMs reconstructed from an automatic pipeline. Both the internal and external validation results suggest that CHESHIRE is a powerful tool for GEM curation to reveal unknown links between reactions and observed metabolic phenotypes.

Introduction

As a mathematical representation of the metabolism for an organism, the GEnome-scale Metabolic model (GEM) offers a comprehensive gene–reaction–metabolite connectivity through two matrices: the stoichiometric matrix associating metabolites with their reactions; and the reaction-gene matrix associating reactions with their corresponding enzymes and genes[1,2]. GEMs are powerful computational tools to predict metabolic fluxes in living organisms[2,3]. Used alone or integrated with high-throughput data, GEMs can produce mechanistic insights and falsifiable predictions that progressively advance various disciplines in biomedical sciences[4,5], including metabolic engineering[6,7], microbial ecology[8], and drug discovery[9]. Recently, the rapid growth in whole-genome sequencing data[10] has triggered a surge in draft GEMs[11,12]. Yet, these initial or draft models generated by automatic reconstruction pipelines require comprehensive manual refinement and validation[13,14] (a time-consuming curation step that also needs substantial expertise) that involve, for example, finding missing reactions due to incomplete genomic and functional annotations (a critical step of the so-called gap-filling procedure)[15,16]. Therefore, the quality of initial draft GEMs has a profound impact on the time spent on the manual curation, the refined model quality, and ultimately its utility in biomedical applications.

A large number of gap-filling algorithms have been designed to tease out missing reactions in draft GEMs[17–20]. Despite wide differences in their input data types, objectives, and algorithms, they generally follow two steps: (1) find dead-end metabolites that cannot be produced or consumed and/or some inconsistencies between the draft model prediction and experimental data

(e.g., growth profiles); and (2) add a set of reactions to resolve the dead-end blocks and/or inconsistencies[18]. Most of these methods are driven by advances in experimental techniques to identify model-data inconsistencies[17,18]. However, experimental data is not readily available for non-model organisms (e.g., microbial species living in the human gut, many of which are not even culturable under laboratory conditions[21]), thus limiting the utility of those tools. A few exceptions, which are entirely topology-based and do not require *a priori* dataset, include (1) classical flux consistency-based approaches GapFind/GapFill[22] and FastGapFill[23]; and (2) state-of-the-art machine learning approaches Neural Hyperlink Predictor (NHP)[24] and Clique Closure-based Coordinated Matrix Minimization (C3MM)[25] (and its alternatives BoostGapFill[26] and CMM[27]). The former simply aims to restore the network connectivity based on flux consistency, while the latter exploits advanced machine learning techniques to predict missing reactions[28].

Machine learning approaches frame the prediction of missing reactions in a GEM as a supervised learning task of hyperlink prediction on a hypergraph[24,25,29], where known metabolic reactions are used to predict the presence of additional reactions based on topological features of the metabolic network in the GEM[30]. Note that metabolic networks or any biochemical reaction networks have a very natural hypergraph representation. Compared to graphs where each link connects two nodes, hypergraphs allow each hyperlink to connect more than two nodes[31–33]. In the hypergraph representation of a metabolic network, each molecular species is a node, while each reaction is a hyperlink connecting all the molecular species involving in it.

As two state-of-the-art machine learning approaches for missing reaction prediction, both

C3MM and NHP have notable limitations. C3MM has limited scalibility and cannot predict unseen reactions, since it includes all candidate metabolic reactions (obtained from a universal reaction pool) in a matrix factorization-based method during training. While the neural network-based method NHP can handle unseen reactions, it approximates hypergraphs using graphs with clique expansion (i.e., replacing each hyperlink with a clique – a complete subgraph where all the nodes belonging to the original hyperlink are connected with each other) in feature initialization, which results in the loss of higher-order information. More importantly, both methods were benchmarked against a handful of GEMs (lacking a comprehensive test) and were only internally validated using artificial gaps introduced by randomly deleting reactions from input GEMs (lacking phenotypic data validation, i.e., external validation).

Here we develop a new method called CHESHIRE (**CHE**byshev **S**pectral **H**yperl**I**nk p**RE**dictor) to overcome limitations of existing machine learning approaches. CHESHIRE takes a metabolic network and a pool of candidate reactions as the input. It outputs confidence scores for candidate reactions. For internal validation, we demonstrate that CHESHIRE significantly outperforms NHP and C3MM in a systematic test of recovering artificially removed reactions from the entire 108 high-quality GEMs in the BiGG database[34]. For external validation, we show that CHESHIRE improves the predictability of metabolic fermentation products over 24 draft GEMs reconstructed using a recent pipeline – CarveMe[11].

A brief overview of CHESHIRE

CHESHIRE is a deep learning-based algorithm that can predict missing reactions in GEMs using topological features of the metabolic networks without any inputs from experimental data. For a set of metabolic reactions (Fig. 1a) and its corresponding metabolic network (Fig. 1b), we use a hypergraph (Fig. 1c) to present its structure, where each hyperlink represents a metabolic reaction and connects participating reactant and product metabolites (*Supplementary Information* Section 1). CHESHIRE requires the incidence matrix of the metabolic network as the sole input for training. For the metabolic network of a given GEM, its incidence matrix (Fig. 1d) has boolean values indicating the presence or absence of any metabolite in any reaction and can be converted from its stoichiometric matrix by binarization of the nonzero values.

