Teasing out Missing Reactions in Genome-scale Metabolic Networks through Graph Convolutional Networks

Can Chen[1,†], Chen Liao[2,†] & Yang-Yu Liu[1,3,*]

[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*

[3]*Center for Artificial Intelligence and Modeling, The Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61801, 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., missing reactions). Existing gap-filling methods typically require phenotypic data as input to tease out missing reactions. We still lack a computational method for rapid and accurate gap-filling of metabolic networks before experimental data is available. Here we present a deep learning-based method — CHEbyshev Spectral HyperlInk pREdictor (CHESHIRE) — to predict missing reactions in GEMs purely from metabolic network topology. We demonstrate that

CHESHIRE outperforms other topology-based methods in predicting artificially removed reactions over 926 high- and intermediate-quality GEMs. Furthermore, CHESHIRE is able to improve the phenotypic predictions of 49 draft GEMs for fermentation products and amino acids secretions. Both types of validation 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 generated by automatic reconstruction pipelines[11,12]. Yet, these draft models contain knowledge gaps and thus require comprehensive manual curation [13,14], e.g., finding missing reactions due to incomplete genomic and functional annotations. 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.

Numerous optimization-based gap-filling methods have been designed to tease out missing reactions in draft GEMs[15–18]. 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[16]. Optimization-based methods often require data as input to identify model simulation-data inconsistencies[15,16]. However, experimental data is not readily available for non-model organisms, thus limiting the utility of those tools. For example, most intestinal organisms are considered "uncultivable" and their functions remain unknown[19]. Even for cultivable organisms, high-throughput phenotypic screening, i.e., searching for organisms with desired phenotypes, relies on the analysis of microbial extracts or genetic modifications, which can become complicated, time-consuming, and expensive. Given the increasing availability of cultivable organisms and their genomes, there is a pressing need for rapid and accurate *in silico* predictions of metabolic phenotypes solely from genomic sequences. Even though the predictions are theoretical, downstream experimental validations could be mush less resource-demanding.

A few gap-filling methods that are entirely topology-based and do not require phenotypic data as input include (1) classical methods that restore the network connectivity based on flux consistency, such as GapFind/GapFill[20] and FastGapFill[21]; and (2) state-of-the-art machine learning methods that exploit advanced machine learning techniques, such as Neural Hyperlink Predictor (NHP)[22] and Clique Closure-based Coordinated Matrix Minimization (C3MM)[23] (and its alternatives BoostGapFill[24] and CMM[25]). Machine learning methods frame the prediction of missing reactions in a GEM as a task of predicting hyperlinks on a hypergraph[22,23,26]. Compared to graphs where each link connects two nodes, hypergraphs allow each hyperlink to connect more than two nodes[27–29]. Notably, metabolic networks or any biochemical reaction networks have a very natural hypergraph representation: each molecular species is a node and each reaction is a hyperlink connecting all the molecular species involved in it.

There are notable limitations associated with existing topology-based machine learning methods. C3MM has an integrated training-prediction process, which includes all candidate reactions (obtained from a reaction pool) during training. Hence, it has limited scalability (i.e., it cannot handle large reaction pools), and the model has to be re-trained for each new reaction pool. While the neural network-based method NHP separates candidate reactions from training, it approximates hypergraphs using graphs in generating node features, 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 validations on predicting metabolic phenotypes, 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 the limitations of existing machine learning methods. CHESHIRE only requires a metabolic network for training and outputs confidence scores for candidate reactions from a reaction pool. For internal validation, we showed that CHESHIRE outperforms NHP and C3MM in systematic tests of recovering artificially removed reactions from 108 BiGG[30] models and 818 AGORA[8] models. For external validation, we assessed the ability of CHESHIRE to predict metabolic phenotypes. Using 49 draft GEMs reconstructed from commonly used pipelines (CarveMe[11] and ModelSEED[31]), we showed that CHESHIRE improves the theoretical predictions of whether fermentation metabolites and amino acids are produced by these GEMs.

A brief overview of CHESHIRE

CHESHIRE is a deep learning-based method that predicts missing reactions in GEMs using topological features of their metabolic networks without any inputs from experimental data (other than genomic sequences). For each metabolic network (Fig. 1a), we use a hypergraph (Fig. 1b) to model its structure, where each hyperlink represents a metabolic reaction and connects participating reactant and product metabolites (Supplementary Information Section 1.1). CHESHIRE takes the incidence matrix of the hypergraph and a decomposed graph (built from the hypergraph of existing or candidate reactions) as input. The former contains boolean values indicating the presence or absence of each metabolite in each reaction. The latter consists of fully connected subgraphs (each subgraph represents a reaction with all its metabolites connected) formed by positive and negative reactions during training and by candidate reactions during prediction (Fig. 1c and d). Positive reactions are those existing in the metabolic network, while negative reactions are fake (do not exist) and created for model-balancing purposes (often referred to as negative sampling). Note that only positive reactions are used to construct the incidence matrix.

