Deep learning-driven automatic reconstruction of genome-scale metabolic networks

Xiaoyi Liu
University of South Carolina
Hongpeng Yang
University of South Carolina
Chengwei Ai
Tianjin University
Yijie Ding
University of Electronic Science and Technology of China
Jijun Tang
University of South Carolina
Fei Guo (guofei@csu.edu.cn)
Central South University https://orcid.org/0000-0001-8346-0798

Article

Keywords: Metabolic network, Gap-filling, Genome-scale metabolic models, Missing annotation, Hyperlink prediction, Hypergraph convolution network, Hypergraph attention network

Posted Date: February 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2605759/v1

Additional Declarations: There is **NO** Competing Interest.

Deep learning-driven automatic reconstruction of genome-scale metabolic networks

Xiaoyi Liu[1], Hongpeng Yang[1], Chengwei Ai[5], Yijing Ding[4*], Jijun Tang[1,3*] and Fei Guo[2*]

[1]Computational Science and Engineering, University of South Carolina, 1244 Blossom Street, Columbia, 29208, South Carolina, USA.

[2]Computer Science and Engineering,, Central South University, 932 Lushan S Rd, Changsha, 410083, Hunan, China.

[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Nanshan, 518055, Shenzhen, China.

[4]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, 324000, Zhejiang, China.

[5]School of Computer Science and Technology, Tianjin University, No.135, Yaguan Road, Tianjin Haihe Education Park, Tianjin, 300350, Tianjin, China.

*Corresponding author(s). E-mail(s): wuxi_dyj@163.com; jtang@cse.sc.edu; guofeieileen@163.com;
Contributing authors: xiaoyil@email.sc.edu; hongpeng@email.sc.edu; ai_chery@163.com;

Abstract

Incomplete knowledge of metabolic processes impairs the accuracy of GEnome-scale Metabolic models (GEMs), hindering advancements in systems biology and metabolic engineering. To close this critical gap, we present CLOSEgaps, a machine learning-based algorithm that considers the hypergraph topology of metabolic networks and hypothetical reactions to predict missing reactions and identify gaps in GEMs. Extensive

results show that CLOSEgaps accurately gap-filled metabolic networks, filling over **96%** of artificially introduced gaps, and enhances the predictability of fermentation products in **24** wild-type GEMs. Furthermore, we integrate CLOSEgaps into a generalized workflow for automated metabolic network reconstruction, hereby named NICEgame, and found a notable improvement in producing four crucial metabolites (Lactate, Ethanol, Propionate, and Succinate) in two organisms. As a broadly applicable solution for any GEM or reaction, CLOSEgaps promises to enhance biotechnological and biomedical applications by improving GEM-based predictions and automating the NICEgame workflow.

Keywords: Metabolic network, Gap-filling, Genome-scale metabolic models, Missing annotation, Hyperlink prediction, Hypergraph convolution network, Hypergraph attention network

1 Introduction

Integrating a comprehensive understanding of biology at the systems level is essential for advancing bio-engineering, drug targeting, and medical therapies [1–3]. In this pursuit, metabolic networks and annotated genomes are leveraged to gain a holistic view of cellular functions. Despite these efforts, gaps still exist in our knowledge of cellular metabolic capabilities. Systematically uncovering these unknown metabolic processes has the potential to catalyze a wide range of medical and biotechnological applications [4].

Genome-scale metabolic models (GEMs) have emerged as a powerful tool for the systematic analysis of cellular metabolic functions [5–9]. With extensive use in the study of model organisms, these models are commonly evaluated through simulation techniques such as flux balance analysis (FBA) [10], which assumes a balanced flux of metabolites in the metabolic network via linear optimization [7, 11]. Recently, the availability of whole-genome sequencing data [12] has opened up new avenues for constructing wild-type GEMs. However, incomplete knowledge of metabolic processes results in incomplete wild-type models, characterized by missing reactions and unannotated gene products [4, 5, 13]. This presents an opportunity for GEM reconstruction through the gap-filling process [14], aimed at minimizing the number of missing components by adding new reactions to the model [15].

Various classic gap-filling algorithms have been developed and reviewed, including constraint-based modeling, GrowMatch, and comparative genomics methods [16]. However, these methods often rely on experimental techniques, making the process time-consuming and resource-intensive. To address these limitations, successful gap-filling needs a more practical approach. Thus, topology-based hypergraph approaches have gained popularity in the field of bioinformatics [17–19]. Neural Hyperlink Predictor (NHP) [20] and Coordinated Matrix Minimization (CMM) [21] are two hypergraph-based methods that can be used to efficiently gap-fill GEMs. However, these methods are

limited to the space of known annotated proteins and biochemistry [4]. Expanding our understanding of metabolic networks to include novel biochemistry requires exploration beyond the space of known biochemistry, a crucial step in advancing our understanding of metabolic network function.

Gap-filling in metabolic networks is a crucial step in understanding the metabolic functions of cells. However, current machine-learning approaches face limitations in accurately predicting metabolic reactions [22]. To overcome these limitations, we introduce a novel framework named hypergraph ConvoLution netwOrk and attention mechaniSm integrated Explorer for GAPS prediction of metabolism (CLOSEgaps). The extensive experimental results demonstrate that this hypergraph-based strategy leads to a significant improvement in gap-filling performance, with accuracy reaching over 96%. The diverse and multimodal nature of CLOSEgaps not only enhances the predictive model but also better reflects the actual metabolic reactions in the GEMs. By integrating the GEMs with the pool of hypothetical reactions predicted by CLOSEgaps, we aim to identify missing metabolic reactions and bridge the gap in understanding metabolic network function.

To further improve the gap-filling process, we present the Network Integrated maChine lEarninG Approach for MEtabolic network reconstruction (NICEgame) workflow, integrated with CLOSEgaps. NICEgame was applied to 24 wild-type GEMs reconstructed from CarveMe [23], suggesting novel biochemistry and significantly improving the predictability of metabolic fermentation products. The combination of the robust CLOSEgaps framework, biochemistry database, and simulation methods fully automates the gap-filling process, offering the potential to accelerate the completion of GEMs and enable effective bioengineering and drug-targeting strategies.

2 Results

2.1 A workflow of CLOSEgaps and NICEgame for metabolic network reconstruction

NICEgame is a workflow for automated metabolic network reconstruction and comprises three stages. The first stage maps metabolites to SMILES [24] using the public database of Biochemistry. The second stage uses CLOSEgaps to rank and add top N reactions to the wild-type GEMs to create gap-filled GEMs. The third stage applies flux simulation to predict metabolic phenotypes. CLOSEgaps tries to match the predictions of the wild-type model with the observed phenotype by adding reactions. If there's a discrepancy between the gap-filled and wild-type GEMs, it indicates unexplored pathways, which are addressed in the workflow using Linear Mixed-Integer Programming (LMIP) to infer the reactions causing the gaps.

CLOSEgaps is used in the NICEgame workflow to predict missing reactions in GEMs. It involves four key steps: negative reaction sampling, feature initialization, feature refinement, and evaluation and ranking (Figure 1), details in the Materials and Methods section 4. The first step (Figure 1A, B) uses

a metabolic network and the ChEBI database [25] to sample negative reactions. The second step maps metabolites to hypernode features and reactions to hyperedge features and applies a fully connected layer for feature initialization (Figure 1B, C). The third step (Figure 1D) refines the metabolic network structure and properties with hypergraph convolution and attention. In the final step (Figure 1E), the hyperedge feature is updated and multiplied by the transposed incidence matrix, then each reaction's feature vector is fed into a neural network to determine its confidence level.

In our study, CLOSEgaps offers a unique advantage in evaluating and ranking gap-filling reactions for wild-type GEMs with increased accuracy. Our method considers not only the likelihood of the reactions being present but also their impact on the metabolic network and its performance. By adding the highest-ranking reactions to reconcile gaps, CLOSEgaps leads to significantly improved predictions of metabolic phenotypes, which is the ultimate goal of reconstructing GEMs of microorganisms [26].

2.2 Metabolic network reconstruction

2.2.1 Evaluating performance of CLOSEgaps

Metabolic network reconstruction is a crucial step in understanding the metabolic processes of an organism. To ensure the accuracy of the predictions, it is imperative to have a robust gap-filling algorithm in place. In this study, we evaluated various training methods to determine the optimal CLOSEgaps model for gap-filling. CLOSEgaps was trained using the curated metabolic network dataset from the recently published Saccharomyces cerevisiae yeast8.5 model [27]. To enhance the accuracy of the predictions, negative reactions (reactions that do not exist) were also sampled from the data set [21, 28].

Therefore, negative reactions were sampled from the metabolites dataset using the public ChEBI database (https://www.ebi.ac.uk), which contains 44,359 metabolites (refer to the Methods Section 4.1 for more information). The GEMs were augmented by adding positive reactions and negative (fake) reactions in a 1 : 1 ratio. The positive and negative mixtures were then randomly split into 60% for training, 20% for validation, and 20% for testing. The performance of the models was evaluated using classical classification performance metrics: the F1 score (the harmonic mean of Recall and Precision), the Area Under the Receiver Operating Characteristic curve (AUC), the Area Under the Precision-Recall (AUPR), and Precision and Recall. To account for the complexity of metabolic reactions, the negative reactions were further sampled using imbalanced atom number and balanced atom number strategies (details can be found in the Methods Section 4.1.1 and Supplementary Information Section A).