The learning architecture of CHESHIRE has four major steps: feature initialization, feature refinement, pooling, and scoring (Fig. 1e). For feature initialization, we employ an encoder-based one-layer neural network[35] to generate a feature vector for each metabolite from the incidence matrix. This initial feature vector encodes the crude information of topological relationship of a metabolite with all reactions in the metabolic network. For feature refinement, to capture the metabolite-metabolite interactions, we treat each reaction as a fully connected subgraph and subsequently use Chebyshev spectral graph convolutional network (CSGCN) to refine the feature vector of each metabolite by incorporating the features of other metabolites from the same reaction. Compared with traditional graph neural networks, CSGCN can extract local and composite metabolite features by exploiting Chebyshev expansion and spectral graph theory[36]. In

particular, it utilizes K spectral filters (resulting in K channels), mathematically represented by Kth-order polynomials of the subgraph Laplacian, to compress large-scale properties into coarse-grained variables. For pooling (i.e., integrating node or metabolite features into hyperlink- or reaction-level representation), we utilize graph coarsening methods to compute a feature vector for each reaction from the feature vectors of its metabolites. We combine two pooling functions, a maximum minimum-based function (as used in NHP[24]) and a Frobenius norm-based function[37] to provide complementary information of metabolite features. Note that the norm-based pooling is efficient at separating reactions in the reaction feature space[37] (and we choose the Frobenius norm among the others due to its low computational cost), while the maximum minimum-based pooling reflects the topological similarities among all the metabolites of a reaction[24]. Finally, for scoring, we feed the feature vector of each reaction into a one-layer neural network to produce a probabilistic score for the reaction that indicates the confidence of its existence. Compared with NHP which shares a similar architecture, CHESHIRE exploits a simple encoder[35], a sophisticated CSGCN[36] and a practical Frobenius norm-based pooling function[37] during training (*Supplementary Information* Section 2 and 3).

In the following, we perform both internal and external validations (Fig. 1f and g) to show the superiority of CHESHIRE over other methods in predicting missing reactions across a large set of microorganisms.

Internal validation of CHESHIRE using artificially introduced gaps in GEMs

We compared CHESHIRE with the state-of-the-art machine learning approaches NHP and C3MM as they have been demonstrated to display superior performances over previous algorithms such as Self Attention-based Graph Neural Networks for Hypergraphs (Hyper-SAGNN)[38], BoostGapFill[26], and CMM[27] in predicting missing reactions. We also included FastGapFill and Node2Vec-mean[24,39] (referred to as NVM below) as two baseline methods. Note that NVM has a relatively simple learning architecture that can generate metabolite and reaction features by Node2Vec (a random walk-based graph embedding method that generates node features) and mean pooling without feature refinement, respectively (*Supplementary Information* Section 2).

All the deep learning-based algorithms (i.e., CHESHIRE, NHP and NVM) require negative sampling of reactions, i.e., creating fake (or "negative") reactions during training, to balance specificity and sensitivity of the trained model. Each negative reaction was created from an existing real (or "positive") reaction in the metabolic network by keeping half of its metabolites (rounding is required if the number of metabolites is odd) and filling the other half with randomly selected metabolites from a universal metabolite pool (*Supplementary Information* Section 3). While negative reactions are not mandatory for algorithm testing, they allow richer types of classification performance metrics for model evaluation. Note that C3MM and FastGapFill do not require negative sampling of reactions during training. However, the former can handle negative reactions during testing, while the latter requires all the testing reactions to be positive.

Below we performed two sets of internal validation based on artificially introduced gaps

(i.e., removing some real reactions) from all 108 high-quality and manually-curated GEMs that constitute the BiGG database[34] (*Supplementary Information* Section 4). Note that for each GEM, we removed the biomass, exchange, demand, and sink reactions, since they do not represent knowledge gaps, and will not be considered in the gap-filling procedure. We used a threshold score of 0.5 to determine whether an unseen reaction is true or false.

First, we evaluated the performances of all the machine learning approaches (i.e., CHESHIRE, NHP, NVM, and C3MM) in predicting missing reactions via four classical classification performance metrics: the Area Under the Receiver Operating Characteristic curve (AUROC), Recall, Precision, and F1 score (the harmonic mean of Recall and Precision). The use of these metrics (except Recall) require negative sampling in testing, so we excluded FastGapFill in this set of internal validation. For all the algorithms except C3MM, we augmented the existing positive reactions in each GEM by negative reactions in a 1:1 ratio and randomly split the positive and negative mixture into 60% training (metabolic network to be gap-filled) and 40% testing (unseen candidate reactions). For fair comparison, we also introduced negative reactions (with 1:1 positive-negative ratio) to the testing set of C3MM. Tested on a total of 108 BiGG GEMs, CHESHIRE achieves the best performance in all the classification metrics mentioned above (Fig. 2a-d). Compared to the overall second best approach NHP, CHESHIRE has a significantly higher level of Precision ($P < 10^{-6}$, Wilcoxon signed rank test) and Recall ($P < 10^{-18}$), suggesting that CHESHIRE can recover the majority of true reactions without sacrificing its ability to distinguish true from fake reactions. For all the performance metrics, the wide distributions indicate that the performance of CHESHIRE (as well as other methods) is GEM-dependent, even

though they are all high-quality models. The most easily gap-filled GEM is a reconstruction from *Phaeodactylum tricornutum* CCAP 1055/1 (model iLB1027_lipid), where CHESHIRE is able to achieve AUROC=0.92, Recall=0.89, Precision=0.82, and F1 score=0.86 on the testing set. Notably, CHESHIRE is not sensitive to the threshold score, the negative sampling strategy, and the negative sampling ratio in this validation, where its performance still prevails over the other methods (*Supplementary Information* Section 5, Fig. S1-S3).