The learning architecture of CHESHIRE has four major steps: feature initialization, feature refinement, pooling, and scoring (Fig. 1e and f). For feature initialization, we employ an encoder-based one-layer neural network[32] 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 use Chebyshev spectral graph convolutional network (CSGCN)[33] on the decomposed graph to refine the feature vector of each metabolite by incorporating the features of other metabolites from the same reaction. For pooling (i.e., integrating node- or metabolite-level features into hyperlink- or reaction-level representation), we utilize graph coarsening methods to compute a feature vector for each reaction (represented by a fully connected subgraph in the decomposed graph) from the feature vectors of its metabolites. We combine two pooling functions, a maximum minimum-based function (as used in NHP[22]) and a Frobenius norm-based function[34] to provide complementary information of metabolite features. 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. In the training phase, the resulting scores are compared to the target scores (one for positive reactions and zero for negative reactions) with a loss function for updating the model parameters (Fig. 1e, Supplementary Information Section 3.3). Compared with NHP which shares a similar architecture, CHESHIRE exploits a simple encoder[32], a sophisticated CSGCN[33], and a practical Frobenius norm-based pooling function[34] (Supplementary Information Section 2.3 and 3.1).

Internal validation of CHESHIRE using artificially introduced gaps

The goal of internal validation is to test the ability of CHESHIRE to recover artificially introduced gaps (i.e., removing existing reactions in metabolic networks). We compared CHESHIRE with the state-of-the-art machine learning methods NHP and C3MM as they have been demonstrated to display superior performances over previous gap-filling (or hyperlink prediction) methods.

We also included Node2Vec-mean[22,35] (referred to as NVM below) as a baseline method. The learning architecture of NVM is relatively simple. It uses Node2Vec (a random walk-based graph embedding method that generates node features) and mean pooling (averaging node features) to generate metabolite and reactions features, respectively, without feature refinement (Supplementary Information Section 2.3).

Below we performed two types of internal validation based on artificially introduced gaps (Fig. 2a). For both types, metabolic reactions in a given GEM were first split into a training set and a testing set over 10 Monte Carlo runs. All the deep learning-based methods (CHESHIRE, NHP, and NVM) require negative sampling of reactions. We created negative reactions at 1:1 ratio to positive reactions for the training and testing sets, respectively, by replacing half (rounded if needed) of the metabolites in each positive reaction with randomly selected metabolites from a universal metabolite pool (Fig. 2a, Supplementary Information Section 3.2). In the first type of internal validation, the training and testing sets of positive reactions and their derived negative reactions were combined and used for training and testing, respectively. For fair comparison, we also introduced negative reactions to the testing set of C3MM. In the second type of internal validation, every step remains the same except that the testing set was not mixed with its derived negative reactions but with real reactions from a universal database.

To perform the first type of internal validation, we tested CHESHIRE on a total of 108 high-quality BiGG GEMs (with 60% training and 40% testing). CHESHIRE achieves the best performance in different classification performance metrics, including the Area Under the Receiver

Operating Characteristic curve (AUROC), Recall, Precision, and F1 score (the harmonic mean of Recall and Precision) (Fig. 2b-e). A threshold score of 0.5 was used to determine whether a test reaction is true or false. Compared to the overall second-best method NHP, CHESHIRE has a significantly higher level of Precision ($P < 10^{-4}$, two-sided paired-sample t-test) and Recall ($P < 10^{-16}$), suggesting that CHESHIRE recovers 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 the other three 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 (Fig. S1-S3, Supplementary Information Section 5.2). We also performed the same type of internal validation on a larger set of metabolic networks, i.e., 818 AGORA models from the Virtual Metabolic Human database[8], and observed similar results (Fig. S4).

To perform the second type of internal validation, we tested CHESHIRE on the same BiGG GEMs (with 90% training and 10% testing) with genus-specific and the universal BiGG reaction pools (Supplementary Information Section 4.2 and 4.3). The former has a relatively small size (200-800 reactions) per GEM, while the latter has almost 17,000 reactions. Since the number of reactions with similar biochemistry mechanisms and thus nearly identical confidence scores scale up with the size of candidate reaction pool, a loose threshold of 0.5 may still predict hundreds or

thousands of candidates as missing reactions. Instead of using a fixed cutoff threshold, we added the top 25, 50, 100, and N reactions with the highest confidence scores (N is the number of artificially removed reactions). We found that CHESHIRE achieves the highest recovery rate at the four cutoffs for both types of reaction pools (Fig. 2f-m). Notably, by adding the top 25 reactions from the genus-specific reaction pools, CHESHIRE identifies more than 40% (on average) artificially removed reactions, significantly outperforming the other three methods (Fig. 2f; $P < 10^{-16}$, two-sided paired-sample t-test). While the performance of CHESHIRE declines when more reactions were added, it is still significantly better than the second-best method C3MM at the top N cutoff (Fig. 2i; $P < 0.05$). Furthermore, as expected, using the entire BiGG universal reaction pool would undermine the performances of recovery rate for all the methods. CHESHIRE nevertheless accomplishes the best performance compared to NHP and NVM (Fig. 2j-m; $P < 10^{-16}$).

Taken together, the two types of internal validations 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 whether gap-filled GEMs by CHESHIRE has improved performance compared to draft GEMs in terms of their predictions of phenotypic data (Fig. 3a, Supplementary Information Section 6). This test is biologically meaningful. After all, a major

rationale for reconstructing GEMs of microorganisms is to provide theoretical predictions of their metabolic phenotypes[36].