As summarized in Table 1, CLOSEgaps achieves F1 score of 96%, 99% AUC, 99% AUPR, 95% Precision, and 96% Recall on the testing set. Table 2 summarizes the results with a balanced atom number negative sampling strategy, which showed a slightly decreased performance compared to the

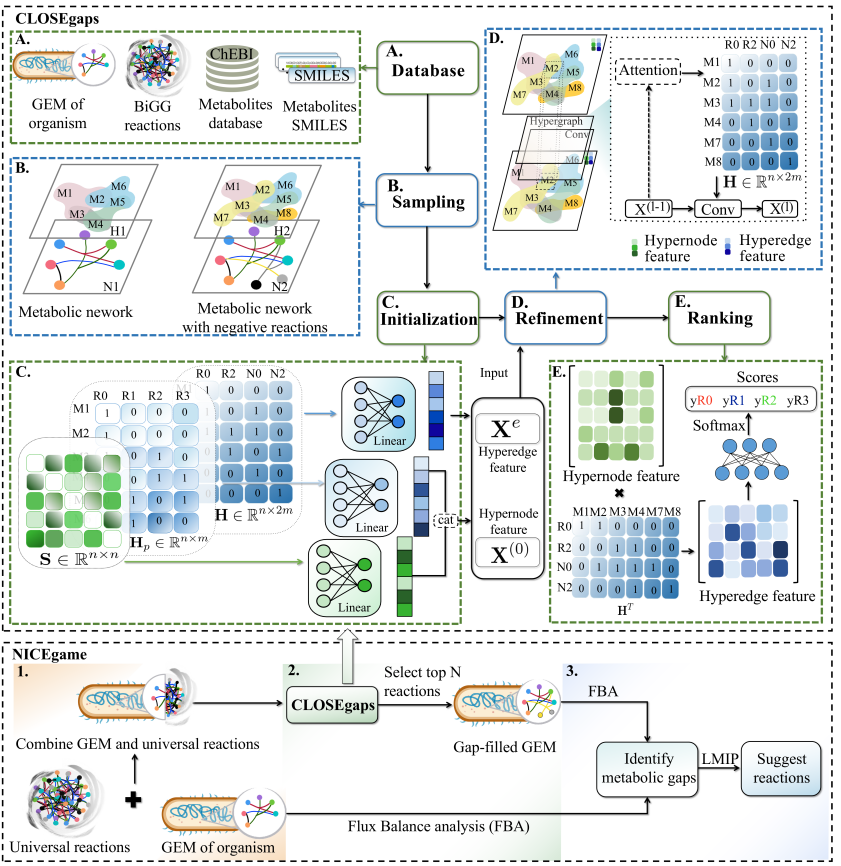


Fig. 1: An overview of NICEgame and CLOSEgaps structure. The NICEgame workflow uses a GEM model and universal reaction pool as input. (*1*) The wild-type GEM is merged with a universal database and (*2*) an essentiality analysis integrated with CLOSEgaps is performed on the wild-type GEM and the expanded network to identify which gaps can be filled. (*3*) FBA is utilized to predict fermentation phenotypes for the gap-filled GEMs and the wild-type GEMs, and LMIP causally suggests the reactions that lead to the false-positive phenotypes' production. (A-E) The architecture of CLOSEgaps. (A) The GEM, BiGG reactions, ChEBI metabolites database, and metabolites are represented by SMILES. (B) The GEM is used to construct a hypergraph, and the processed ChEBI database is used for negative sampling. (C) The incidence matrix of the hypergraph (positive incidence matrix and incidence matrix), and the similarity of metabolites in the metabolic network, are used to initialize features through a fully connected layer. Then, the positive incidence matrix feature and similarity matrix representations are concatenated as input to the next level. (D) The hypergraph convolution and hypergraph attention network are used to refine features. (E) The ranking module predicts missing reactions by refining hyperedge features, multiplying the transpose of the incidence matrix, feeding the result into a fully connected neural network, and using softmax for prediction. This module can both predict missing reactions and rescue reactions for gap-filling by leveraging the BiGG biochemistry data set (CLOSEgaps).

results in Table 1. This outcome is anticipated because swapping metabolites with ones with the same number of atoms can decrease the variability of the models' initial randomization [21]. Nevertheless, CLOSEgaps is robust and still achieves the best performance with other methods [17–20, 28–30]. Table 1 and Table 2 also showed that CLOSEgaps consistently outperformed other topology-based methods on the yeast8.5 model. The combination of CLOSEgaps with the hypergraph and SMILES representation demonstrates the power of hypergraph-integrated deep learning methods for predicting missing metabolic reactions. In conclusion, CLOSEgaps with the unbalanced negative sample was selected as the basic missing reactions prediction model for the rest of the study.

Table 1: Comparison of model performance in predicting missing reactions in the Saccharomyces cerevisiae metabolic network with negative sampling irrespective of reactant and product atom number balance.

Methods	F1 Score	AUC	AUPR	Precision	Recall
Node2Vector	0.7018	0.7422	0.7644	0.6778	0.7276
GCN	0.7097	0.7812	0.7774	0.7331	0.6879
NHP	0.6435	0.5301	0.5134	0.5063	0.8827
RGCN	0.7843	0.9343	0.9530	0.6657	0.9543
HGNN	0.7644	0.9111	0.8965	0.8828	0.6740
GraghSAGE	0.8499	0.9285	0.9443	0.7836	0.9284
CHESHIRE	0.9131	0.9570	0.9531	0.8874	0.9404
CLOSEgaps	**0.9574**	**0.9899**	**0.9885**	**0.9545**	**0.9602**

Table 2: Comparison of model performance in predicting missing reactions in the Saccharomyces cerevisiae metabolic network with negative sampling regarding reactant and product atom number balance.

Methods	F1 Score	AUC	AUPR	Precision	Recall
Node2Vector	0.6840	0.7415	0.7743	0.6636	0.7058
GCN	0.6836	0.7936	0.8346	0.5569	0.8847
NHP	0.6353	0.5265	0.5218	0.495	0.8867
RGCN	0.8140	0.9415	0.9567	0.7130	**0.9483**
HGNN	0.8330	0.9315	0.9261	**0.9165**	0.7634
GraghSAGE	0.8751	0.9304	0.9398	0.8292	0.9264
CHESHIRE	0.9102	0.9543	0.9413	0.8944	0.9264
CLOSEgaps	**0.9197**	**0.9735**	**0.9722**	0.8962	0.9443

2.2.2 Assessment of generalizability across various GEMs

To systematically evaluate the generalizability of CLOSEgaps, we selected four high-quality GEMs from the BiGG database [31] and annotated them using

a manual curation process (details in Methods). The selected GEMs include iMM904 (Saccharomyces cerevisiae S288C), iAF692 (Methanosarcina barkeri str. Fusaro), iAF1260b (Escherichia coli str. K-12 substr. MG1655), and iJO1366 (Escherichia coli str. K-12 substr. MG1655). We tested CLOSEgaps using 3,000 and 8,000 organic reactions, retrieved from the USPTO data set [32], the largest available organic chemical reaction library (details in Methods).

CLOSEgaps performed best with the highest average > 97% AUC and AUPR (as shown in Figure 2). On the iMM904 GEM, CLOSEgaps achieved an AUC and AUPR of 98.38% and 98.28%, respectively, demonstrating its accuracy in annotating gaps with any Saccharomyces cerevisiae GEMs. CLOSEgaps also demonstrated its generalized capability with an AUC and AUPR of 98.21%, 98.27% and 97.55%, 97.58% on the iAF1260b and iJO1366 Escherichia coli models, respectively. Additionally, CLOSEgaps achieved a 97.84% AUC and 97.83% AUPR for annotating missing reactions with the iAF692 Methanosarcina barkeri model. Significantly, CLOSEgaps consistently outperformed topology-based models such as CHESHIRE, GraphSAGE, HGNN, RGNN, NHP, GCN, and Node2Vec, with increases of up to 7.35% and 6.84% in AUC and AUPR, respectively (as shown in Figure 2).

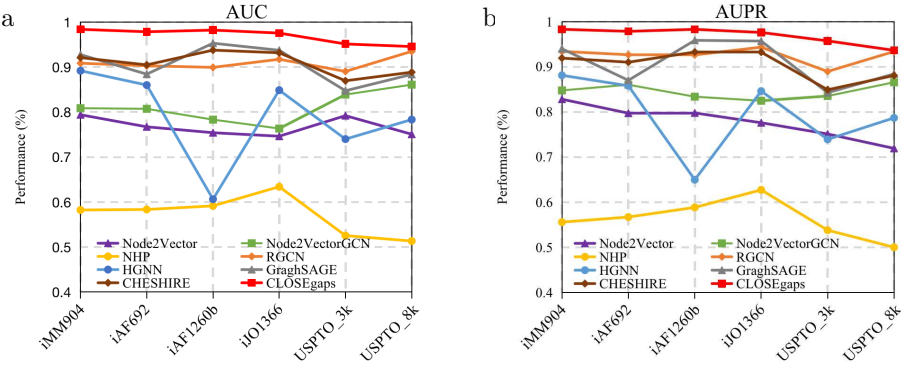


Fig. 2: Performance of CLOSEgaps evaluated via **a** AUC and **b** AUPR metrics and compared to CHESHIRE, GraphSAGE, HGNN, RGNN, NHP, GCN, and Node2Vector on four metabolic networks and the USPTO chemical reaction data set.