Second, to evaluate the robustness of CHESHIRE and other methods against the size of training data, we varied the percentage of reactions (between 0.2 and 0.8) randomly removed from metabolic networks. The remaining and removed reactions were respectively used as the training and testing sets. We calculated the percentage of recovered reactions over a representative subset of BiGG GEMs from four different species (Fig. 2e-h), including iAF1260b (*Escherichia coli str. K-12 substr.* MG1655), iMM1415 (*Mus musculus*), iPC815 (*Yersinia pestis* CO92), and RECON1 (*Homo sapiens*). Similar to the first set of internal validation, we augmented the positive reactions by negative reactions in a 1:1 ratio in the training set for all the methods except C3MM and FastGapFill. Since no negative reaction was added to the testing reaction set, we included FastGapFill in this set of internal validation and replaced its default universal KEGG reaction pool[40] with the testing set. We found that CHESHIRE achieves the highest recovery rate (>80%) for the four representative GEMs. More importantly, its performance is remarkably stable across different reaction removal percentages. By contrast, the other machine learning approaches (particularly C3MM) not only have much less recovery rates but their performances start to decline drastically when too many reactions are removed. Consistent with a previous

study[26], FastGapFill ranked the worst with its recovery rates varying between 0 and 10%, and its performance increases, rather than decreases, when more reactions are removed. Since FastGapFill is not a machine learning-based method, a larger training size does not necessarily translate to a better performance. In fact, the removed reactions, if not many, do not create blocked reactions and dead-end metabolites due to network redundancy, and hence are not considered as "gaps" by FastGapFill. This explains why its performance was lower when a smaller fraction of reactions was removed as artificial gaps.

Taken together, the two sets of internal validation demonstrate that CHESHIRE outperforms other topology-based methods and hence is more promising for predicting missing reactions in draft GEMs.

External validation of CHESHIRE via phenotypic prediction

Compared to internal validation that tests the predictions by using artificially removed reactions as the ground truth, external validation tests the predicted missing reactions from the current GEMs, which is more challenging due to the lack of such ground truth data. Despite the difficulty to validate each individual reaction predicted to be missing, we can validate them as a group by testing whether adding these reactions together to the GEM can improve the GEM's power in predicting metabolic phenotypes such as fermentation. This test is biologically meaningful. After all, a major rationale for reconstructing GEMs of microorganisms is to provide theoretical predictions of their metabolic phenotypes[41].

To test whether our method can improve GEMs' power of phenotypic prediction, we designed a workflow that uses CHESHIRE to gap-fill draft GEMs (Fig. 3a, *Supplementary Information* Section 6). Unlike conventional gap-filling methods which need fermentation data to identify gaps, our approach is completely unsupervised without seeing any data during the process of gap-filling. Briefly, CHESHIRE is trained on the entire reaction set of a draft GEM and ranks the candidate reactions (taken from a reaction pool, e.g., the BiGG database[34]) by their likelihoods of being present in the draft GEM. The top reactions in the ranking are added to the draft GEM to produce a gap-filled GEM. Using parsimonious flux balance analysis[42] and flux variability analysis[43], we identify inconsistent phenotypic predictions made between the draft GEM and the gap-filled GEM and the inconsistency suggests potentially missing reactions in the draft GEM. Particularly, for fermentation phenotypes predicted by the gap-filled GEM but not by the draft GEM, the workflow uses Linear Mixed-Integer Programming to infer the reactions that causally result in the hypothetical gaps. The model quality of the gap-filled GEM in predicting fermentation can be evaluated by comparison with experimental data.

We applied the workflow to a compiled dataset including fermentation profiles of 9 metabolites from 24 bacterial organisms (*Supplementary Information* Section 4, Table S1) grown under anaerobic conditions[12]. The draft GEMs of those organisms were reconstructed using a recent automatic reconstruction pipeline CarveMe[11]. Given that the BiGG universal reaction pool has more than 10,000 reactions (*Supplementary Information* Section 4), a very stringent threshold of 0.99999 may still predict thousands of reactions as missing reactions, depending on the draft GEMs (Fig. S4). This is expected because the number of reactions with similar biochemistry

mechanisms and nearly identical confidence scores scale up with the size of candidate reaction pool. Instead of using a fixed cutoff score, we therefore gap-fill GEMs by adding the top 200 or 500 reactions with the highest confidence scores. For any reaction causing energy-generating cycles (EGCs)[44], the reaction was included if EGCs can be eliminated by changing its flux bounds and otherwise skipped (*Supplementary Information* Section 6).

As mentioned above, the external validation evaluates model performance by formalizing a classification or prediction problem of fermentation profiles. This is different from the internal validation which classifies presence and absence of reactions directly. Using the same set of classification performance metrics (AUROC, Recall, Precision, and F1 score), we compared four different groups of models: the original GEMs reconstructed from CarveMe (CarveMe), gap-filled GEMs by adding top 200 or 500 reactions predicted by CHESHIRE and NHP (CHESHIRE-200/500 and NHP-200/500), and gap-filled GEMs by randomly adding 500 reactions from the universal BiGG reaction pool (Random-500). FastGapFill and NVM were not included here due to their poor performances in internal validation. C3MM was not considered either, because it requires infeasible computational time for such a large candidate reaction pool (with 17,298 reactions).

We first observed a high variability across the draft CarveMe models to predict fermentation profiles (Fig. 3b-i, gray dots): a few models correctly predicted all phenotypes but some others failed to predict any. Second, adding 200 or 500 NHP-predicted reactions barely improves the phenotypic predictions and the improvement is even worse than that after randomly adding 500

reactions (Fig. 3b-e). To the contrary, by adding 200 or 500 CHESHIRE-predicted reactions in the CarveMe draft models, the mean performances significantly increase (Fig. 3f-i, Fig. S5). We further demonstrated that the improved performance is not simply due to more reactions by showing significantly better performances between CHESHIRE-500 and Random-500. The randomly added 500 reactions did improve the phenotypic predictions, but the level of improvement is similar to that after adding the top 200 CHESHIRE-predicted reactions.