Briefly, CHESHIRE is trained on the entire reaction set of a draft GEM, and candidate reactions (taken from a reaction pool, e.g., the BiGG database[30]) are ranked based on both confidence and similarity scores. The confidence score, returned by CHESHIRE, quantifies the probability of a candidate reaction being present in the GEM. The similarity score measures the maximum correlation between a candidate reaction and all existing reactions in the GEM. Given our rationale that dissimilar reactions are more likely to be functionally complementary to the existing ones, all candidate reactions whose confidence scores ≥ 0.9995 are ranked by their similarity scores (least similar to most similar). The top 200 reactions are added to the draft GEM to produce a gap-filled GEM (Supplementary Information Section 6.1). Particularly, any reaction causing energy-generating cycles (EGCs)[37] is included if EGCs can be eliminated by changing its flux bounds and otherwise skipped. Given culture medium (Supplementary Information Section 6.2), the simulated phenotypes of both GEMs (Supplementary Information Section 6.3) are then compared to experimental observation for validation.

We applied the workflow to a compiled dataset including fermentation profiles of 9 metabolites from 24 bacterial organisms (Table S1, Supplementary Information Section 4.4) grown under anaerobic conditions[12]. The draft GEMs of those organisms were reconstructed using a recent automatic reconstruction pipeline CarveMe[11]. The same set of performance metrics as used in the internal validation was used, except that AUROC was replaced by AUPRC (the Area Under

Precision-Recall Curve) for unbalanced datasets. We compared four different groups of models: the draft GEMs reconstructed from CarveMe (CarveMe), gap-filled GEMs by adding the top 200 reactions predicted by CHESHIRE and NHP (CHESHIRE-200 and NHP-200), and gap-filled GEMs by randomly adding 200 reactions from the universal BiGG reaction pool (Random-200).

We observed a high variability across the draft CarveMe models to predict fermentation profiles. One model correctly predicts all phenotypes (F1 = 1.0) but 14 others fail to predict any (F1 = 0.0). Adding the 200 NHP-predicted reactions barely improves the phenotypic predictions and the improvement is even worse than that after randomly adding 200 reactions (Fig. 3b-e). To the contrary, CHESHIRE-200 increases the mean performances significantly (Fig. 3b-e, Fig. S5) and, in particular, the F1 score for 11 of the 24 draft GEMs ($P < 0.01$, two-sided paired-sample t-test). Compared across the 9 fermentation metabolites, the biggest improvement of CHESHIRE over CarveMe draft models was observed on acetic acid followed by lactic acid production. Among the 24 draft GEMs, CHESHIRE increases the correct predictions of acetic acid and lactic acid phenotypes from 8 to 17 and 14 to 20, respectively. Finally, we demonstrated that the improved performance is not simply due to more reactions by showing a significantly better performance of CHESHIRE-200 than that of Random-200 ($P < 0.05$, two-sided paired-sample t-test).

To test whether the improvement of CHESHIRE over CarveMe-reconstructed draft models is generalizable to GEMs from other sources, we performed the same fermentation test on ModelSEED-reconstructed draft models. Similar results were observed, where CHESHIRE-200 improves the predictions over draft models, and draft models plus randomly selected 200

reactions (Fig. S6). NHP, again, shows no improvement. Our results suggest that CHESHIRE enables consistent improvement of phenotypic prediction over GEMs reconstructed from different pipelines.

The fermentation test comprises nine metabolites that are very close to central carbon pathways. To test if CHESHIRE can fill other types of gaps, we assessed CHESHIRE for predicting secretions of amino acids, substrate utilization for growth, and gene essentiality. The dataset of amino acid secretions measures production profiles of 20 amino acids for 25 bacterial GEMs (Table S2, Supplementary Information Section 4.5). This dataset is highly unbalanced with 478 positive phenotypes and 22 negative phenotypes. Similar to the fermentation test described above, CHESHIRE-200 outperforms Random-200 (F1 score: $P < 10^{-5}$, two-sided paired-sample t-test), and NHP-200 shows no improvement at all (Fig. 3f-i, Fig. S5). In particular, CHESHIRE increases the correct predictions of 67 amino acid secretions that are knowledge gaps in draft GEMs. Despite the significant improvement, the recall remains very low at about 20%, suggesting many remaining false-negative gaps.

Each of the substrate utilization and gene essentiality tests contains 5 GEMs (Table S3, Supplementary Information Section 4.6 and 4.7). The utilization of various carbon-, nitrogen-, phosphorus-, and sulfur-substrates for growth were tested using Biolog phenotype arrays[38] in a high-throughput manner. Essential genes were identified using gene knockout experiments and a gene is essential if its deletion is lethal and causes cell death. When tested on both datasets, CHESHIRE, however, fails to fill gaps in nearly all 5 GEMs, except for the growth phenotypes of

Bucillus subtilis where CHESHIRE-200 increases the F1 score of CarveMe draft model from 0.58 to 0.87 (Fig. S7). Notably, NHP fails in all the tests, including the growth phenotypes of *B. subtilis*.