Furthermore, CLOSEgaps was also evaluated on a large chemical reactions dataset, the USPTO, to assess its ability to predict large-scale chemical reactions. The results show that CLOSEgaps outperformed other models and achieved the best classification performance on both the four GEMs and the USPTO data set, as depicted in Figures 2 USPTO$_3k$ and USPTO$_8k$. Specifically, CLOSEgaps had the highest AUC and AUPR values of 95.13% and 95.72% on the USPTO data set with 3,000 reactions (USPTO$_3k$), and 94.56%

and 93.62% on the USPTO data set with 8,000 reactions (USPTO$_8k$). Additionally, CLOSEgaps had the highest F1 score, Precision, and Recall results, indicating its reliability and accuracy in annotating gaps in various reaction networks (as shown in Figuer 3).

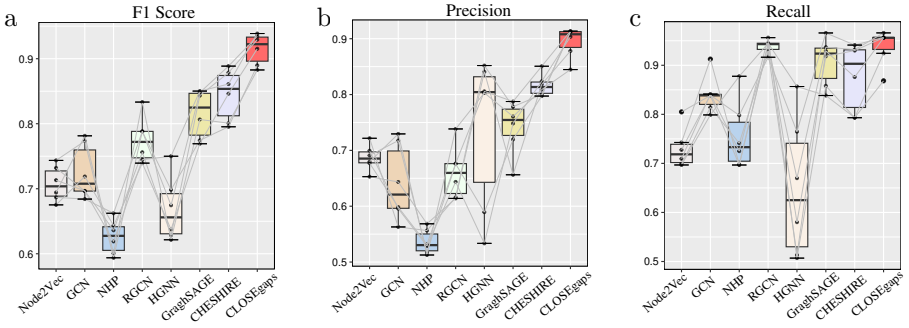


Fig. 3: Performance Comparison of CLOSEgaps and Other Methods on Four BiGG GEMs: Boxplots Showing **a** F1 Score, **b** Precision, **c** Recall. Each Dot Represents a GEM. Central Line in the boxplot Indicates the Median.

2.2.3 Assessment of robustness on artificially introduced gaps

To evaluate the robustness of CLOSEgaps for metabolic network reconstruction even when the initial reconstruction phase involves a highly incomplete network. Experiments were conducted on four metabolic networks from three species including yeast8.5, iMM904, iAF692, and iAF1260b. Statistics of the data sets used in the experiments are presented in Table 3. The negative reactions were also sampled with the ChEBI database with each metabolic network. The recovery rate is used to evaluate the performance of CLOSEgaps in terms of the number of true-positive predictions made among the top N missing reactions. Therefore, 20%, 40%, 60%, and 80% of metabolic reactions were randomly removed to introduce hypothetical gaps in each of the GEMs. The reactions that remained were utilized as the training set, while those that were removed served as the testing set.

The evaluation of CLOSEgaps for metabolic network reconstruction is presented in Figure 4. The results showed that CLOSEgaps displayed a remarkable performance with a recovery rate surpassing 96% across all evaluated GEMs. In the yeast8.5 model, when 20% of reactions were removed as gaps, CLOSEgaps successfully filled 96.82% of them, outperforming CHESHIRE [28] which recovered 94.23% of the missing reactions. This outcome highlights the superiority of CLOSEgaps' hypergraph-based architecture and its ability to effectively fill metabolic gaps. For example, the use of Node2Vector [17] with embedding initialization resulted in the filling of only 70.58% of the gaps.

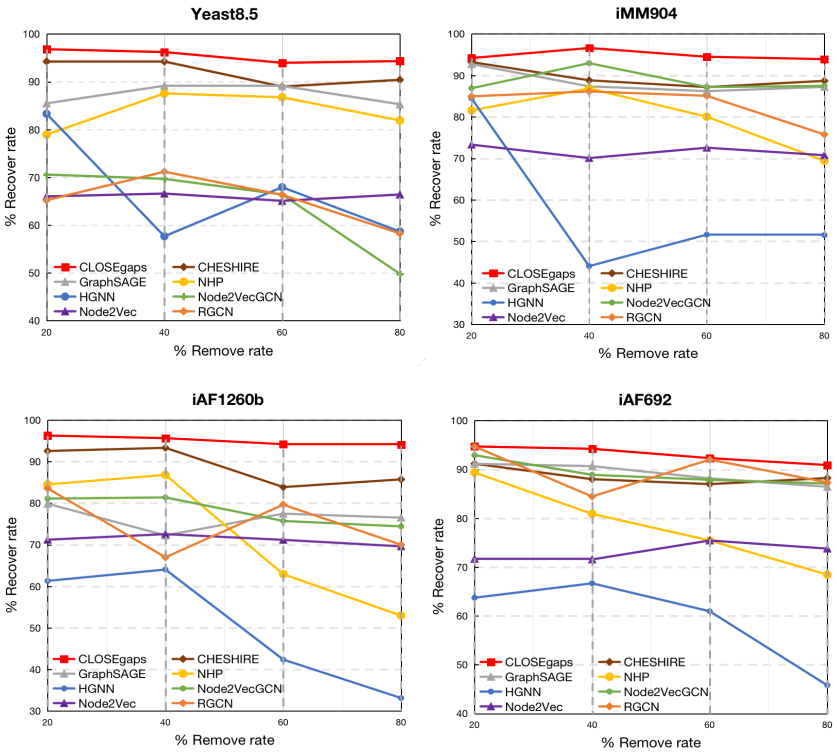


Fig. 4: Comparison of CLOSEgaps with other methods (CHESHIRE, GraphSAGE, HGNN, RGNN, NHP, GCN, and Node2Vec) in the recovery of reactions in metabolic networks from four GEMs. Reactions were removed randomly from the GEMs and treated as unobserved in the testing set.

Additionally, CLOSEgaps displayed remarkable stability across different reaction removal percentages, indicating its ability to produce accurate results even with limited training data. These results demonstrate the practicality and potential for widespread application of CLOSEgaps in metabolic network reconstruction and gap-filling GEMs, even in cases of substantial gaps or incomplete networks during the initial reconstruction phase.

2.3 Fermentation process improvement

2.3.1 Simulation study of anaerobic growth

The NICEgame workflow was applied to a data set comprising fermentation profiles of 24 bacterial organisms grown under anaerobic conditions [13] as described in the Supplementary Information Section E and Table D1. As CLOSEgaps is a key component of NICEgame, the CLOSEgaps algorithm was utilized to identify metabolic gaps between the silico phenotypes and available experimental data. This was achieved through the simulation of anaerobic

growth conditions and capturing of false-positive phenotypes. To fill these gaps, we targeted reactions that were crucial for growth in the wild-type GEM, but are currently unexplored, by integrating CLOSEgaps with the hypothetical reactions pool (BiGG universal pool). CLOSEgaps successfully identified 13 false-positive phenotypes compared to available experimental data. It is noteworthy that the SMILES information for these 24 GEMs was excluded from the analysis. The results, as shown in the first table of Figure 5, demonstrate the capability of CLOSEgaps to identify metabolic gaps, indicated by orange boxes. The reactions rescued for specific phenotype growth were essential for growth in the wild-type GEM and filled previously unknown pathways in the GEM with the BiGG data set.

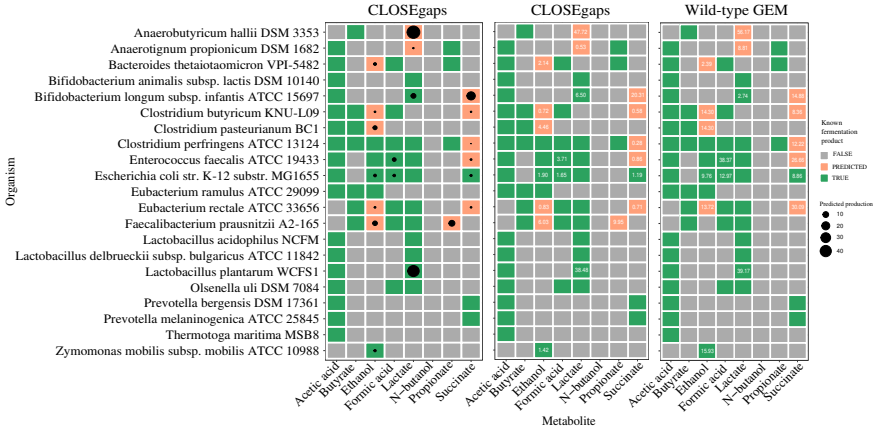


Fig. 5: Predicted anaerobic fermentation products for 24 bacterial organisms: A comparison of CarveMe and CLOSEgaps models (Excluding Cutibacterium acnes KPA171202, Clostridium acetobutylicum ATCC 824, and Aminobacterium colombiense DSM 12261). Green highlights indicate known fermentation products as reported in literature, while the orange box demonstrates CLOSEgaps' correction of false-positive product predictions in gap-filled GEMs. The size of the points represents the predicted metabolite production (columns) for each organism (row), with predictions made using Minimize-Total-Flux (MTF) flux balance analyses and depicted in black.