To understand how CHESHIRE gap-fills the GEMs, we compared the model predictions between CHESHIRE-500 and CarveMe on a per GEM and per metabolite basis. CHESHIRE-500 improves the F1 score for 19 of the 24 draft GEMs. Among the rest five GEMs, CHESHIRE-500 achieved zero F1 scores (Fig. 3i) on two GEMs (*Anaerobutyricum hallii* DSM 3353 and *Zymomonas mobilis subsp. mobilis* ATCC 10988) where each produces a single metabolite but CHESHIRE failed to gap-fill. Compared across the 9 fermentation metabolites, the biggest improvement of CHESHIRE over CarveMe draft models was observed on acetic acid followed by lactic acid. CHESHIRE increases the correct predictions of acetic acid phenotype from 8 to 22 of the 24 draft GEMs. The gap-filling guided by CHESHIRE corrected 15 false-negative predictions by adding a single acetic acid transport reaction between intracellular and extracellular or periplasm space via proton symporter (e.g., *Prevotella bergensis* DSM 17361, see Fig. 3j). This is not surprising as the poor performance of draft GEMs is mostly due to poor annotation of transporter genes[45–47]. Notably, the acetic acid transport reaction is not unique. We found that the citrate/sodium symporter, which does not involve acetic acid molecule, can rescue the phenotype of acetic acid production in the GEM of *Enterococcus faecalis* ATCC 19433 by coupling with a

citrate/acetate antiporter identified by the CarveMe pipeline. For another short-chain fatty acid–butyric acid, CHESHIRE corrected two false-negatives for *Eubacterium rectale* ATCC 33656 and *Clostridium butyricum* KNU-L09 by adding a reaction from butyryl-CoA to butyrate (Fig. 3k).

CHESHIRE can also identify false-positive predictions. For example, the draft GEM of *Anaerobutyricum hallii* has positive maximum flux of lactate production at fastest growth, while experimental data shows no production. By contrast, lactate consumption, rather than production, is preferred by the gap-filled GEM which incorporates 500 CHESHIRE-predicted reactions because these reactions facilitate lactate utilization and increase the maximum growth rate. We further found that the efficient use of lactate for faster growth of the gap-filled GEM is enabled by two NAD(P)H-mediated redox reactions that do not directly facilitate lactate uptake (Fig. 3l). Since lactate utilization reduces NAD^+ to NADH (oxidized and reduced forms of nicotinamide adenine dinucleotide) and both reactions involve NAD^+ or $NADP^+$ (phorsphorylated form of NAD^+), we hypothesize that the two reactions overcome the redox imbalance in the lactate metabolism of the draft GEM. This example shows that CHESHIRE can identify missing reactions that have consequences on distant fermentation pathways via a global, systematic effect.

Discussion

GEM gap-filling has been long considered as a process of fitting a GEM to observed data[48]. This problem is typically formulated by a mixed-integer linear programming that minimizes the number of added reactions under the constraint that the observed phenotypes are satisfied. Therefore, the

majority of GEM gap-filling algorithms falls short of predicting metabolic gaps in both network connections and functions without knowing experimental phenotypes *a priori*. FastGapFill, as one of a few exceptions, fits a specific task of gap-filling to resolve dead-ends and blocked reactions[23]. Both our (Fig. 2e-h) and previous studies[26,49,50] have shown that FastGapFill exhibits a poor performance in filling artificially introduced gaps. Although gap-filling with experimental data is critically important, it is limited to understanding the gene-reaction-phenotype mappings in conditions where the data was measured. The environmental conditions are combinatorially complex; as a theoretical tool, the primary purpose of GEMs is to rapidly offer theoretical predictions of metabolic activities over a large array of environmental conditions where data has not been measured.

We therefore present a new approach, CHESHIRE, which uses deep learning techniques to predict missing reactions in GEMs solely based on network topology. It is completely unsupervised without using any phenotypical data. The performance of CHESHIRE has been rigorously examined through both internal and external validations over GEMs of a large set of microorganisms. Compared to previous gap-filling algorithms, CHESHIRE adopts the concept of hypergraphs with advanced graph convolutional networks to accurately learn the geometrical patterns of metabolic networks and predict missing metabolic reactions without inputs from any experimental data. To this end, CHESHIRE meets the demand by efficiently curating GEMs to improve *in silico* predictions of missing metabolic reactions and functions, which can be readily tested by *in vitro* experiments. Despite CHESHIRE is purely topology-based, it has a flexible architecture that can potentially incorporate metabolomics data to further boost the gap-filling

performance. For example, CHESHIRE can be trained on the incidence matrix of metabolite correlation-based networks alone or jointly with reaction-based GEMs to predict novel, yet unidentified, metabolic pathways[51].

A highlight of our study is that we benchmarked our approach using fermentation data in addition to artificially introduced gaps. To the best of our knowledge, this benchmark has not been performed for previous unsupervised topology-based gap-filling algorithms. The task of improving metabolic phenotype prediction is generally more challenging than correctly identifying missing metabolic reactions: metabolic networks are highly connected and a missing phenotype may be due to the absence of multiple reactions. Despite the challenge, we show that CHESHIRE improves the prediction of fermentation products over 24 bacterial organisms based on draft GEMs reconstructed from one of the three mostly used automatic pipelines, i.e., CarveMe[11]. We selected CarveMe because: (1) the generated draft GEMs have higher quality than those generated by ModelSeed[52]; and (2) Gapseq[12], the most recent pipeline, has an innate gap-filling tool using a filtered pool of candidate reactions supported by genetic evidences. Although a fair comparison between CHESHIRE and Gapseq is very difficult to establish, Gapseq points out a promising direction to further improve gap-filling-based GEM curation by integrating the existing automatic reconstruction pipelines with the standalone deep learning-based gap-filling tools. The integration would allow substantial reduction of candidate reactions for gap-filling from a reaction universe to a small subset of those supported by homology-based search. Indeed, 11% (>1,000) reactions from the BiGG universal reaction pool have confidence scores at least 99.9% on average across all 24 GEMs (Fig. S4). In the future, we foresee the integration of CHESHIRE and other metabolic

network gap-filling tools to fully automate the gap-filling step of GEM refinement.

Author Contributions Y.-Y.L. conceived and designed the project. C.C. developed the CHESHIRE algorithm and performed internal validations. C.L. performed the external validation and interpreted the results. C.C. and C.L. prepared the manuscript. Y.-Y.L. edited and approved the manuscript.