To understand how CHESHIRE gap-fills the GEMs, we used Mixed Integer Linear Programming to infer which reaction(s) from the top 200 candidates that causally fill the gaps (Fig. 3a, Supplementary Information Section 6.4). The gap-filling guided by CHESHIRE corrected 10 false-negative predictions by adding a single acetic acid transport reaction between intracellular and extracellular via proton symporter, e.g., *Prevotella bergensis* DSM 17361 (Fig. 3j). This is not surprising as the poor performance of draft GEMs is mostly due to poor annotation of transporter genes[39–41]. Beside false negatives, CHESHIRE can also identify false-positive predictions. For example, the draft GEM of *Anaerobutyricum hallii* has lactate dehydrogenase and can theoretically produce or use lactate for growth when lactate is present in the culture medium. The draft model predicts lactate production, since lactate utilization is dispensable for maximal biomass production. However, this prediction contradicts experimental data. CHESHIRE-200 fills this gap by adding two NAD(P)H-mediated redox reactions (Fig. 3k) that enable maximization of growth rate by consuming lactate. Supported by previous reports of lactate consumption[42,43], this example shows that CHESHIRE can identify missing reactions that have consequences on distant fermentation pathways via a global and systematic effect. For gaps in amino acid secretions, causal reaction inference indicates that these gaps were solved by adding amino acid transportation reactions (Fig. 3l and m).

To assess the false positives added by CHESHIRE-200, we counted the number of

fermentation products and amino acids that were added by CHESHIRE but not observed experimentally. For the fermentation test, CHESHIRE adds a phenotype in 27 simulations (i.e., combinations of genome and metabolite), where 15 are true positives and 12 are false positives. For the amino acid test, all 67 simulations that predicted a gain of phenotype are true positives. These results suggest that the risk of introducing false positive phenotypes may depend on the model and phenotype. We further assessed whether these false positives may be linked to the bias of CHESHIRE to score specific types of reactions higher than the others. For nearly all enzymatic functional classes (Supplementary Information Section 6.5), we found a huge variability in the rankings of reactions catalyzed by enzymes that belong to each individual class (Fig. S8). Relatively, reactions catalyzed by dinucleosidetriphosphatase, hydratase, and cyclase are scored higher on average.

Discussion

Optimization-based GEM gap-filling has been long considered as a process of fitting a GEM to observed data[44]. 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 methods 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[21]. Previous studies[24,45,46] 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 collected. 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 collected.

We therefore present a new method CHESHIRE which uses deep learning techniques to resolve gaps in GEMs at reaction and phenotypic levels solely based on network topology. It is completely unsupervised without any input phenotypic data and thus improper to be compared to data-driven optimization-based methods directly. 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 methods, 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. In addition, CHESHIRE is computationally efficient than C3MM and NHP (Table S4, Supplementary Information Section 3.4). Most importantly, CHESHIRE has been validated on realistic biological datasets. To our best knowledge, such benchmark has not been performed for previous topology-based gap-filling methods. We showed that CHESHIRE significantly improves the phenotypic predictions of fermentation products and amino acids secretions over a total of 49 draft GEMs reconstructed from a mostly used automatic reconstruction pipeline CarveMe[11]. Despite the success, we found that substrate utilization and gene essentiality are gap-filling resistant cases for CHESHIRE and, broadly, the topology-based gap-filling methods.

Even in this worst scenario, CHESHIRE shows better performance than the competitive method NHP in the prediction of substrate utilization for *B. subtilis*. Since only 5 GEMs were used in each of the two tests, a comprehensive assessment of CHESHIRE over a larger GEM collection may be more informative of how much CHESHIRE struggles with these tasks.

Although CHESHIRE advances phenotypic predictions, the use of a universal pool and the top 200 reactions risk of adding reactions that do not exist (false positives). Correct predictions of phenotypes do not necessarily mean correct inference of missing reactions. It is likely that different enzymes carry the same metabolic functions (e.g., fermentation metabolism) from different substrates. While adding reactions is expedited by increasingly advanced machine learning methods, it is still a heavily manual task to trim reactions that are wrongly added. This challenge, termed as content removal, has been recognized as one of the two fundamental bottlenecks for GEM quality improvement[47]. This bottleneck is largely overlooked and highly time-consuming: a significant amount of time is required to identify which reactions should be removed. Initial community-driven efforts have been made to mitigate this challenge, including building new GEM quality standards such as MEMOTE[48] and developing novel frameworks for GEM quality assessment and improvement.

Alternative to content removal, the number of false positives can be reduced by limiting the database size for candidate reactions. Though at its infancy, database reduction is a valuable technology that has broad utility for both optimization- and topology-based methods. There are many possible routes of database reduction. First, a universal database can be split into genus-,

species-, or even phylogroup-specific databases by aggregating all reactions in GEMs that belong to individual taxa. Despite the pioneering efforts in AGORA models, we are still lacking a large-scale database of high-quality GEMs that cover a wide range of taxonomic diversity. Second, GapSeq[12] points out a promising direction to use genomic information for reducing a universal database to a small subset of reactions supported by gene annotations. By lowering the threshold for sequence homology, comparing protein domains and sequence signatures, and mapping content to distantly related organisms, more genes and candidate reactions can be annotated[47]. Finally, we should not ignore the possibility that gaps may be filled by altering the directionality of reactions without adding new ones. The current version of CHESHIRE does not consider reaction directionality and thus cannot fill gaps caused by wrong directions. A systematic integration of available thermodynamic data with GEMs will reduce the number of gaps and thus the total number of incorrectly introduced reactions. This deserves dedicated efforts, and we leave it as a future work.