2.3.2 CLOSEgaps-assisted optimization of fermentation pathways

The CLOSEgaps tool is capable of identifying missing reactions that can impact distant fermentation pathways through a systematic and global approach [13]. Our study focused on the metabolic network of Faecalibacterium prausnitzii A2-165, which was expected to have a positive maximum flux for ETOHtex and PPAt2pp, but experimentally showed otherwise. CLOSEgaps then predicted the addition of two critical reactions, Ethanol transport via diffusion and H+/Propionate symporter (periplasm), which led to an increase

in the metabolic network's maximum growth rate and improved fermentation product production.

Additionally, previous research (Zimmermann et al., 2021; Bernstein et al., 2021; Chen et al., 2022) has demonstrated the potential to optimize metabolic pathways in GEMs for the production of valuable fermentation products through simulation of reaction fluxes. CLOSEgaps' gap-filling process has the capability of enhancing the utilization of metabolic networks and producing more products. In our study, CLOSEgaps improved the production of four metabolites (Lactate, Ethanol, Propionate, and Succinate) in two organisms, Bifidobacterium longum subsp. infantis ATCC 15697 and Faecalibacterium prausnitzii A2-165 (as shown in Figure 5, the last two tables). In the organism Faecalibacterium prausnitzii A2-165, Ethanol production improved to 6.03, and Propionate production improved to 9.95. In the organism Bifidobacterium longum subsp. infantis ATCC 15697, Succinate production increased by 5.43, and Lactate production increased by 4.03. However, CLOSEgaps failed to gap-fill three GEMs: Cutibacterium acnes KPA171202, Clostridium acetobutylicum ATCC 824, and Aminobacterium colombiense DSM 12261.

2.3.3 Predicting fermentation products in anaerobic GEMs

Significantly, this biological experiment demonstrates that CLOSEgaps enhances the predictability of metabolic fermentation products for 24 wild-type GEMs reconstructed from CarveMe [23] grown in anaerobic conditions. The CLOSEgaps method scores the likelihood of each hypothetical reaction being present in the wild-type GEM [33] by solely analyzing GEMs, offering an efficient and automated approach to gap-filling. This approach is superior to relying on experimental data, which is limited by the availability of high-quality metabolic pathways and reaction measurements [13]. The wild-type GEMs were reconstructed using CarveMe [23]. We downloaded all 11,893 reactions from BiGG (http://bigg.ucsd.edu) to form a candidate reaction pool. Instead of adopting a fixed cutoff score, we included the 200 reactions with the highest confidence scores in the GEMs, ensuring that reactions causing energy-generating cycles (EGCs) [34] were only added if the EGCs could be resolved by adjusting their flux bounds.

We achieved a significant improvement in performance (as shown in Figure 6) by integrating 200 reactions predicted by CLOSEgaps into the draft models reconstructed from CarveMe [23]. These results demonstrate the effectiveness of our topology-based gap-filling strategy. CLOSEgaps outperformed CarveMe with a higher mean F1 score of 64.38% and AUC of 92.90%, compared to CarveMe's mean F1 score of 30.51% and AUC score of 72.41%. Despite this remarkable performance, CLOSEgaps was unable to gap-fill three GEMs (Cutibacterium acnes KPA171202, Clostridium acetobutylicum ATCC 824, and Aminobacterium colombiense DSM 12261) that each produced a single metabolite.

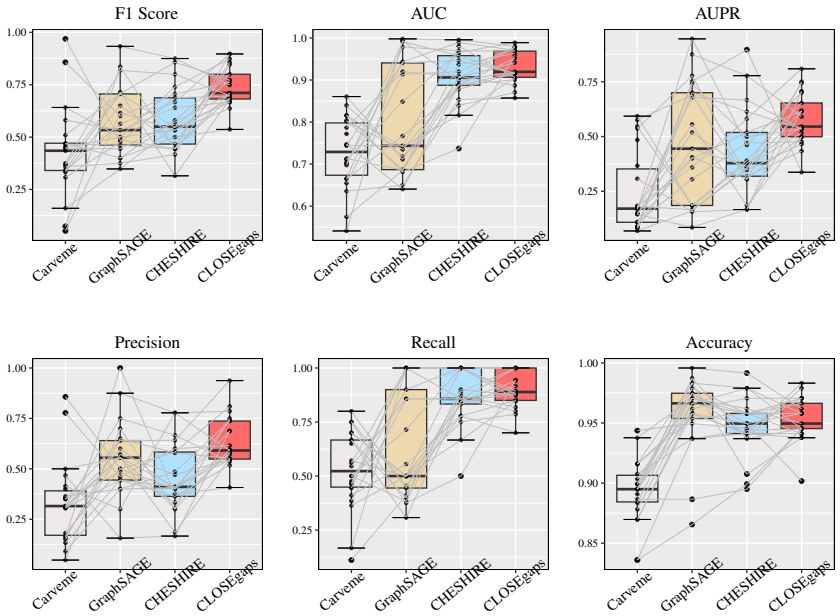


Fig. 6: Performance comparison on metabolic networks gap-filled by GraphSAGE, CHESHIRE, and CLOSEgaps. Model performance was measured by F1 score, AUC, AUPR, Precision, Recall, and Accuracy on 24 GEMs, each represented by a dot, which was gap-filled using GraphSAGE, CHESHIRE, and CLOSEgaps. The "CarveMe" data represents the draft models reconstructed from the CarveMe pipeline, while the gap-filled model represents the draft models augmented with an additional 200 reactions predicted by models. The central line in each boxplot represents the median.

2.3.4 Pathway visualization and potential reaction heatmap

Our model offers a unique advantage in studying the central metabolic network of Saccharomyces cerevisiae. Three pathways including glycolysis, pentose phosphate, the tricarboxylic acid cycle (TCA cycle), and the amino acid pathway are visualized in Figure 7a. Our model's feature embedding result, shown in Figure 7b, provides a more in-depth interpretation of the metabolic pathways. The heatmap in the figure provides a visual representation of the relationships between metabolites and highlights potential reaction clusters based on the clustering analysis. The clustering results suggest that metabolites within the same cluster are likely to participate in the same reaction. The heatmap provides a clear and intuitive way to identify potential reactions based on the relationships between metabolites.

The formation of clusters in the visualization is a strong indicator of metabolic reactions occurring between these metabolites. The glycolysis process, for example, is demonstrated in two stages, starting with the dehydrogenation of D-Glucose 6-phosphate and ending with the formation of Coenzyme A in the TCA cycle. In contrast, the lack of cluster formation in

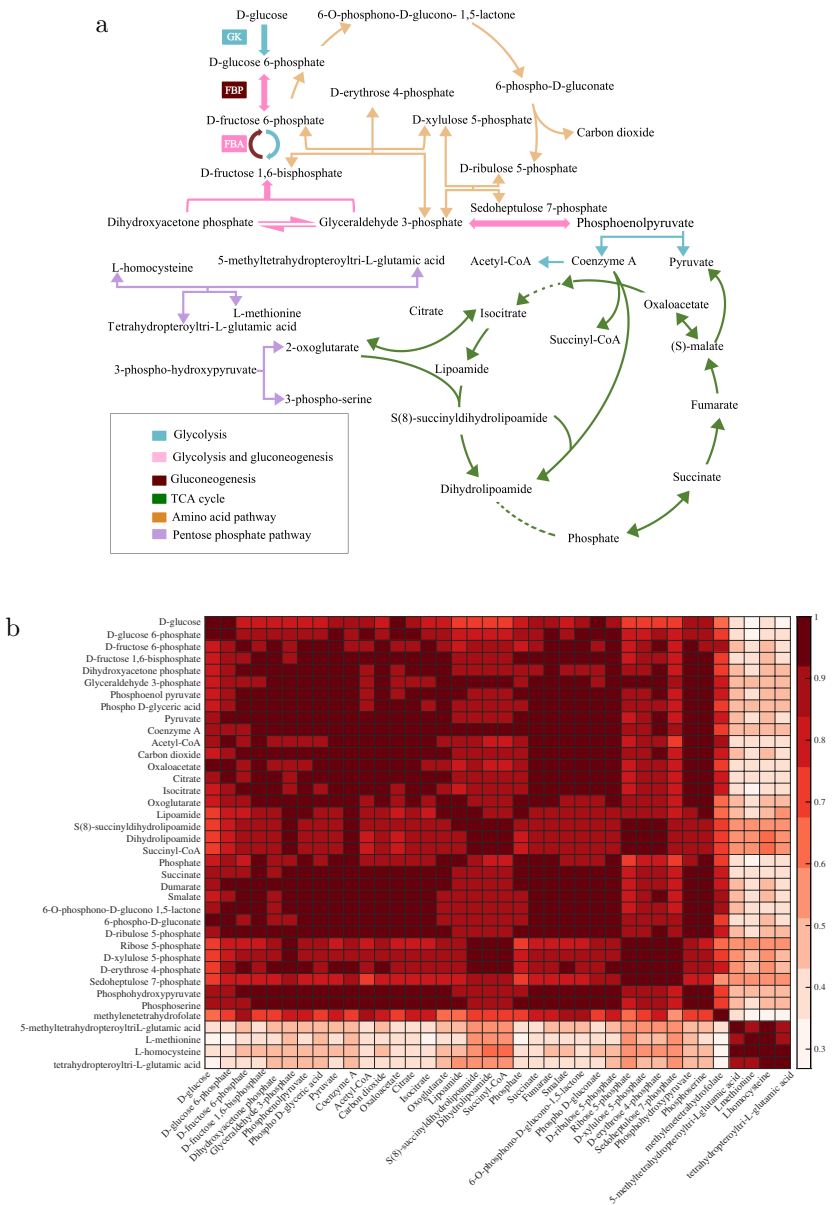


Fig. 7: **a** Pathway visualization of central carbon metabolism pathways and critical metabolites in Saccharomyces cerevisiae. The metabolic pathways of glycolysis are depicted in light blue, the reversible steps of glycolysis/gluconeogenesis are depicted in pink, the dedicated steps of gluconeogenesis are depicted in red, the tricarboxylic acid (TCA) cycle is depicted in green, and the glyoxylate shunt is depicted in orange. Pathways dedicated to amino acids are depicted in purple. Key enzymes are labeled as GK (glucokinase), FBP (fructose bisphosphatase), and FBA (fructose bisphosphate aldolase). **b** Heatmap representing the relationships between metabolites and potential reaction clusters. The color scale represents the strength of the relationships between metabolites, with darker colors indicating a stronger relationship. Clustering analysis was performed on the heatmap, grouping similar metabolites together and highlighting potential reaction clusters.