Acknowledgements Y.-Y.L. acknowledges grants from the National Institutes of Health (R01AI141529, R01HD093761, RF1AG067744, UH3OD023268, U19AI095219, and U01HL089856).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Dr. Yang-Yu Liu (email: yyl@channing.harvard.edu).

Reference

1. Hao Wang, Jonathan L Robinson, Pinar Kocabas, Johan Gustafsson, Mihail Anton, Pierre-Etienne Cholley, Shan Huang, Johan Gobom, Thomas Svensson, Mattias Uhlen, et al. Genome-scale metabolic network reconstruction of model animals as a platform for translational research. *Proceedings of the National Academy of Sciences*, 118(30), 2021.

2. Xin Fang, Colton J Lloyd, and Bernhard O Palsson. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nature Reviews Microbiology*, 18(12):731–743, 2020.

3. Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

4. Changdai Gu, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. Current status and applications of genome-scale metabolic models. *Genome Biology*, 20(1):1–18, 2019.

5. Almut Heinken, Arianna Basile, Johannes Hertel, Cyrille Thinnes, and Ines Thiele. Genome-scale metabolic modeling of the human microbiome in the era of personalized medicine. *Annual Review of Microbiology*, 75:199–222, 2021.

6. Steinn Gudmundsson and Juan Nogales. Recent advances in model-assisted metabolic engineering. *Current Opinion in Systems Biology*, 28:100392, 2021.

7. Dongsoo Yang, Seon Young Park, Yae Seul Park, Hyunmin Eun, and Sang Yup Lee. Metabolic engineering of escherichia coli for natural product biosynthesis. *Trends in Biotechnology*, 38(7):745–765, 2020.

8. Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, 2017.

9. Jonathan L Robinson and Jens Nielsen. Anticancer drug discovery through genome-scale metabolic modeling. *Current Opinion in Systems Biology*, 4:1–8, 2017.

10. Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of earth's microbiomes. *Nature Biotechnology*, 39(4):499–509, 2021.

11. Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, 2018.

12. Johannes Zimmermann, Christoph Kaleta, and Silvio Waschina. gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology*, 22(1):1–35, 2021.

13. Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010.

14. Charles J Norsigian, Xin Fang, Yara Seif, Jonathan M Monk, and Bernhard O Palsson. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. *Nature Protocols*, 15(1):1–14, 2020.

15. Jennifer L Reed, Trina R Patel, Keri H Chen, Andrew R Joyce, Margaret K Applebee, Christopher D Herring, Olivia T Bui, Eric M Knight, Stephen S Fong, and Bernhard O Palsson. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences*, 103(46):17480–17484, 2006.

16. Ottar Rolfsson, Bernhard Ø Palsson, and Ines Thiele. The human metabolic reconstruction recon 1 directs hypotheses of novel human metabolic functions. *BMC Systems Biology*, 5(1):1–16, 2011.

17. Jeffrey D Orth and Bernhard Ø Palsson. Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*, 107(3):403–412, 2010.

18. Shu Pan and Jennifer L Reed. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Current Opinion in Biotechnology*, 51:103–108, 2018.

19. Pratip Rana, Carter Berry, Preetam Ghosh, and Stephen S Fong. Recent advances on constraint-based models by integrating machine learning. *Current Opinion in Biotechnology*, 64:85–91, 2020.

20. David B Bernstein, Floyd E Dewhirst, and Daniel Segre. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. *Elife*, 8:e39733, 2019.

21. Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 2019.

22. Vinay Satish Kumar, Madhukar S Dasika, and Costas D Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):1–16, 2007.

23. Ines Thiele, Nikos Vlassis, and Ronan MT Fleming. Fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17):2529–2531, 2014.

24. Naganand Yadati, Vikram Nitin, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Talukdar. Nhp: Neural hypergraph link prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1705–1714, 2020.

25. Govind Sharma, Prasanna Patil, and M Narasimha Murty. C3mm: clique-closure based hyperlink prediction. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pages 3364–3370, 2020.

26. Tolutola Oyetunde, Muhan Zhang, Yixin Chen, Yinjie Tang, and Cynthia Lo. Boostgapfill: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods. *Bioinformatics*, 33(4):608–611, 2017.

27. Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. Beyond link prediction: Predicting hyperlinks in adjacency space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4430–4437, 2018.

28. Yeji Kim, Gi Bae Kim, and Sang Yup Lee. Machine learning applications in genome-scale metabolic modeling. *Current Opinion in Systems Biology*, 25:42–49, 2021.

29. Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009.

30. Diogo M Camacho, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.

31. Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

32. Can Chen and Indika Rajapakse. Tensor entropy for uniform hypergraphs. *IEEE Transactions on Network Science and Engineering*, 7(4):2889–2900, 2020.

33. Can Chen, Amit Surana, Anthony Bloch, and Indika Rajapakse. Controllability of hypergraphs. *IEEE Transactions on Network Science and Engineering*, 8(2):1646–1657, 2021.

34. Charles J Norsigian, Neha Pusarla, John Luke McConn, James T Yurkovich, Andreas Dräger, Bernhard O Palsson, and Zachary King. Bigg models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Research*, 48(D1):D402–D406, 2020.

35. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

36. Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29:3844–3852, 2016.

37. Caglar Gulcehre, Kyunghyun Cho, Razvan Pascanu, and Yoshua Bengio. Learned-norm pooling for deep feedforward and recurrent neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 530–546. Springer, 2014.

38. Ruochi Zhang, Yuesong Zou, and Jian Ma. Hyper-sagnn: a self-attention based graph neural network for hypergraphs. In *The International Conference on Learning Representations*, 2020.

39. Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

40. Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

41. David B Bernstein, Snorre Sulheim, Eivind Almaas, and Daniel Segrè. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology*, 22(1):1–22, 2021.

42. Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(1):390, 2010.