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

Data Avaliability All biological datasets analyzed in this study are publicly available. All processed and simulation data supporting the findings of this study are available within the article and its supplementary materials.

Code Avaliability The source code of our computational framework is available at `https://github.`

com/canc1993/cheshire-gapfilling.

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

16. 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.

17. 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.

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

19. 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.

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

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

22. 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.

23. 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.

24. 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.

25. 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.

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

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

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

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

30. 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.

31. Samuel MD Seaver, Filipe Liu, Qizhi Zhang, James Jeffryes, José P Faria, Janaka N Edirisinghe, Michael Mundy, Nicholas Chia, Elad Noor, Moritz E Beber, et al. The modelseed biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic acids research*, 49(D1):D575–D588, 2021.

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

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. WS Mauchline and CW Keevil. Development of the biolog substrate utilization system for identification of legionella spp. *Applied and Environmental Microbiology*, 57(11):3345–3349, 1991.

39. 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.

40. 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.

41. 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.

42. Sudarshan A Shetty, Simone Zuffa, Thi Phuong Nam Bui, Steven Aalvink, Hauke Smidt, and Willem M De Vos. Reclassification of *Eubacterium hallii* as *Anaerobutyricum hallii* gen. nov., comb. nov., and description of *Anaerobutyricum soehngenii* sp. nov., a butyrate and propionate-producing bacterium from infant faeces. *International Journal of Systematic and Evolutionary Microbiology*, 68(12):3741–3746, 2018.

43. Sylvia H Duncan, Petra Louis, and Harry J Flint. Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Applied and environmental microbiology*, 70(10):5810–5817, 2004.

44. 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.

45. 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.

46. 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.

47. Yara Seif and Bernhard Ørn Palsson. Path to improving the life cycle and quality of genome-scale models of metabolism. *Cell Systems*, 12(9):842–859, 2021.

48. Christian Lieven, Moritz E Beber, Brett G Olivier, Frank T Bergmann, Meric Ataman, Parizad Babaei, Jennifer A Bartell, Lars M Blank, Siddharth Chauhan, Kevin Correia, et al. Memote for standardized genome-scale metabolic model testing. *Nature biotechnology*, 38(3):272–276, 2020.

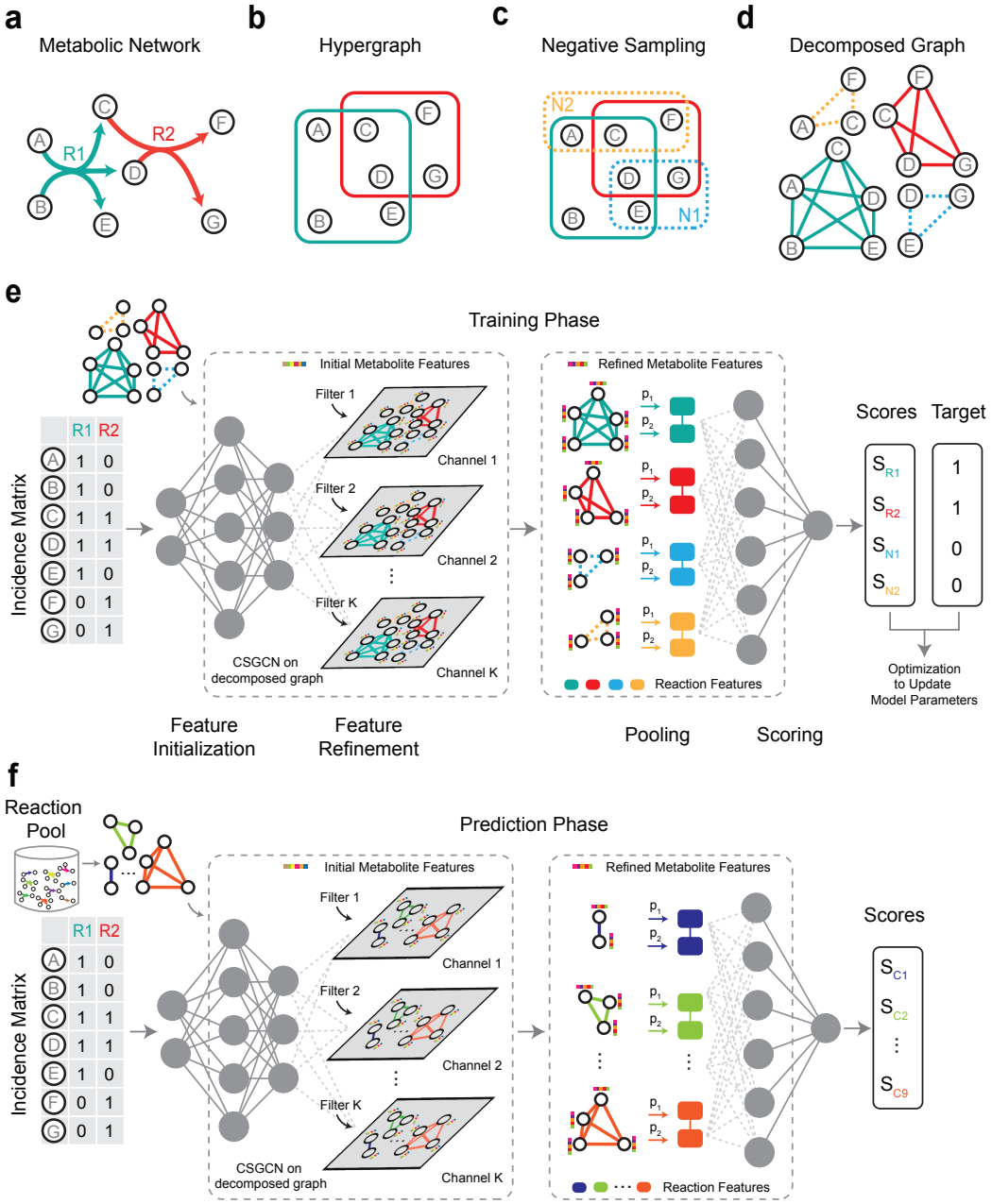
a
Metabolic Network
R1
R2
b
Hypergraph
c
Negative Sampling
N2
N1
d
Decomposed Graph
e
Training Phase
Incidence Matrix
R1 R2
Initial Metabolite Features
Filter 1
Channel 1
Filter 2
Channel 2
Filter K
Channel K
CSGCN on decomposed graph
Refined Metabolite Features
Reaction Features
Scores Target
Optimization to Update Model Parameters
Feature Initialization
Feature Refinement
Pooling
Scoring
f
Reaction Pool
Prediction Phase
Incidence Matrix
R1 R2
Initial Metabolite Features
Filter 1
Channel 1
Filter 2
Channel 2
Filter K
Channel K
CSGCN on decomposed graph
Refined Metabolite Features
Reaction Features
Scores