the amino acid pathway suggests that the metabolites do not react with each other. In addition, CLOSEgaps distinguishes itself from conventional graph embedding methods by being able to learn a high-order, relation-aware embedding for link prediction. To better illustrate the difference between CLOSEgaps and other baseline methods, such as Node2Vec and GCN, we present a visual demonstration of the learned reaction embedding through the t-SNE tool [35], which is explained in detail in the Supplementary Information Section C, Figure C2.

3 Discussion

In this study, we introduce a new and innovative metabolic network reconstruction workflow and machine learning model, NICEgame and CLOSEgaps, which is aimed at predicting missing reactions. CLOSEgaps is a fully data-driven model that is built on the foundation of curated GEMs and hypothetical reaction data. The comprehensive evaluations of both internal and external test sets reveal that our method effectively curates GEMs, leading to improved predictions of missing metabolic reactions and functions that can later be verified through experimentation. Our approach demonstrated a high mean recovery rate of 95.34% when benchmarked on the yeast8.5 model using artificially introduced gaps. Additionally, by integrating mixed-integer linear programming, we further benchmarked CLOSEgaps using fermentation data and improved the prediction of fermentation products for 24 bacterial organisms, achieving a mean F1 score of 74.24% based on wild-type GEMs reconstructed from CarveMe [23]. The use of CLOSEgaps resulted in improved production of Ethanol and Propionate in Faecalibacterium prausnitzii A2-165, and of Succinate and Lactate in Bifidobacterium longum subsp. infantis ATCC 15697. These results highlight the potential of CLOSEgaps as a valuable tool in optimizing fermentation pathways and metabolic network reconstruction.

CLOSEgaps and NICEgame address a crucial need for efficient GEM curation to enhance in silico predictions of missing metabolic reactions and functions. There is room for future improvement through the provision of additional information, advanced enzyme prediction [36, 37], and design tools [38–41] to aid in experimental analysis. This work has significant implications for the study of metabolic networks. CLOSEgaps can be applied to any existing GEM, advancing the fields of biotechnology and biomedicine. The key reactions predicted by CLOSEgaps can serve as a valuable resource for identifying new ways to improve strain performance, such as increasing biomass or product yield. CLOSEgaps holds tremendous potential for identifying metabolic network gaps, reconstructing metabolic networks, and rational design.

4 Methods

4.1 Data collection and preprocessing

We utilized CLOSEgaps to predict missing reactions in both metabolic networks and chemical reaction datasets. We applied it to the metabolic network of Saccharomyces cerevisiae using the yeast8.5 metabolic network data set [27]. To prepare the data, we removed the reaction location markers, converted reversible reactions into two separate reactions when reactants and products were different, and manually collected SMILES strings for each metabolite from several public data sources, as shown in Table 4. The data collection and cleaning process is depicted in Figure 8. We obtained 52,960 SDF files for metabolites from ChEBI (https://www.ebi.ac.uk), and after filtering out invalid SMILES strings, we had a collection of 44,359 metabolites. Additionally, we downloaded 11,893 reactions from BiGG (http://bigg.ucsd.edu) [31], which were collected from 79 metabolic networks of various organisms, forming a candidate reaction pool for the external experiments. We then eliminated candidate reactions that were biomass, exchange, demand, sink, or already present in the network. This process was repeated for other metabolic networks of three species: Saccharomyces cerevisiae, E.coli, and M.barkeri, as summarized in Table 4.

The USPTO chemical reaction dataset consisting of 1.8 million text-mined reaction equations in SMILES notation, recorded from 1976 to 2016 [42]. We randomly selected 3,000 and 8,000 reactions from the USPTO dataset [32] which comprises 20,000 chemical reactions, was acquired from Lowe, D. M. [42] for testing. The statistics of each dataset are presented in Table 3.

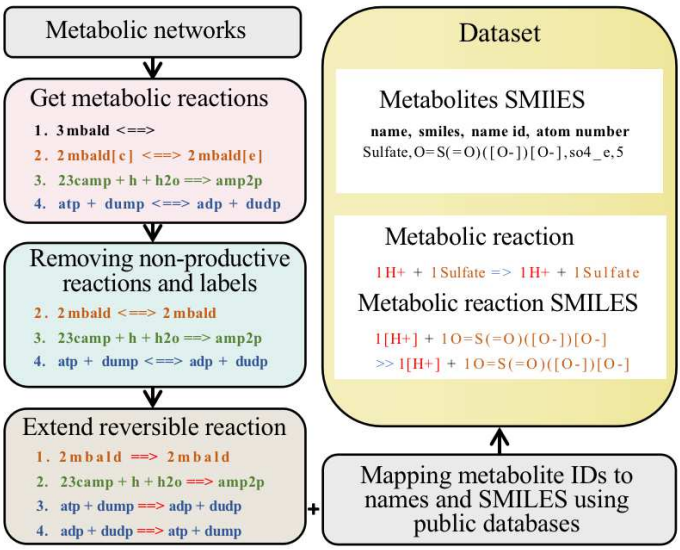


Fig. 8: Overview of the data cleaning process

Table 3: Metabolic network and chemical reaction dataset statistics.

Dataset	Species	Metabolites (vertices)	Reactions (hyperlinks)
Yeast8.5	Saccharomyces cerevisiae (Jul. 2021)	1136	2514
iMM904	Saccharomyces cerevisiae S288C (Oct. 2019)	533	1026
iAF1260b	Escherichia coli str.K-12 substr.MG1655	765	1612
iJO1366	Escherichia coli str.K-12 substr.MG1655	812	1713
iAF692	Methanosarcina barkeri str.Fusaro	422	562
USPTO_3k	Chemical reaction	6706	3000
USPTO_8k	Chemical reaction	15405	8000

Table 4: Summary of metabolic networks and candidate reactions public databases.

Name	Website
ChEBI	https://www.ebi.ac.uk/
BiGG	http://bigg.ucsd.edu/
Kegg	https://www.genome.jp/kegg/kegg2.html
Metanetx	https://www.metanetx.org/
Pubmed	https://pubmed.ncbi.nlm.nih.gov/17202168/
Metacys	https://metacyc.org/

4.1.1 Negative sampling strategies

To achieve accurate missing reaction prediction, we balanced the specificity and sensitivity of our model by sampling negative reactions during training, following Zhang et al. [21]. We generated corresponding negative hyperlinks for each positive hyperlink in the hypergraph by using half of the nodes from positive hyperlinks and half from processed ChEBI data, as per Chen et al.'s findings that such a method was unlikely to result in another valid reaction [28]. Additionally, we ensured reactant-product atomic number balance by creating atom-balanced negative reaction samples. Experiments were conducted with varying ratios of atom selection and negative sampling, including 0.5, 0.2, and 0.8 percent metabolite replacement and $(1:1)$, $(1:2)$, and $(1:3)$ negative to positive hyperlink ratios (details in Supplementary Information Section A).

4.1.2 Metabolite similarity calculation

To express the similarity relationship between metabolites comprehensively and accurately, the chemical structure of the metabolites was used to calculate their similarity. Effective methods for determining fingerprint-based similarity include the Tanimoto coefficient, gaussian, and cosine similarity (Supplementary Information Section A, Figure A1c). The chemical structure of the metabolites was represented in the standard SMILES format.

4.2 CLOSEgaps

The prediction of missing reactions in GEMs is accomplished through the use of a machine learning algorithm named CLOSEgaps, which consists of three

key steps in its learning architecture: feature initialization, feature refinement, and hypothetical reaction ranking. The overview of CLOSEgaps is illustrated in Figure 1.

4.2.1 Feature initialization

For a set of metabolic reactions and their corresponding metabolic network. To describe the complex relationships between metabolic in a GEM. We use a powerful graph model hypergraph to present its structure, where a hyperedge can connect more than two vertices indicating each hyperlink represents a metabolic reaction and connects corresponding reactant and product metabolites. Let $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$, with vertices $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ and hyperedges $\mathcal{E} = \{e_1, e_2, ..., e_m\}$, $e_i \subseteq \mathcal{V}, i = 1, 2, ..., m$ for the metabolic network of a given GEM. In general, the hypergraph $\mathcal{H}$ can be represented by its incidence matrix $\mathbf{H}_p \in \mathbb{R}^{n \times m}$, hereby named as positive incidence matrix, capturing the presence or absence of each metabolite in any given reaction through the use of boolean values. This can be obtained from the stoichiometric matrix by converting its non-zero values into binary ones through a process of binarization. The number of metabolites is denoted by n while the number of reactions is represented by m. Negative sampling of reactions is also represented as an incidence matrix $\mathbf{H_n}$. Suppose $\mathbf{H} = [\mathbf{H}_p\mathbf{H}_n] \in \mathbb{R}^{n \times 2m}$ including n metabolic and $2m$ reactions as Figure 1B.