43. I Thiele and S Gudmundsson. Computationally efficient flux variability analysis. *BMC Bioninformatics*, 11(489):1–3, 2010.

44. Claus Jonathan Fritzemeier, Daniel Hartleb, Balázs Szappanos, Balázs Papp, and Martin J Lercher. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Computational Biology*, 13(4):e1005494, 2017.

45. Nachon Raethong, Jirasak Wong-Ekkabut, Kobkul Laoteng, and Wanwipa Vongsangnak. Sequence-and structure-based functional annotation and assessment of metabolic transporters in *Aspergillus oryzae*: a representative case study. *BioMed Research International*, 2016.

46. Marie Schöpping, Paula Gaspar, Ana Rute Neves, Carl Johan Franzén, and Ahmad A Zeidan. Identifying the essential nutritional requirements of the probiotic bacteria *Bifidobacterium animalis* and *Bifidobacterium longum* through genome-scale modeling. *NPJ Systems Biology and Applications*, 7(1):1–15, 2021.

47. Elena Vinay-Lara, Joshua J Hamilton, Buffy Stahl, Jeff R Broadbent, Jennifer L Reed, and James L Steele. Genome–scale reconstruction of metabolic networks of *Lactobacillus casei* atcc 334 and 12a. *PloS One*, 9(11):e110785, 2014.

48. Gregory L Medlock and Jason A Papin. Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. *Cell Systems*, 10(1):109–119, 2020.

49. Sylvain Prigent, Clémence Frioux, Simon M Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, et al. Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. *PLoS Computational Biology*, 13(1):e1005276, 2017.

50. Edward Vitkin and Tomer Shlomi. Mirage: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biology*, 13(11):1–11, 2012.

51. David Toubiana, Rami Puzis, Lingling Wen, Noga Sikron, Assylay Kurmanbayeva, Aigerim Soltabayeva, Maria del Mar Rubio Wilhelmi, Nir Sade, Aaron Fait, Moshe Sagi, et al.

Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Communications Biology*, 2(1):1–13, 2019.

52. Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9):977–982, 2010.

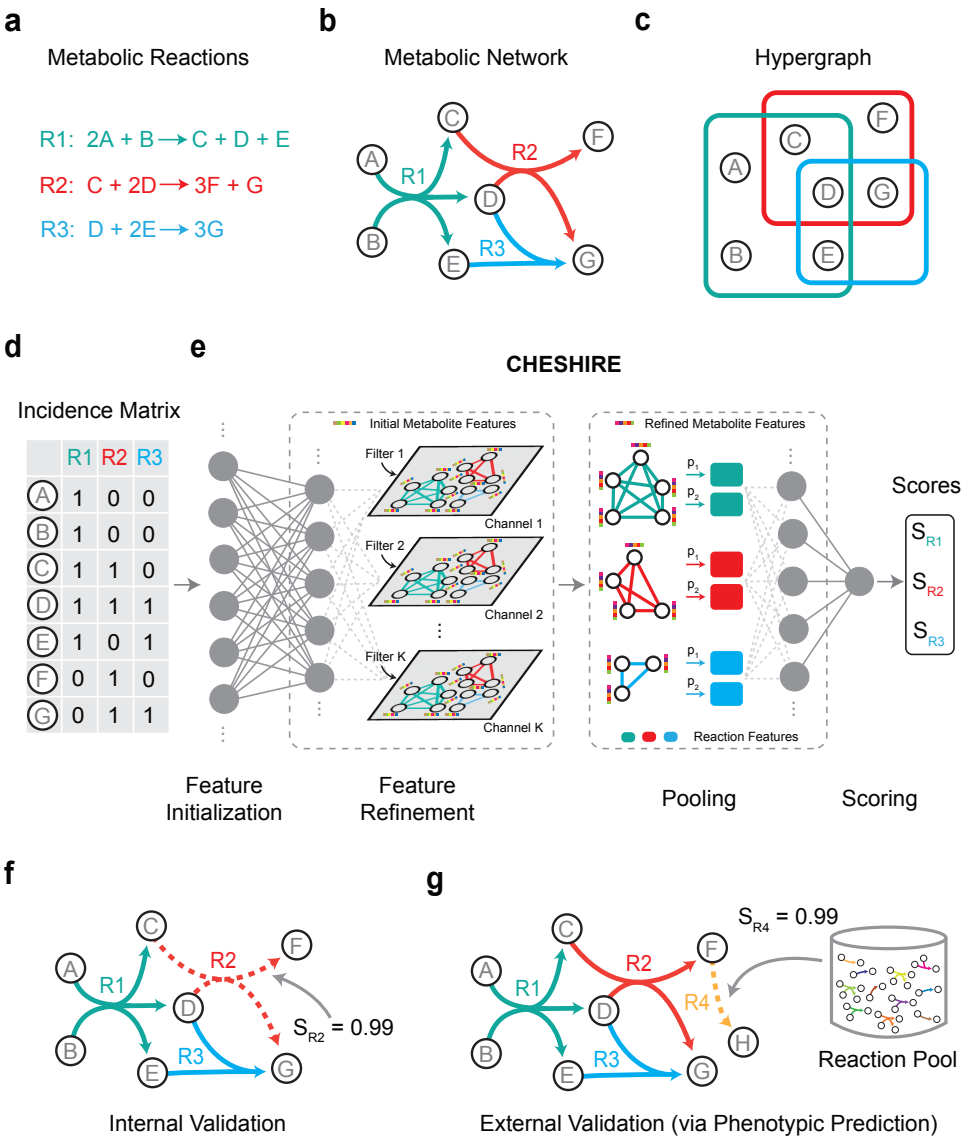
a
Metabolic Reactions
R1: 2A + B → C + D + E
R2: C + 2D → 3F + G
R3: D + 2E → 3G
b
Metabolic Network
c
Hypergraph
d
Incidence Matrix
R1 R2 R3
A 1 0 0
B 1 0 0
C 1 1 0
D 1 1 1
E 1 0 1
F 0 1 0
G 0 1 1
e
CHESHIRE
Initial Metabolite Features
Filter 1
Channel 1
Filter 2
Channel 2
Filter K
Channel K
Refined Metabolite Features
Reaction Features
Scores
S R1
S R2
S R3
Feature Initialization
Feature Refinement
Pooling
Scoring
f
S R2 = 0.99
Internal Validation
g
S R4 = 0.99
Reaction Pool
External Validation (via Phenotypic Prediction)