Fig. 1: CHESHIRE workflow. **a.** Schematic representation of a metabolic network. **b.** Hypergraph representation of the metabolic network. The hypergraph is undirected where each hyperlink connects metabolites that participate the same reaction. **c.** Negative sampling of the metabolic network. Solid and dashed boxes represent positive and negative reactions (e.g., N1, N2), respectively. **d.** Decomposed graph of the metabolic network, where each reaction (either positive or negative) is treated as a fully connected subgraph (solid and dashed lines represent positive and negative reactions, respectively). **e.** The architecture of CHESHIRE during training. The deep neural network takes the incidence matrix and the decomposed graph **d** as input, and consists of an encoder layer, 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 output confidence scores are compared to the target scores for updating model parameters. The gray dots represent the hidden neurons. **f.** The architecture of CHESHIRE during prediction. The neural network takes the incidence matrix and a decomposed graph built from candidate reactions as input and outputs confidence scores for candidate reactions based on the trained model parameters.

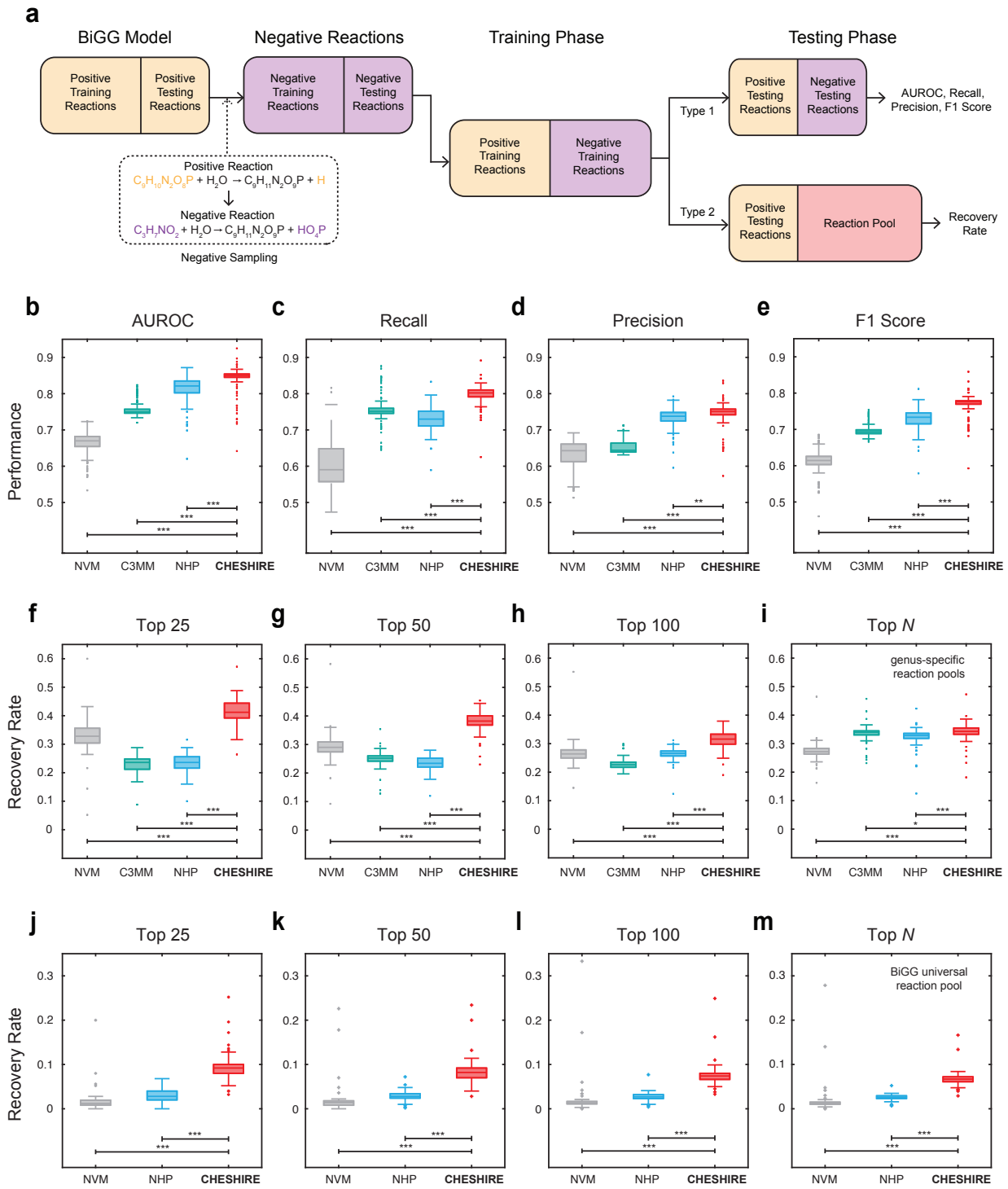
a
BiGG Model
Negative Reactions
Training Phase
Testing Phase
Positive Training Reactions
Positive Testing Reactions
Negative Training Reactions
Negative Testing Reactions
Positive Reaction
$C_9H_{10}N_2O_8P + H_2O \rightarrow C_9H_{11}N_2O_9P + H$
Negative Reaction
$C_3H_7NO_2 + H_2O \rightarrow C_9H_{11}N_2O_8P + HO_4P$
Negative Sampling
