

1                   Deep learning-driven automatic  
2                   reconstruction of genome-scale metabolic  
3                   networks

4                   Xiaoyi Liu<sup>1</sup>, Hongpeng Yang<sup>1</sup>, Chengwei Ai<sup>5</sup>, Yijing  
5                   Ding<sup>4\*</sup>, Jijun Tang<sup>1,3\*</sup> and Fei Guo<sup>2\*</sup>

6                   <sup>1</sup>Computational Science and Engineering, University of South  
7                   Carolina, 1244 Blossom Street, Columbia, 29208, South Carolina,  
8                   USA.

9                   <sup>2</sup>Computer Science and Engineering,, Central South University,  
10                   932 Lushan S Rd, Changsha, 410083, Human, China.

11                   <sup>3</sup>Shenzhen Institute of Advanced Technology, Chinese Academy  
12                   of Sciences, 1068 Xueyuan Avenue, Nanshan, 518055, Shenzhen,  
13                   China.

14                   <sup>4</sup>Yangtze Delta Region Institute (Quzhou), University of  
15                   Electronic Science and Technology of China, Quzhou, 324000,  
16                   Zhejiang, China.

17                   <sup>5</sup>School of Computer Science and Technology, Tianjin University,  
18                   No.135, Yaguan Road, Tianjin Haihe Education Park, Tianjin,  
19                   300350, Tianjin, China.

20                   \*Corresponding author(s). E-mail(s): [wuxi\\_dyj@163.com](mailto:wuxi_dyj@163.com);  
21                   [jtang@cse.sc.edu](mailto:jtang@cse.sc.edu); [guofeieileen@163.com](mailto:guofeieileen@163.com);  
22                   Contributing authors: [xiaoyil@email.sc.edu](mailto:xiaoyil@email.sc.edu);  
23                   [hongpeng@email.sc.edu](mailto:hongpeng@email.sc.edu); [ai\\_chery@163.com](mailto:ai_chery@163.com);

## 2 Automated annotating and curating gaps with CLOSEgaps

24

**Supplementary information**

25

**1 Negative sampling strategies**

26

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.

32

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.

39

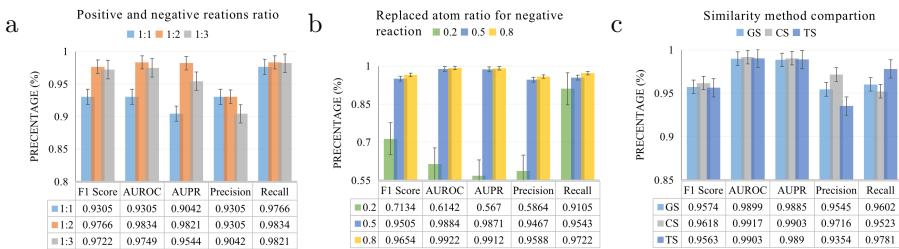
An interesting observation is the decreased performance of all methods when using negative sampling considering reactions with a balanced atom number. The F1 scores dropped nearly 4%. This poor performance may be attributed to the narrowing of the selection when metabolites are replaced with ones with 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.

49

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 1b). 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 the "Metabolites Similarity Measurements" section of Figure 1c. 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.

67

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 1a, 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.



**Fig. 1:** 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.

## 2 Training strategy

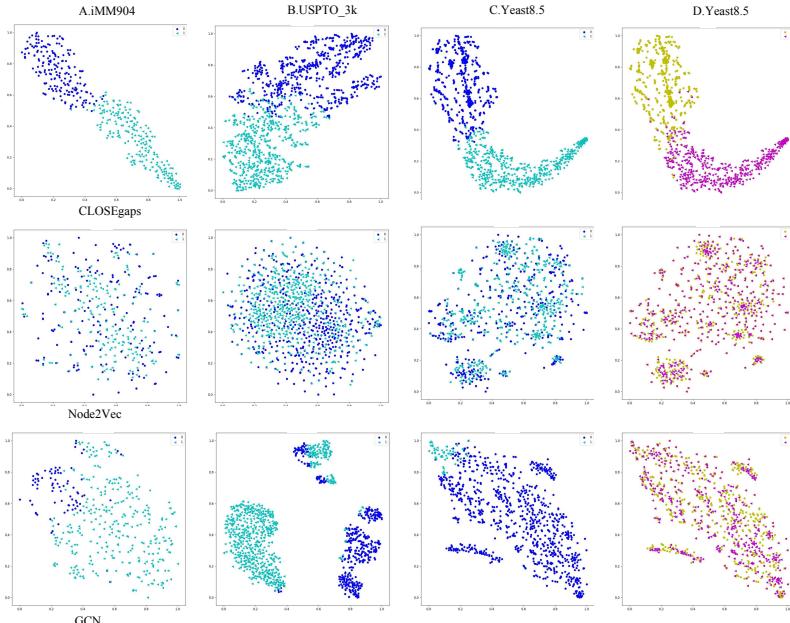
We use the efficient Adam optimization algorithm [1] with a learning rate of 0.01 and weight decay of  $5 \times 10^{-4}$  to train CLOSEgaps. During training, CLOSEgaps minimizes the loss function, thereby maximizing the scores for positive reactions. In testing, CLOSEgaps applies the learned weights to calculate the probability score for unseen reactions in either a testing set or the BiGG data set.

## 3 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 [2] to visualize the learned reaction feature embedding over the yeast8.5 data set. The results, as shown in Figure 2, demonstrate that the prediction of our model (CLOSEgaps) is better clustered and separated

#### 4 Automated annotating and curating gaps with CLOSEgaps

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. 2:** The 2D 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.

## 4 Hyperparameter selection

The key hyperparameters of CLOSEgaps include the encoder feature dimension ( $d_{enc}$ ), the graph convolutional feature dimension ( $d_{conv}$ ), 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  $d_{enc} = 64, 128, 256$ ,  $d_{conv} = 64, 128, 256$ ,  $L = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ , 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 1:** Table of the 24 bacterial genomes utilized in phenotypes prediction

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

## 5 Fermentation test data

A collection of 24 bacterial genomes from a previous study [3] was used to conduct a fermentation product test, as reported in Table 1. 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

## 6      *Automated annotating and curating gaps with CLOSEgaps*

120 reconstructed using the CarveMe tool [4], and the growth medium conditions  
121 were also included in the collection.

## 122     **References**

- 123 [1] Jais, I. K. M., Ismail, A. R. & Nisa, S. Q. Adam optimization algorithm for  
124 wide and deep neural network. *Knowledge Engineering and Data Science*  
125 **2** (1), 41–46 (2019) .
- 126 [2] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach.*  
127 *Learn. Res.* **9** (11) (2008) .
- 128 [3] Zimmermann, J., Kaleta, C. & Waschyna, S. gapseq: informed prediction  
129 of bacterial metabolic pathways and reconstruction of accurate metabolic  
130 models. *Genome Biol.* **22** (1), 1–35 (2021) .
- 131 [4] Thiele, I., Vlassis, N. & Fleming, R. M. fastgapfill: efficient gap filling in  
132 metabolic networks. *Bioinformatics* **30** (17), 2529–2531 (2014) .