Fig. 1: Brief summary of CHESHIRE workflow. **a.** Examples of metabolic reactions. **b.** Schematic representation of a metabolic network. **c.** Hypergraph representation of the metabolic network. The hypergraph is undirected where each hyperlink connects metabolites that participate the same reaction. **d.** Incidence matrix of the hypergraph shown in (a). Each value of the incidence matrix is a boolean variable indicating whether the metabolite on the row participates in the reaction on the column. **e.** Deep learning architecture of CHESHIRE during training. The deep neural network consists of an encoder layer that takes the incidence matrix as the only input, a Chebyshev spectral graph convolutional layer with K filters (resulting in K channels), a pooling layer with two pooling functions, and a final scoring layer. The grey disks represent the hidden neurons, and the output confidence scores show the probabilities of given reactions that have been missed by the input metabolic network. **f.** Schematic illustration of internal validation. We tested model performance by predicting the gaps artificially created in GEMs. **g.** Schematic illustration of external validation. We identified gaps of intact draft GEMs by comparing model predictions with phenotypic data.

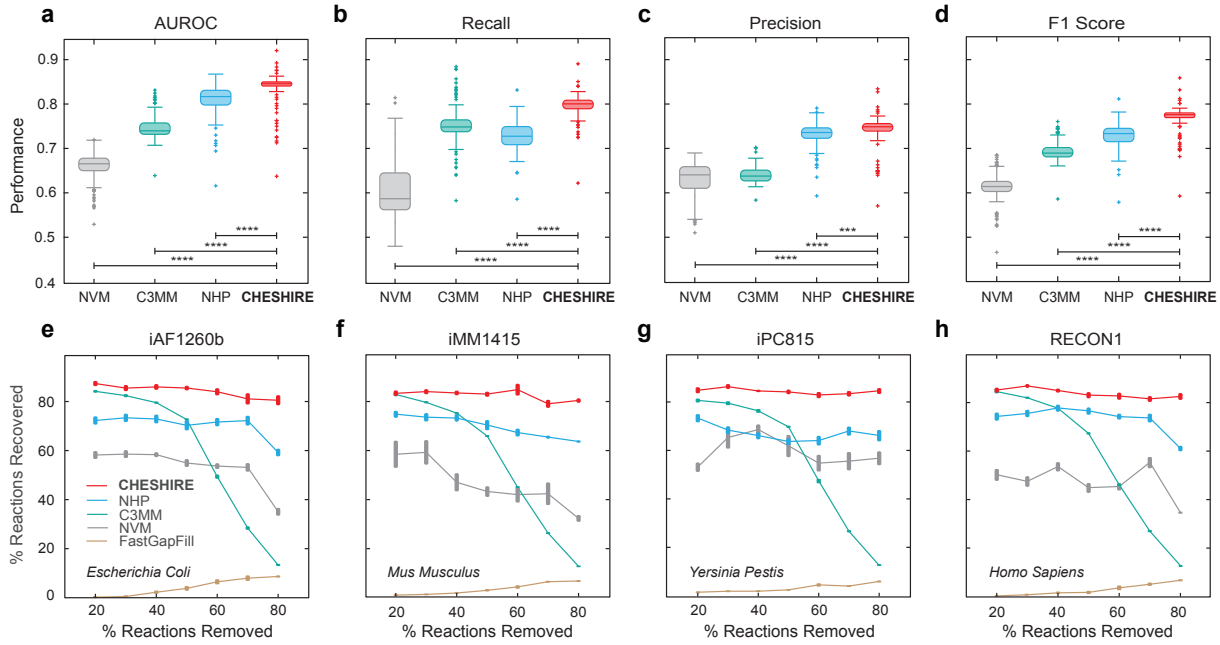

Fig. 2: Internal validation using artificially introduced gaps. **a-d.** Boxplots of the performance metrics (AUROC, Recall, Precision, and F1 score) calculated on 108 BiGG GEMs (each dot represents a GEM) for CHESHIRE vs. NHP, C3MM, and NVM. **e-h.** Reaction recovery rate of CHESHIRE vs. NHP, C3MM, NVM, and FastGapFill for gap-filling the four selected BiGG GEMs. For all the panels, reactions were removed at random from the GEMs and treated as unseen reactions in the testing set. Each data point represents the mean statistic over 10 Monte Carlo runs. Error bars: standard error of the mean. Wilcoxon signed rank tests: $***P < 10^{-6}$; $****P < 10^{-18}$.

