Appendix for Heterogeneous Causal Metapath Graph Neural Network for Gene-Microbe-Disease Association Prediction

Kexin Zhang [1], **Feng Huang** [1], **Luotao Liu** [1], **Zhankun Xiong** [1], **Hongyu Zhang** [1,2], **Yuan Quan** [1,2*] and **Wen Zhang** [1,2*]

[1] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

[2] Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China

{kexinzhang, fhuang233, luotaoliu, xiongzk}@webmail.hzau.edu.cn,

{zhy630, quanyuan, zhangwen}@mail.hzau.edu.cn

1 Interpretation of Causal Metapaths

Numerous studies have confirmed causal interaction modes that constitute the GMD triple-wise associations. We believe that the interaction modes contain the key information beneficial for GMD association prediction, which are hence modeled as the causal metapaths. The differences between the causal metapaths and common metapaths are mainly reflected in structure and semantics. The common metapath usually is in a symmetric or palindrome structure (e.g., *A-P-V-P-A*) implying the semantics that describes a high-order relation between the nodes from the same types (e.g., authors (A)), hence the directionality has no significance, while the causal metapath refers to the ordered relation sequence connecting different types of entities (nodes) that have causal interactions with each other and different orderings imply different causal semantics. The descriptions and the examples regarding the causal metapaths derived from causal relations of genes, microbes, and diseases are shown in Table S1.

2 Biological Features of Entities

According to previous research, we calculated the similarity matrices for genes, microbes, and diseases based on knowledge in the biological field to obtain the initial features of nodes. For genes, we employed the method provided by [Yang *et al.*, 2021], utilizing the overlap of Gene Ontology (GO) annotations between genes to compute gene functional similarity matrix. For microbes and diseases, we respectively adopted the approaches proposed by [Ma and Jiang, 2020] and [Wang *et al.*, 2010] to calculate microbe taxonomic similarity matrix and disease semantic similarity matrix by gathering semantic information from the NCBI Taxonomy database and the Medical Subject Heading (MeSH) database.

3 Implementation of the baseline methods

All the baseline methods were implemented using their publicly available source codes, adopting either their best-performing or default parameters. Hyper-parameter settings for mTransH, RAM, HAN, MAGNN, and HypergraphSynergy remained consistent with the provided in original papers. For the hyper-parameters of RF, MLP, XGBoost, Tucker, CP, and NeurTN were carefully tuned to achieve optimal performance. Notably, HAN and MAGNN are metapath-based heterogeneous graph neural network models used for tasks such as node clustering and pairwise association prediction. HAN learns node embeddings from different metapath-based graphs that contains only one node type and can be regarded as an undirected homogeneous graph, where the intermediate nodes along the metapath are ignored. In order to make it suitable for our triple-wise association prediction task, we modified the metapath types to *G-M-G*, *G-D-G*, *M-G-M*, *M-D-M*, *D-G-D*, and *D-M-D* according to the format of the original paper (the length and number of the metapaths are consistent with the metapaths we utilized). Similar to it, MAGNN also learns the contextual semantic information of the target node by defining metapaths of symmetric structure, so we employed the same method to implement the triple-wise association prediction task.

We run HCMGNN and other baselines on our workstation with 48 Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz.

4 Hyper-parameter Sensitivity Analysis

In this section, we conducted hyper-parameter sensitivity analysis in 5-fold CV to investigate the impact of several hyperparameters on the performance of HCMGNN. We chose the learning rate lr, the hidden layer dimension d and the balance coefficient γ in the loss function for analysis.

4.1 Effect of learning rate lr.

We chose lr from the range of {0.0001, 0.001, 0.0025, 0.005, 0.0075}. The results in Figure S1 (a) demonstrate that HCMGNN achieves the best performance when lr is set to 0.005. Therefore, we selected 0.005 as the optimal parameter value for the learning rate.

4.2 Effect of hidden layer dimension d.

We selected the hidden layer dimension d from the set {8, 16, 32, 64, 128}. Figure S1 (b) illustrates the results, indicating that HCMGNN attains optimal performance when d is set to 64. Consequently, 64 was chosen as the optimal value for the hidden layer dimension.