In addition, to compare the similarity of metabolites, their chemical structures were utilized. Previous research has found that the Tanimoto coefficient is a useful measure of fingerprint-based similarity [43]. The closer the chemical structures of two metabolites are, the higher their similarity score will be. In order to perform this comparison, the chemical structures of metabolites were extracted in standard SMILES form and used to construct the Tanimoto similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$.

Then, we apply the fully connected layer to $\mathbf{H}$, $\mathbf{H}_p$, and $\mathbf{S}$. Next, the positive incidence and similarity matrix representations are concatenated as a hypernode feature $\mathbf{X}^{(0)}$ and hyperedge feature $\mathbf{X}^e = \{x_1^e, x_2^e, ..., x_m^e\}$ inputting to the hypergraph convolution network. This initial feature representation encodes crude information regarding the topological relationship between a metabolite and all the reactions within the metabolic network. The formula is as follows:

$$\mathbf{X}^e = Linear(\mathbf{H}), \tag{1a}$$

$$\mathbf{X}^{(0)} = Cat(Linear(\mathbf{H}_p), Linear(\mathbf{S})) \tag{1b}$$

4.2.2 Feature refinement

In the feature refinement step of our study, we leveraged hypergraph convolution and hypergraph attention to treat each reaction as a fully-connected subgraph. We utilized a multi-channel hypergraph convolution network and multi-heads attention module to learn deep embeddings on high-order graph-structured data, as introduced in [30]. Hypergraph convolution defines a basic

formulation for performing convolution on a hypergraph, which facilitates efficient information propagation between vertices by exploiting the high-order relationship and local clustering structure. This approach facilitated the enhancement of the feature representation of each metabolite by taking into account the attributes of other metabolites involved in the same reaction. For a hypergraph incidence matrix $\mathbf{H}$, each hyperedge ϵ is assigned a positive weight $W_{\epsilon\epsilon}$, with all the weights stored in a diagonal matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$. The degree of vertex v_i is the number of hyperlinks containing that node, which can be computed as Eq 2, the diagonal node degree matrix of a hypergraph by $\mathbf{D} \in \mathbb{R}^{m\times m}$.

$$D_{ii} = \sum_{\epsilon=1}^{m} W_{\epsilon\epsilon}\mathbf{H}_{i\epsilon} \tag{2}$$

And, the hyperedge degree is defined as Eq 3, the diagonal edge degree matrix of a hypergraph by $\mathbf{B} \in \mathbb{R}^{m\times m}$

$$B_{\epsilon\epsilon} = \sum_{\epsilon=1}^{n} \mathbf{H}_{i\epsilon} \tag{3}$$

Then, one step of hypergraph convolution is defined as Eq. 4

$$x_i^{(l+1)} = \delta(\sum\sum_{j=1}^{n}\sum_{\epsilon=1}^{m} H_{i\epsilon}H_{j\epsilon}W_{\epsilon\epsilon}X_j^{(l)}\mathbf{P}), \tag{4}$$

where $x_i^{(l)}$ is the embedding of the ith vertex in the lth layer. $\delta(\cdot)$ is a non-linear activation function, here we utilize ReLU [44]. $\mathbf{P} \in \mathbb{R}^{F^l\times F^{l+1}}$ is the weight matrix between the lth and $(l+1)$th layer.

So, Eq. 4 can be written in a matrix form as Eq. 5.

$$\mathbf{X}^{(l+1)} = \delta(\mathbf{H}\mathbf{W}\mathbf{H}^T\mathbf{X}^l\mathbf{P}), \tag{5}$$

where $\mathbf{X}^l \in \mathbb{R}^{n\times F^l}$ and $\mathbf{X}^{l+1} \in \mathbb{R}^{n\times F^{l+1}}$ are the input of the (l)th and $(l+1)$th layer, respectively.

Alternatively, in contrast to the propagation is directional and asymmetric, a row normalization is also feasible as Eq 6

$$\mathbf{X}^{(l+1)} = \delta(\mathbf{D}^{-1}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^T\mathbf{X}^l\mathbf{P}) \tag{6}$$

Here, $\mathbf{D}$ and $\mathbf{B}$ are the vertex and hyperedge degree matrices in a hypergraph, respectively.

Song et al. [45] proposed a hypergraph attention mechanism that enables the learning of a dynamic incidence matrix to represent the connections between vertices and assigns numerical values to reflect the intensity of these connections. For the lth hypergraph convolution layer, $x_i^{(l)}$ is the embedding of the v_i vertex and x_ϵ^e is the embedding of its associated hyperedge ϵ, the

attentional score is defined as Eq 7

$$H_{i\epsilon} = \frac{exp\left(\delta\left(sim\left(x_i^{(l)}\mathbf{P}, x_\epsilon^e\mathbf{P}\right)\right)\right)}{\sum_{k\epsilon\mathcal{N}_i} exp\left(\delta\left(sim\left(x_i^{(l)}\mathbf{P}, x_k^e\mathbf{P}\right)\right)\right)}, \tag{7}$$

where $\mathcal{N}_i$ is the neighborhood set of v_i. A similarity function denoted as sim $(\cdot)$ is employed to compute the pairwise similarity between each pair of vertices.

4.2.3 Hypothetical reaction ranking

Finally, the prediction of missing reactions in the metabolic network was formulated as a binary classification task. To incorporate the metabolite features into a hyperlink-level representation, the output feature vector was multiplied with the transpose of the incidence matrix using matrix multiplication (Equation 8a). This was followed by the application of a fully-connected layer, and the feature vector of each hyperlink was fed into a softmax function to predict the probability distribution over two classes: existence or non-existence. The softmax function provided a convenient way to convert the raw predictions of the model into a probability distribution, allowing for more informed decisions about the existence of a reaction in the metabolic network.

$$\mathbf{Y} = Softmax(Linear(\mathbf{H}^T\mathbf{X})), \tag{8a}$$

where $Y = \{y_1, y_2, ..., y_m\}$ represents the prediction score for each hyperlink.

4.3 Loss function

We use the cross-entropy loss function as Eq. 9

$$Loss = \frac{1}{n+m}\left(\sum_{e_i\in\mathcal{E}_p} \log(y_i) + \sum_{e_i\in\mathcal{E}_n} \log(1-y_i)\right), \tag{9}$$

where ε_p is the set of positive hyperlinks, ε_n is the set of negative hyperlinks.

4.4 Gap-filling process

4.4.1 Ranking hypothetical reactions

We used the Fritzemeier et al. method [34] to detect EGCs in gap-filled GEMs from wild-type GEMs. The method creates 15 energy dissipation reactions for ATP, CTP, GTP, and other energy metabolites. If a dissipation reaction has non-zero flux, it indicates the presence of an EGC. If the reaction is reversible, we restrict its flux. If it is irreversible, we skip it. We also skip reactions involving oxygen due to anaerobic growth conditions. This process continues until all 200 reactions have been added.

4.4.2 Assessing anaerobic fermentation fluxes

In order to assess anaerobic fermentation fluxes of various metabolites, we employed a method similar to that described in Zimmermann et al. [13] through comparative analysis of GEMs using parsimonious Flux Balance Analysis (pFBA) [46] and Flux-Variability-Analysis (FVA), on both wild-type and gap-filled GEMs. Using pFBA, we optimized for biomass production while avoiding extraneous nutrient influxes, and used pFBA solutions to constrain import fluxes. FVA was employed to predict the potential range of fermentation outcomes and determine the maximum amount of individual fermentation products that could be produced under different FBA solutions.

Metabolites with a secretion flux normalized to biomass greater than 10^{-5} were considered as produced by the GEM. Using the predictions of CLOSEgaps, we classified each fermentation metabolite as produced or not produced, enabling a direct comparison to observed fermentation phenotypes and evaluation of the accuracy and reliability of the GEM in predicting metabolic processes. To assess the performance of CLOSEgaps-filled GEMs, we compared their classification accuracy (AUC, Recall, Precision, and F1 score) with that of wild-type GEMs.

4.4.3 Identifying key reactions in metabolite secretion

Furthermore, we adopt the same strategies described by Chen et al. [28] to identify key reactions in gap-filled GEMs for metabolite secretion using linear mixed-integer programming. This process enables us to determine the minimum set of reactions added during gap-filling responsible for enabling the secretion phenotype. To specifically identify the reactions that enable the secretion of a particular metabolite in a gap-filled GEM but not in its corresponding wild-type GEM, we used LMIP. For each predicted reaction, where each positive hyperlink $e \in \mathcal{E}$ has a corresponding negative hyperlink f, we described its flux activity using a binary variable A, and imposed two linear constraints: *(1)* $f - f_{min}A \geq 0$, and *(2)* $f - f_{max}A \leq 0$, where f represents the flux of the reaction, and f_{min} and f_{max} were set to -1000 and 1000, respectively, to indicate that the reaction has an unconstrained flux ($f \in [-1000, 1000]$). If $A = 1$, the reaction carries a non-zero flux, otherwise it carries zero flux ($f = 0$). We then minimized the sum of all binary indicator variables while ensuring that the secretion flux of the metabolite was positive (with a threshold of 0.1). The minimum number of reactions required to gap-fill the wild-type GEM and produce a particular metabolite can be determined by finding the minimal sum of the binary indicator variables. The identities of the key reactions responsible for this process can be obtained accordingly.