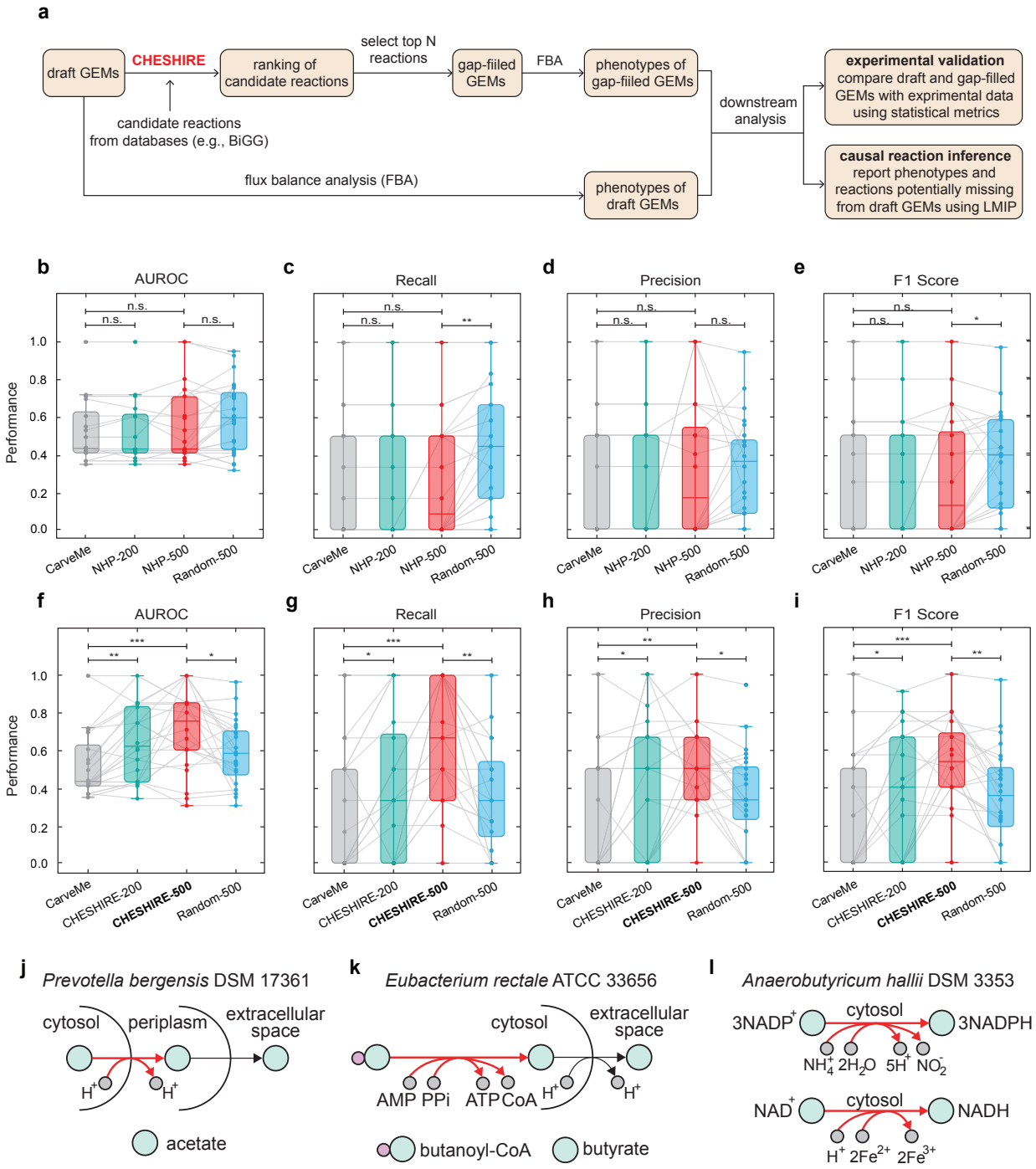
a
draft GEMs
CHESHIRE
ranking of candidate reactions
select top N reactions
gap-fiiled GEMs
FBA
phenotypes of gap-fiiled GEMs
candidate reactions from databases (e.g., BiGG)
flux balance analysis (FBA)
phenotypes of draft GEMs
downstream analysis
experimental validation
compare draft and gap-filled GEMs with exprimental data using statistical metrics
causal reaction inference
report phenotypes and reactions potentially missing from draft GEMs using LMIP
b AUROC
c Recall
d Precision
e F1 Score
f AUROC
g Recall
h Precision
i F1 Score
Performance
CarveMe
NHP-200
NHP-500
Random-500
CHESHIRE-200
CHESHIRE-500
n.s.
j Prevotella bergensis DSM 17361
cytosol
periplasm
extracellular space
acetate
k Eubacterium rectale ATCC 33656
AMP PPi ATP CoA
butanoyl-CoA
butyrate
l Anaerobutyricum hallii DSM 3353
3NADP+
3NADPH
NAD+
NADH

Fig. 3: External validation by predicting fermentation product profiles. **a**. Flowchart of external validation. The predicted fermentation profiles from CHESHIRE-gapfilled GEMs are validated by comparison to experimental observation. For fermentation phenotypes correctly predicted by gap-filled GEMs but missed by draft GEMs, we also identify the causal reactions from CHESHIRE-predicted set that improve the phenotypic prediction using Linear Mixed-Integer Programming (LMIP). **b-i.** Model performance (AUROC, Recall, Precision, and F1 score) tested on 24 GEMs (each dot represents a GEM) gap-filled by NHP (b-e) and CHESHIRE (f-i). "CarveMe" represents the draft models reconstructed from CarveMe. "NHP-200/CHESHIRE-200" and "NHP-500/CHESHIRE-500" represent draft models plus 200 and 500 reactions predicted by NHP/CHESHIRE, respectively. For "Random-500", 500 reactions randomly selected from the universal reaction pool were added to the draft models and each dot averaged over three Monte Carlo runs. Boxplot: central line represents the median, box limits represent the first and third quartiles, and whiskers extend to the smallest and largest values or at most to 1.5× the interquartile range, whichever is smaller. Wilcoxon signed rank tests: n.s., not significant; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. **j-l.** Examples of CHESHIRE-predicted reactions (red arrows) that causally gap-fill the observed phenotypes of acetate (j), butyrate (k) and lactate (l) fermentation. Note that the two reactions in panel-l are not involved in the lactate metabolism but have consequences on the lactate fermentation via a systemic effect of redox coupling. Black arrows represent reactions in the draft GEMs and gray circles represent cofactors. Abbreviations of cofactors: adenosine triphosphate (ATP); adenosine phosphate (AMP); inorganic pyrophosphate (PPi); Coenzyme A (CoA); oxidized/reduced nicotinamide adenine dinucleotide (NAD^+/NADH); oxidized/reduced nicotinamide adenine dinucleotide phosphate ($NADP^+$/NADPH).