Type 1
Type 2
Reaction Pool
AUROC, Recall, Precision, F1 Score
Recovery Rate
b AUROC
c Recall
d Precision
e F1 Score
Performance
f Top 25
g Top 50
h Top 100
i Top N
genus-specific reaction pools
Recovery Rate
j Top 25
k Top 50
l Top 100
m Top N
BiGG universal reaction pool
NVM
C3MM
NHP
CHESHIRE

Fig. 2: Internal validation using artificially introduced gaps. **a.** Flowchart of internal validation. Two types of internal validation were performed. The former mixes artificially removed positive reactions and their derived negative reactions as candidate reactions, while the latter uses artificially removed positive reactions and real reactions from a universal reaction database as candidate reactions. **b-e.** 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. **f-i.** Reaction recovery rate of CHESHIRE vs. NHP, C3MM, and NVM with genus-specific reaction pools for gap-filling the BiGG GEMs by adding the top 25, 50, 100, and N reactions with the highest confidence scores (N is the number of artificially removed reactions). **j-m.** Reaction recovery rate of CHESHIRE vs. NHP and NVM with the entire BiGG universal reaction pool for gap-filling the BiGG GEMs by adding the top 25, 50, 100, and N reactions with the highest confidence scores. For **f-m**, only the BiGG models with over 100 reactions were tested, and C3MM was only validated on the genus-specific reaction pools due to the issue of scalability. Each data point represents the mean statistic over 10 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. Two-sided paired-sample t-test: $*P < 0.05$; $**P < 10^{-4}$; $***P < 10^{-16}$.