5 Data availability

The raw data is collected from ChEBI (https://www.ebi.ac.uk/), and BiGG (http://bigg.ucsd.edu/). The processed data used to train and test the model is

available at CLOSEgaps [https://github.com/guofei-tju/CLOSEgaps]. Source data are provided in this paper.

6 Code availability

The code for performing the analyses in this manuscript is available at [https://github.com/guofei-tju/CLOSEgaps]. This repository contains a detailed README with instructions on generating the results presented in the manuscript.

Supplementary information.

Appendix A Negative sampling strategies

The CLOSEgaps approach enables the filtering of metabolic missing reactions and improves the performance of metabolic reconstruction. Negative sampling is crucial for machine learning approaches in hyperlink prediction. In this work, we evaluated the performance of the CLOSEgaps gap-filling predictor using different negative sampling strategies with the most recently published Saccharomyces cerevisiae yeast8.5 metabolic network data set.

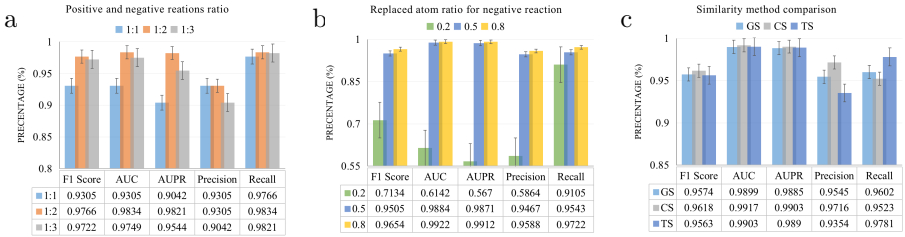
Generating negative sampling strategies is significantly important. We experimented with an alternative negative sampling strategy. For each positive hyperlink $e \in \mathcal{E}$, we generated a corresponding negative hyperlink that balanced the number of atoms. We evaluated the performance of all the methods in predicting missing reactions on artificially introduced gaps in the Saccharomyces cerevisiae yeast8.5 data set using classical classification performance metrics.

An interesting observation is the decreased performance of all methods when using negative sampling considering reactions with a balanced atom number. The F1 scores dropped nearly 4%. This poor performance may be attributed to the narrowing of the selection when metabolites are replaced with ones with the same atom number, leading to the potential repeated selection of some metabolites. Despite the closer proximity of the negative reactions to the truth, the performance still declined. Nevertheless, CLOSEgaps still achieved the best performance among other methods across the majority of metrics. Henceforth, we will refer to the model with the random (imbalanced) negative sampling strategy as CLOSEgaps.

Furthermore, we generate negative reactions for each positive reaction by replacing half of the metabolites involved. The effect of the atom selection ratio is evaluated by changing the percentage of replaced metabolites, with lower percentages indicating that the negative reactions are closer to reality. To assess CLOSEgaps' sensitivity to this negative sampling strategy, we test its performance at 20% and 80% replacement levels. Results show that CLOSEgaps performs well across all evaluation metrics, regardless of the replacement level. When 80% of the metabolites are replaced, the negative

reactions become more random, making it easier for the algorithm to differentiate between fake reactions (as shown in Figure A1b). Conversely, when 20% of the metabolites are replaced, the negative reactions become more similar to the truth, making it more challenging for the algorithm to differentiate. Moreover, we compare the results of different similarity score calculation methods, as depicted in Figure A1c. Although the Consign method achieves the highest F1 score, it is computationally intensive to calculate the similarity for 40,000 metabolites. Hence, we opt for the Gaussian function for further analysis. It is worth noting that CLOSEgaps performs consistently well across different similarity calculation methods.

The performance of the model can be influenced by the ratio of positive and negative reactions. In our previous experiment, the positive reactions were augmented with negative samples in a $1:1$ ratio. Firstly, we evaluate the model's performance with changes in the ratio to $1:2$ and $1:3$. In Figure A1a, the AUC is not affected by the size of the negative sample. On the other hand, the F1 score, precision, and recall show a slight decline as the size of the negative sample increases. Despite potential variations in the negative sampling approach, the stability of CLOSEgaps is maintained and its superiority in terms of performance is upheld.



	F1 Score	AUC	AUPR	Precision	Recall
1:1	0.9305	0.9305	0.9042	0.9305	0.9766
1:2	0.9766	0.9834	0.9821	0.9305	0.9834
1:3	0.9722	0.9749	0.9544	0.9042	0.9821

	F1 Score	AUC	AUPR	Precision	Recall
0.2	0.7134	0.6142	0.567	0.5864	0.9105
0.5	0.9505	0.9884	0.9871	0.9467	0.9543
0.8	0.9654	0.9922	0.9912	0.9588	0.9722

	F1 Score	AUC	AUPR	Precision	Recall
GS	0.9574	0.9899	0.9885	0.9545	0.9602
CS	0.9618	0.9917	0.9903	0.9716	0.9523
TS	0.9563	0.9903	0.989	0.9354	0.9781

Fig. A1: Performance evaluation of CLOSEgaps on yeast8.5 data set using F1 score, AUC, AUPR, Precision, and Recall metrics for different sampling strategies. **a** Sensitivity of CLOSEgaps to positive and negative reaction ratios, with $1:1$, $1:2$, and $1:3$ ratios considered. **b** Effect of replacing 0.2, 0.5, and 0.8 of the metabolites involved in each reaction. **c** Comparison of three similarity calculation methods, GS: Gaussian similarity, CS: Cosine similarity, and TS: Tanimoto similarity.

Appendix B Training strategy

We use the efficient Adam optimization algorithm [47] with a learning rate of 0.01 and weight decay of 5×10^{-4} to train CLOSEgaps. During training, CLOSEgaps minimizes the loss function, thereby maximizing the scores for positive reactions. In testing, CLOSEgaps applies the learned weights to calculate the probability score for unseen reactions in either a testing set or the BiGG data set.

Appendix C Interpretability analysis

In order to evaluate the effectiveness of the proposed CLOSEgaps method compared to traditional graph embedding methods Node2Vec and GCN, we utilized the t-SNE tool [35] to visualize the learned reaction feature embedding over the yeast8.5 data set. The results, as shown in Figure C2, demonstrate that the prediction of our model (CLOSEgaps) is better clustered and separated compared to Node2Vec and GCN, indicating that high-order relation awareness enhances the embedding's ability to capture the structure and semantic proximity among reactions. The visualization of the true label also shows a strong similarity between the prediction and ground truth positions, further supporting our method's improved link prediction performance compared to the original data label.

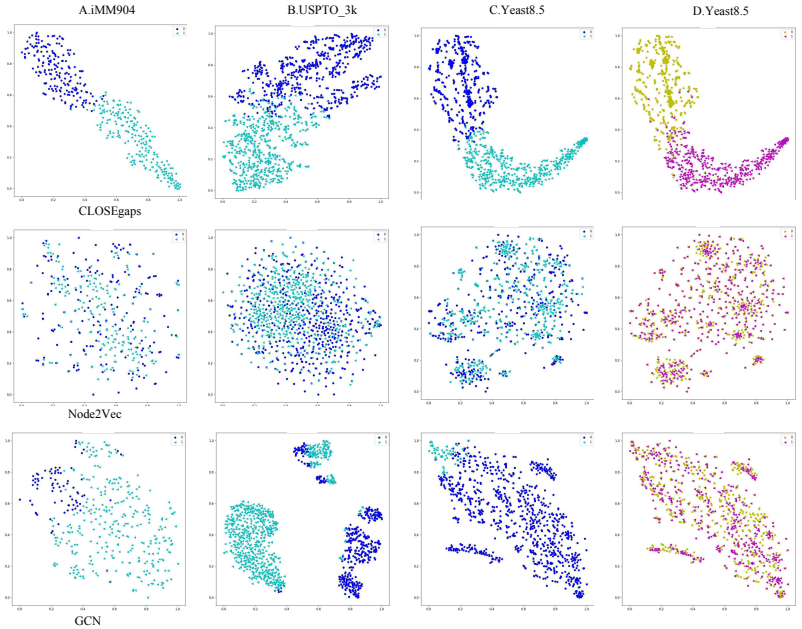


Fig. C2: The 2*D* t-SNE visualization of the latent embeddings for the iMM904, USPTO and Yeast8.5 data sets are presented and compared against two baseline methods, Node2Vec and GCN. The first row depicts the distribution of reactions modeled by CLOSEgaps, represented by blue and green points that indicate predicted labels of 0 and 1, respectively. The true label of the original yeast8.5 data set is shown in Column D, where yellow and pink points correspond to 0 and 1. Rows 2 and 3 display the results from Node2Vec and GCN, respectively.

Appendix D Hyperparameter selection

The key hyperparameters of CLOSEgaps include the encoder feature dimension (denc), the graph convolutional feature dimension (dconv), the number of layers (L), the number of attention mechanism heads (h), and the learning rate (l). These hyperparameters were tuned using a grid search algorithm over the yeast8.5 data set, with searching ranges of $denc = 64, 128, 256$, $dconv = 64, 128, 256$, $L = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, and $l = 0.1, 0.01, 0.001$.