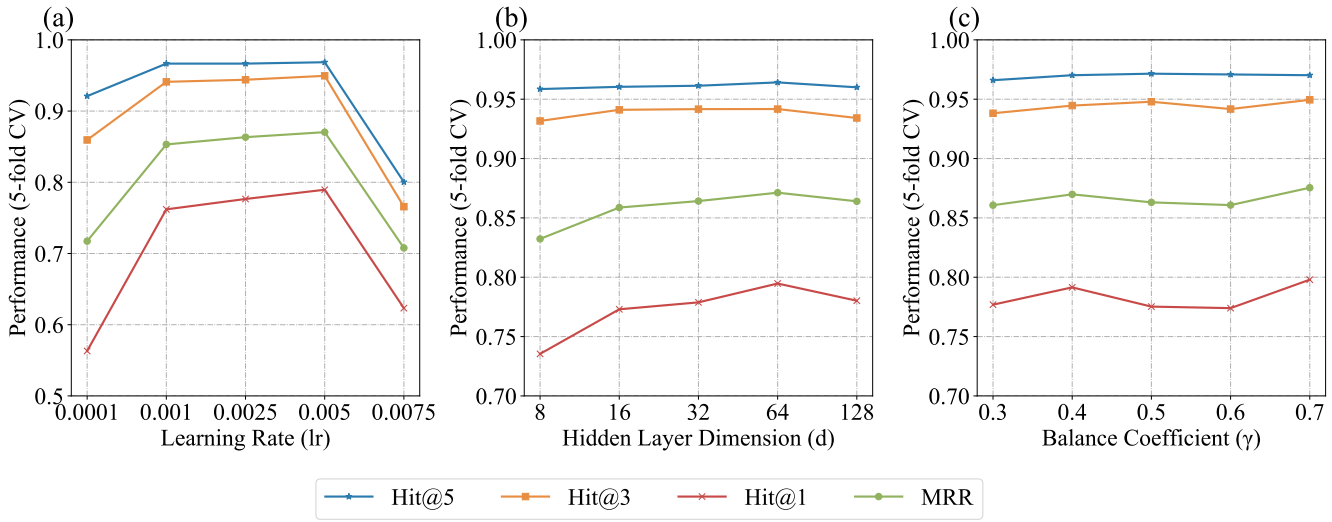

Figure S1: The performance of HCMGNN on 5-fold CV with different hyper-parameters. (a) The impact of the learning rate lr. (b) The impact of the hidden layer dimension d. (c) The impact of the loss balance coefficient γ.

Type	Description	Example
G-M-D	Genes regulate microbial changes to affect complex diseases	*NOD2-Enterobacteriaceae*-IBD
G-D-M	Genes cause changes in disease, which in turn alter the microbes	*SORL1*-AD-*Dialister*
D-M-G	Diseases affect the host gene function by regulating microbes	CRC-*Blautia-STEAP2*
D-G-M	Diseases affect changes in microbes by modulating gene function	RA-*PHRF1-Oxalobacteraceae*
M-D-G	Microbial influences on diseases lead to changes in the function of genes	*Christensenellaceae*-DR-*SHB*
M-G-D	Microbes influence complex disease by modulating gene function	*Morganella-PDE1A*-MDD

Table S1: Descriptions of the six predefined causal metapaths. Inflammatory Bowel Disease is abbreviated as IBD, Alzheimer's Disease is abbreviated as AD, colorectal cancer is abbreviated as CRC, Rheumatoid Arthritis is abbreviated as RA, Diabetic Retinopathy is abbreviated as DR, Major Depressive Disorder is abbreviated as MDD. The evidence of the examples of six causal interactions is as follows: [Luca *et al.*, 2018], [Hughes *et al.*, 2020], [Ni *et al.*, 2022], [Kurilshikov *et al.*, 2021], [Liu *et al.*, 2022] and [Qin *et al.*, 2022].

4.3 Effect of balance coefficient γ.

The balance coefficient γ was searched in the range of $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. We can observe from Figure S1 (c) that HCMGNN gets the optimal performance when γ is set to 0.7.