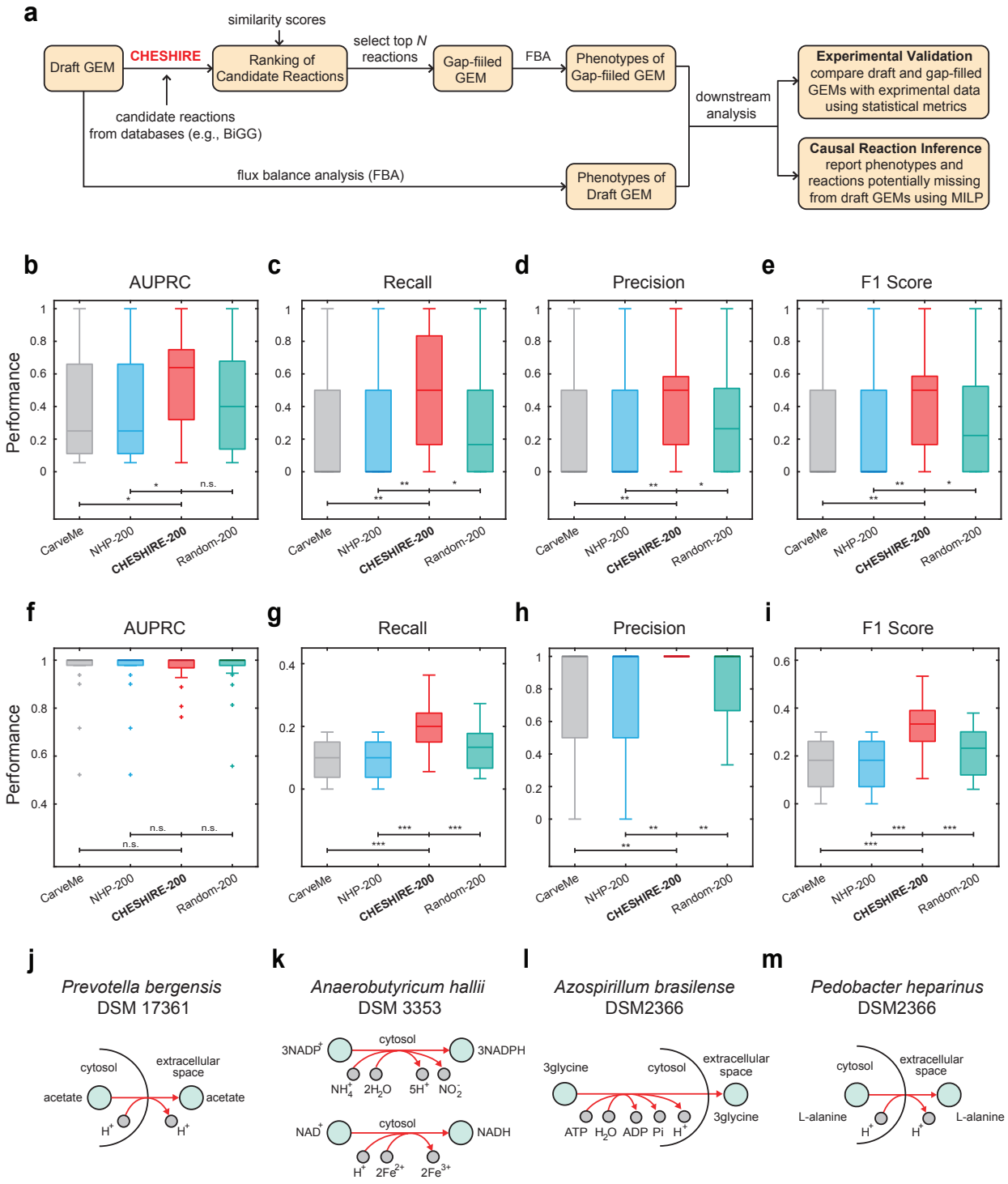
a
similarity scores
Draft GEM
CHESHIRE
Ranking of Candidate Reactions
select top N reactions
Gap-fiiled GEM
FBA
Phenotypes of Gap-fiiled GEM
downstream analysis
Experimental Validation
compare draft and gap-filled GEMs with exprimental data using statistical metrics
Causal Reaction Inference
report phenotypes and reactions potentially missing from draft GEMs using MILP
candidate reactions from databases (e.g., BiGG)
flux balance analysis (FBA)
Phenotypes of Draft GEM
b
AUPRC
c
Recall
d
Precision
e
F1 Score
Performance
CarveMe
NHP-200
CHESHIRE-200
Random-200
f
AUPRC
g
Recall
h
Precision
i
F1 Score
j
Prevotella bergensis
DSM 17361
cytosol
extracellular space
acetate
k
Anaerobutyricum hallii
DSM 3353
3NADP+
3NADPH
NAD+
NADH
l
Azospirillum brasilense
DSM2366
3glycine
ATP
ADP
Pi
m
Pedobacter heparinus
DSM2366
L-alanine

Fig. 3: External validation by predicting metabolic phenotypes. **a**. Flowchart of external validation. The predicted phenotypes from CHESHIRE-gapfilled GEMs are validated by comparison to experimental observation. For 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 Mixed Integer Linear Programming (MILP). **b-i.** Performance (AUPRC, Recall, Precision, and F1 score) of CHESHIRE and NHP in filling gaps in (**b-e**) 24 bacterial GEMs for fermentation metabolite production and (**f-i**) 25 bacterial GEMs for amino acid secretions. NVM was not included here due to its poor performance in internal validation. C3MM was not considered either, because it requires infeasible computational time for such a large candidate reaction pool (with almost 17,000 reactions). "CarveMe" represents the draft models reconstructed from CarveMe. "NHP-200" and "CHESHIRE-200" represent draft models plus 200 reactions predicted by NHP and CHESHIRE, respectively. For "Random-200", 200 randomly selected reactions from the universal BiGG database were added to the draft models. 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. Two-sided paired-sample t-test: n.s., not significant; $*P < 0.05$; $**P < 0.01$; $***P < 10^{-5}$. **j-m.** Examples of CHESHIRE-predicted reactions (red arrows) that causally gap-fill the observed phenotypes of acetate production (j), lactate production (k), and amino acid secretions (l and m). Abbreviations of cofactors: adenosine triphosphate (ATP); adenosine diphosphate (ADP); adenosine phosphate (AMP); phosphate (Pi); inorganic pyrophosphate (PPi); Coenzyme A (CoA); oxidized/reduced nicotinamide adenine dinucleotide (NAD^+/NADH); oxidized/reduced nicotinamide adenine dinucleotide phosphate ($NADP^+$/NADPH).