It was observed that even though different GEMs had different optimal hyperparameter sets, the performance using the optimal hyperparameters from one GEM was very similar to that using the optimal hyperparameters from other GEMs. As a result, a universal hyperparameter set was used for all the GEMs to save computational resources. Additionally, CLOSEgaps was found to be not sensitive to the feature dimensions denc and dconv within their searching ranges, and similar performance was achieved with different combinations of these hyperparameters. Hence, denc and dconv were set to 64 and 128, respectively. The attention mechanism head number was set to 6 and the number of graph convolutional layers was set to 3. Finally, the learning rate was set to 0.01 based on the grid search results.

Table D1: Table of the 24 bacterial genomes utilized in phenotypes prediction

NCBI Assembly	Taxonomy
GCF 001561955.1	*Anaerotignum propionicum DSM 1682*
GCF 001456065.2	*Clostridium butyricum KNU-L09*
GCF 000469345.1	*Eubacterium ramulus ATCC 29099*
GCF 000392875.1	*Enterococcus faecalis ATCC 19433*
GCF 000389635.1	*Clostridium pasteurianum BC1*
GCF 000203855.3	*Lactobacillus plantarum WCFS1*
GCF 000175255.2	*Zymomonas mobilis subsp. mobilis ATCC 10988*
GCF 000173975.1	*Anaerobutyricum hallii DSM 3353*
GCF 000162015.1	*Faecalibacterium prausnitzii A2-165*
GCF 000160535.1	*Prevotella bergensis DSM 17361*
GCF 000144405.1	*Prevotella melaninogenica ATCC 25845*
GCF 000143845.1	*Olsenella uli DSM 7084*
GCF 000056065.1	*Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842*
GCF 000025885.1	*Aminobacterium colombiense DSM 12261*
GCF 000022965.1	*Bifidobacterium animalis subsp. lactis DSM 10140*
GCF 000020605.1	*Eubacterium rectale ATCC 33656*
GCF 000020425.1	*Bifidobacterium longum subsp. infantis ATCC 15697*
GCF 000013285.1	*Clostridium perfringens ATCC 13124*
GCF 000011985.1	*Lactobacillus acidophilus NCFM*
GCF 000011065.1	*Bacteroides thetaiotaomicron VPI-5482*
GCF 000008765.1	*Clostridium acetobutylicum ATCC 824*
GCF 000008545.1	*Thermotoga maritima MSB8*
GCF 000008345.1	*Cutibacterium acnes KPA171202*
GCF 000005845.2	*Escherichia coli str. K-12 substr. MG1655*

Appendix E Fermentation test data

A collection of 24 bacterial genomes from a previous study [13] was used to conduct a fermentation product test, as reported in Table D1. The evaluation focused on producing eight metabolites (acetic acid, butyric acid, ethanol, formic acid, lactic acid, butanol, propionic acid, and succinic acid) during the fermentation process. The corresponding GEMs for these assemblies were reconstructed using the CarveMe tool [5], and the growth medium conditions were also included in the collection.

Acknowledgments.

Acknowledgments. This study was supported by the grant from National Key R&D Program of China (2021YFC2100700), National Natural Science Foundation of China (NSFC 62172296, 62172076, 61972280), Excellent Young Scientists Fund in Hunan Province (2022JJ20077), Scientific Research Fund of Hunan Provincial Education Department (22A0007), and also Sponsored by CCF-Tencent Open Fund (NO. IAGR20220109), Zhejiang Lab Open Research Project (NO. K2022PE0AB07), and Shenzhen Science and Technology Program (No.KQTD20200820113106007), and Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020003), and the Municipal Government of Quzhou (Grant No. 2022D006).

Declarations

F.G. conceived and designed the project. X.L. constructed the database, X.L. and H.Y. developed the algorithm. X.L., H.Y., and C.A. trained the reported data with different ML algorithms. All authors analyzed the results. X.L. prepared the manuscript. F.G. edited and approved the manuscript.

Conflict of interest

The authors declare no competing financial interests.

References

[1] Robinson, J. L. & Nielsen, J. Anticancer drug discovery through genome-scale metabolic modeling. *Current Opinion In Systems Biology* **4**, 1–8 (2017) .

[2] Kim, Y., Kim, G. B. & Lee, S. Y. Machine learning applications in genome-scale metabolic modeling. *Current Opinion In Systems Biology* **25**, 42–49 (2021) .

[3] Xu, Y. *et al.* De novo biosynthesis of rubusoside and rebaudiosides in engineered yeasts. *Nat. Commun.* **13** (1), 3040 (2022) .

[4] Vayena, E. *et al.* A workflow for annotating the knowledge gaps in metabolic reconstructions using known and hypothetical reactions. *Proc. Natl. Acad. Sci.* **119** (46), e2211197119 (2022) .

[5] Thiele, I., Vlassis, N. & Fleming, R. M. fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics* **30** (17), 2529–2531 (2014) .

[6] O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using genome-scale models to predict biological capabilities. *Cell* **161** (5), 971–987 (2015) .

[7] Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol.* **20** (1), 1–18 (2019) .

[8] Li, G. *et al.* Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *Nat. Commun.* **12** (1), 190 (2021) .

[9] Li, F. *et al.* Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints. *Nat. Commun.* **13** (1), 2969 (2022) .

[10] Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28** (3), 245–248 (2010) .

[11] Domenzain, I. *et al.* Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using gecko 2.0. *Nat. Commun.* **13** (1), 1–13 (2022) .

[12] Nayfach, S. *et al.* A genomic catalog of earth's microbiomes. *Nat. Biotechnol.* **39** (4), 499–509 (2021) .

[13] Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* **22** (1), 1–35 (2021) .

[14] Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5** (1), 93–121 (2010) .

[15] Pan, S. & Reed, J. L. 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) .

[16] Orth, J. D. & Palsson, B. Ø. Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* **107** (3), 403–412 (2010) .

[17] Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks 855–864 (2016) .

[18] Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks 593–607 (2018) .

[19] Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems* **30** (2017) .

[20] Yadati, N. *et al.* Nhp: Neural hypergraph link prediction 1705–1714 (2020) .

[21] Zhang, M., Cui, Z., Jiang, S. & Chen, Y. Beyond link prediction: Predicting hyperlinks in adjacency space **32** (1) (2018) .

[22] Oftadeh, O. *et al.* A genome-scale metabolic model of saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. *Nat. Commun.* **12** (1), 1–10 (2021) .

[23] Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46** (15), 7542–7553 (2018) .

[24] Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** (1), 31–36 (1988) .

[25] Degtyarenko, K. *et al.* Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36** (suppl_1), D344–D350 (2007) .

[26] Bernstein, D. B., Sulheim, S., Almaas, E. & Segrè, D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biol.* **22** (1), 1–22 (2021) .

[27] Lu, H. *et al.* A consensus s. cerevisiae metabolic model yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10** (1), 1–13 (2019) .

[28] Chen, C., Liao, C. & Liu, Y.-Y. Teasing out missing reactions in genome-scale metabolic networks through deep learning. *bioRxiv* (2022) .

[29] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016) .

[30] Feng, Y., You, H., Zhang, Z., Ji, R. & Gao, Y. Hypergraph neural networks **33** (01), 3558–3565 (2019) .

[31] Norsigian, C. J. *et al.* Bigg models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **48** (D1), D402–D406 (2020) .

[32] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2** (1), 015016 (2021) .

[33] Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* **11** (1), 1–10 (2010) .

[34] Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B. & Lercher, M. J. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS computational biology* **13** (4), e1005494 (2017) .

[35] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9** (11) (2008) .

[36] Levin, I., Liu, M., Voigt, C. A. & Coley, C. W. Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nat. Commun.* **13** (1), 7747 (2022) .

[37] Li, F. *et al.* Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* **5** (8), 662–672 (2022) .

[38] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596** (7873), 583–589 (2021) .

[39] Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537** (7620), 320–327 (2016) .

[40] Huang, B. *et al.* A backbone-centred energy function of neural networks for protein design. *Nature* **602** (7897), 523–528 (2022) .

[41] Li, F., Chen, Y., Anton, M. & Nielsen, J. Gotenzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Res.* **51** (D1), D583–D586 (2023) .

[42] Lowe, D. M. *Extraction of chemical structures and reactions from the literature.* Ph.D. thesis, University of Cambridge (2012).

[43] Ma, Y. & Ma, Y. Hypergraph-based logistic matrix factorization for metabolite–disease interaction prediction. *Bioinformatics* **38** (2), 435–443 (2022) .

[44] Maas, A. L., Hannun, A. Y., Ng, A. Y. *et al.* Rectifier nonlinearities improve neural network acoustic models **30** (1), 3 (2013) .

[45] Bai, S., Zhang, F. & Torr, P. H. Hypergraph convolution and hypergraph attention. *Pattern Recogn.* **110**, 107637 (2021) .

[46] Lewis, N. E. *et al.* Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6** (1), 390 (2010) .

[47] Jais, I. K. M., Ismail, A. R. & Nisa, S. Q. Adam optimization algorithm for wide and deep neural network. *Knowledge Engineering and Data Science* **2** (1), 41–46 (2019) .

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sourcedata.xlsx
- SupplementaryInformation.pdf