5 Performance of Independent Test Set

The independent test performances of HCMGNN and baselines evaluated by Hit@n, NDCG@n, and MRR are shown in Table S2.

6 Additional details in Section 4.4

We conducted a statistical analysis of node degrees in the heterogeneous graph and averaged the degrees for the nodes contained in each triplet to obtain the average node degree (N) for each triplet in the independent test set. We found that the triplet numbers under different N distributionis only 1 in (0,10], 3 in [10,20], 4 in [20,30], and 6 in [30,40]. Considering that too few triplets may cast doubt on the reliability of the results, we decided to choose triplets under the N distribution (start with 40 and end with 140) where [N-10,N] has more than 10 triplets, so we divided the independent test set into 12 groups. As shown in Figure 4, the yellow bars represent numbers of triplets located at [N-10,N], and gray bars represent numbers of triplets located at (0,N]. We can observe that the relation between N and the number of triplets is similar to normal distribution, where the closer N is to the minimum and maximum, the smaller the number of triplets. To highlight the advantages of HCMGNN over other baseline methods in handling low-degree triples in graph data, we compared HCMGNN with several graph neural network baselines, including HAN, MAGNN, and HypergraphSynergy. As shown in Figure S2, HCMGNN consistently outperforms these baselines across 12 groups of test groups (interval (0, N]) on low-degree samples. This further underscores HCMGNN's advantage in consolidating sparse node information, which helps to mitigate the problem of association sparsity.

In addition, we divided the independent test set into (0,40], [41,80], [81,120], and [121, $+\infty$) four groups by setting the interval of average node degrees to 40. Next, we evaluate the performances of HCMGNN and HCMGNN without the intra-subgraph sharing message passing (w/o MP) by Hit@1 on the test set in the same group. We observe that the results in Figure S1 are consistent with those in Figure 4, HCMGNN has superior predictive performance in low-degree triplets (e.g., the groups of [0,40] and [41,80]), which again demonstrates

Type	Method	Hit@1	Hit@3	Hit@5	NDCG@1	NDCG@3	NDCG@5	MRR
Machine Learning	RF	0.5306	0.8309	0.9038	0.5306	0.7082	0.7386	0.6943
	MLP	0.5714	0.8571	0.9417	0.5714	0.7395	0.7744	0.7252
	XGBoost	0.5918	0.8688	0.9242	0.5918	0.7555	0.7789	0.7389
Tensor Decomposition	Tucker	0.8192	0.8542	0.8571	0.8192	0.8398	0.8409	0.8435
	CP	0.8047	0.8601	0.8746	0.8047	0.8381	0.8440	0.8430
	NeurTN	0.6181	0.8746	0.9417	0.6181	0.7693	0.7972	0.7540
Knowledge Graph Embedding	mTransH	0.7055	0.8426	0.8980	0.7055	0.7859	0.8085	0.7871
	RAM	0.7425	0.8552	0.8980	0.7425	0.8091	0.8270	0.8121
Graph Neural Network	HAN	0.7230	0.8805	0.9359	0.7230	0.8166	0.8392	0.8128
	MAGNN	0.7726	0.8805	0.9413	0.7726	0.8353	0.8520	0.8390
	HypergraphSynergy	0.8251	0.9475	0.9650	0.8251	0.9000	0.9069	0.8913
	HCMGNN	**0.8455**	**0.9650**	**0.9796**	**0.8455**	**0.9178**	**0.9239**	**0.9073**

Table S2: Performances of HCMGNN and baselines evaluated by Hit@n, NDCG@n, and MRR in independent test set.

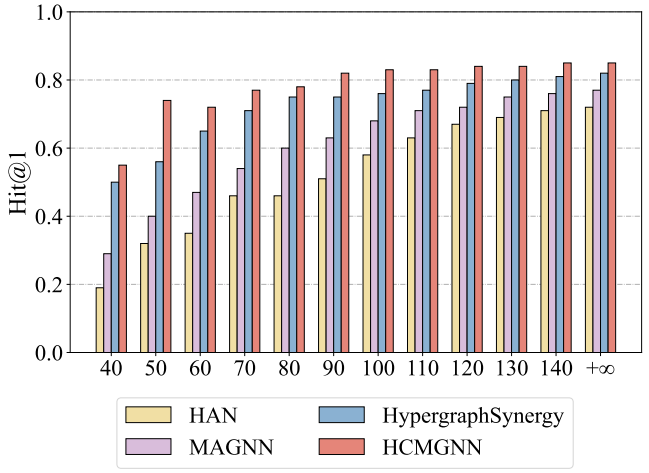

Figure S2: Performances of HCMGNN and baselines in test triplets with different average node degree (N) distributions.

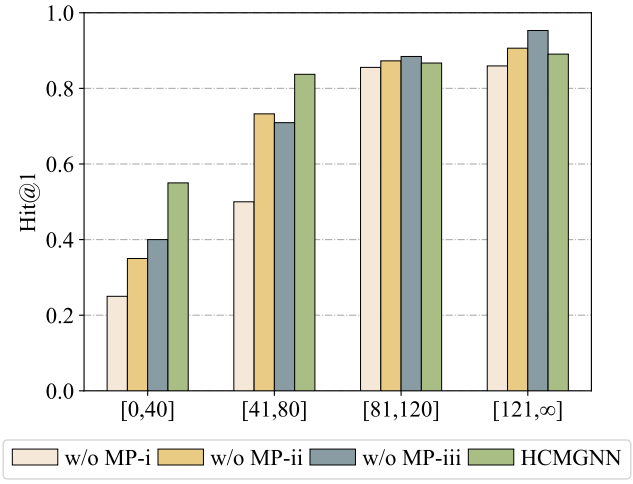

Figure S3: Performances of HCMGNN and w/o MP in test triplets with different average node degree (N) distributions.

that HCMGNN alleviates the association sparsity issue in triple-wise association prediction. Furthermore, HCMGNN exhibits a slight performance drop in the groups of [81,120] and [121,+∞), which may be due to the information overload caused by the semantic sharing strategy on non-sparse nodes. However, the results do not affect our concluding that HCMGNN has the advantage of dealing with sparse association, as we can see from Figure 4 that the performance of HCMGNN increases with the number of triplets.

7 Attention Coefficient Analysis of Causal Metapaths

In order to intuitively observe the change of the attention coefficients of the causal metapaths during the model iteration process, we retained the attention coefficients for every 10 epochs and the optimal epoch determined by early stopping mechanism, and obtained the heat map visualization result. As shown in Figure S4, it can be observed that HCMGNN initially cannot determine the relative importance of each metapath, but with the training of the model, the attention coefficients gradually tilt toward certain metapaths (such as *G-M-D*). The highest attention coefficient of the *G-M-D* allows us to speculate that the causal relation of genes regulate microbial changes to affect complex diseases may play a non-negligible role in GMD association prediction. Evidence in support of this idea is the finding by [Hall *et al.*, 2017] that some host genetic variants make the microbiota dysbiosis, leading to metabolic and immune diseases (such as inflammatory bowel disease). In addition, [Quan *et al.*, 2023] found that gut microbes may be an important mediator to link diseases with the evolution of human genome. Therefore, by analyzing the importance of different metapaths, it will help us to further understand the causal relations of GMD associations.

References

[Hall *et al.*, 2017] Andrew Brantley Hall, Andrew C Tolonen, and Ramnik J Xavier. Human genetic variation and the gut microbiome in disease. *Nature Reviews Genetics*, 18(11):690–699, 2017.

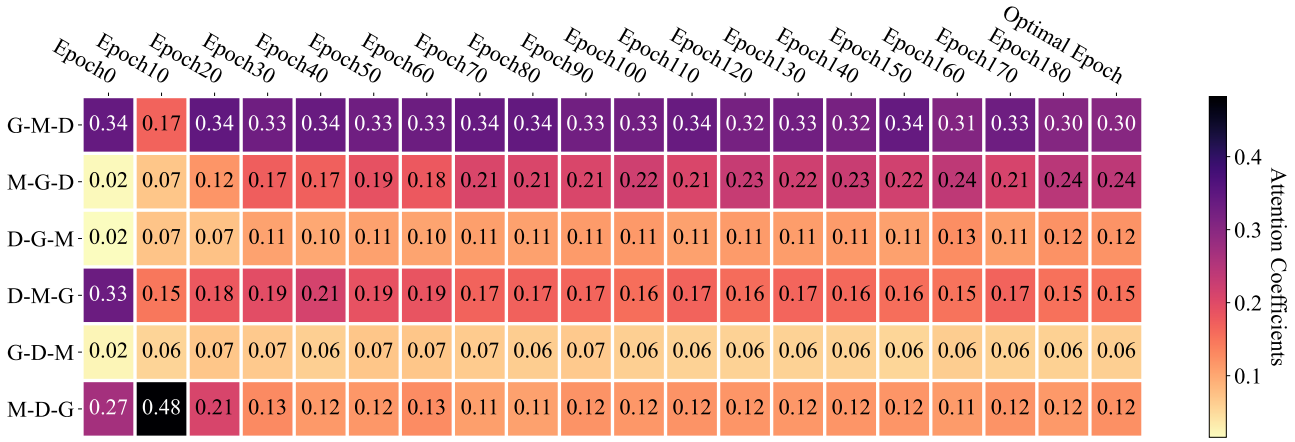

Figure S4: Visualization of attention coefficients for causal metapaths.

[Hughes *et al.*, 2020] David A Hughes, Rodrigo Bacigalupe, Jun Wang, Malte C Rühlemann, Raul Y Tito, Gwen Falony, Marie Joossens, Sara Vieira-Silva, Liesbet Henckaerts, Leen Rymenans, et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature microbiology*, 5(9):1079–1087, 2020.

[Kurilshikov *et al.*, 2021] Alexander Kurilshikov, Carolina Medina-Gomez, Rodrigo Bacigalupe, Djawad Radjabzadeh, Jun Wang, Ayse Demirkan, Caroline I Le Roy, Juan Antonio Raygoza Garay, Casey T Finnicum, Xingrong Liu, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature genetics*, 53(2):156–165, 2021.

[Liu *et al.*, 2022] Kangcheng Liu, Jing Zou, Huimin Fan, Hanying Hu, and Zhipeng You. Causal effects of gut microbiota on diabetic retinopathy: A mendelian randomization study. *Frontiers in Immunology*, 13:930318, 2022.

[Luca *et al.*, 2018] Francesca Luca, Sonia S Kupfer, Dan Knights, Alexander Khoruts, and Ran Blekhman. Functional genomics of host–microbiome interactions in humans. *Trends in Genetics*, 34(1):30–40, 2018.

[Ma and Jiang, 2020] Yuanjing Ma and Hongmei Jiang. Ninimhmda: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe–disease association. *Bioinformatics*, 36(24):5665–5671, 2020.

[Ni *et al.*, 2022] Jing-Jing Ni, Xiao-Song Li, Hong Zhang, Qian Xu, Xin-Tong Wei, Gui-Juan Feng, Min Zhao, Zi-Jia Zhang, Lei Zhang, Gen-Hai Shen, et al. Mendelian randomization study of causal link from gut microbiota to colorectal cancer. *BMC cancer*, 22(1):1371, 2022.

[Qin *et al.*, 2022] Youwen Qin, Aki S Havulinna, Yang Liu, Pekka Jousilahti, Scott C Ritchie, Alex Tokolyi, Jon G Sanders, Liisa Valsta, Marta Brożyńska, Qiyun Zhu, et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nature genetics*, 54(2):134–142, 2022.

[Quan *et al.*, 2023] Yuan Quan, Ke-Xin Zhang, and Hong-Yu Zhang. The gut microbiota links disease to human genome evolution. *Trends in Genetics*, 39(6):451–461, 2023.

[Wang *et al.*, 2010] Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics*, 26(13):1644–1650, 2010.

[Yang *et al.*, 2021] Hongpeng Yang, Yijie Ding, Jijun Tang, and Fei Guo. Identifying potential association on gene-disease network via dual hypergraph regularized least squares. *BMC genomics*, 22:1–16, 2021.