Supplementary Material of Revealing Herb-Symptom Associations and Mechanisms of Action in Protein Networks Using Subgraph Matching Learning

Menglu Li#, Yongkang Wang#, Yujing Ni, Hui Xiong, Zhinan Mei*, Wen Zhang*

In this supplementary material, we first analyze the density distribution of 1-hop protein neighbors for herbs and symptoms (Fig S1), and then discuss the sensitivity of parameters of GraphHSA, including the number of GCN layers L, the coefficients λ_{CL_1}, λ_{CL_2}, and λ_{Pred}, and the hop H. The optimal parameter values are determined sequentially, aligning with the order of their analysis. When analyzing the influence of a certain parameter, the remaining parameters are fixed to default values (the default values $L = 2$, $\lambda_{\text{CL}_1} = \lambda_{\text{CL}_2} = \lambda_{\text{Pred}} = 1.0$, and $H = 1$). If the parameter has been analyzed, the default value is equal to its optimal value. Additionally, we analyze the high-attention targets on Huo Xiang Zheng Qi Liquid. Finally, we provide the results of GraphHSA and baselines on the cross-validation set (Table S1), the results of GraphHSA and its variants in the scenarios of the warm start (Fig S6), the cold start for herb (Fig S7) and the cold start for symptom (Fig S8).

I. Analysis of the Density Distribution of Protein Neighbors for Herbs and Symptoms

Due to different distributions of protein neighbors for herbs and symptoms, we uniformly sample a fixed-size subset in our study. Fig S1 displays the density distribution of 1-hop protein neighbors for herbs and symptoms, i.e., the size of $\mathcal{N}^1(h)$ and $\mathcal{N}^1(s)$. Most $\mathcal{N}^1(s)$ are less than 200, while $\mathcal{N}^1(h)$ ranges from 0 to 1,500. In addition, we utilize a two-sided Mann-Whitney U test to assess the statistical significance between the size of $\mathcal{N}^1(h)$ and $\mathcal{N}^1(s)$ (p-value=1.5046e-05). A two-sided p-value of <0.05 was considered statistically significant [1].

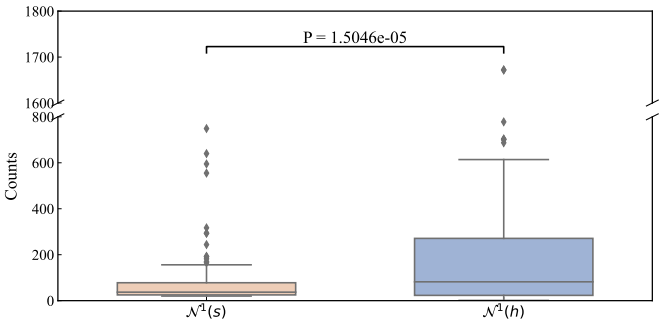

Fig. S1. The density distribution of 1-hop protein neighbors for herbs and symptoms. The p-value is calculated using the two-sided Mann-Whitney U test.

II. Parameter Analysis

We first analyze the effect of the number of GCN layers L, as illustrated in Fig. S2. We vary L within the range {1, 2, 3, 4}. It can be observed that both AUC and AUPR exhibit slight fluctuations as L increases, indicating that increasing the number of GCN layers has a minor impact on performance in the scenario of the cold start for herb/symptom. When L increases, both metrics show a noticeable decline in the scenario of the cold start for pair. This suggests that deeper GCN layers may not effectively capture useful information in the scenario of the cold start for pair and could even introduce noise, leading to performance degradation. In this study, we adopt $L = 1$ to analyze the effect of other parameters.

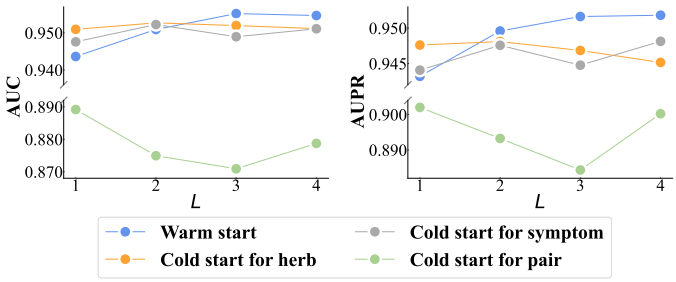

Fig. S2. Results of GraphHSA with different values of the number of GCN layers L on the cross-validation set.

Then, we investigate the influence of coefficients λ_{CL_1}, λ_{CL_2}, and λ_{Pred} by varying these coefficients in the range of $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. This analysis allows us to delve deeper into the significance of self-supervised contrastive loss, supervised contrastive loss, and PPI- and HSA-based classification losses. As shown in Fig. S3, GraphHSA achieves optimal performance when $\lambda_{\text{CL}_1} = 1.0$, $\lambda_{\text{CL}_2} = 0.8$, and $\lambda_{\text{Pred}} = 0.6$, indicating that: (1) Introducing a dual-contrastive learning strategy to maximize agreement between different graphs of the same herb-symptom pairs is crucial for accurate prediction of HSAs. (2) Introducing supervised contrastive loss

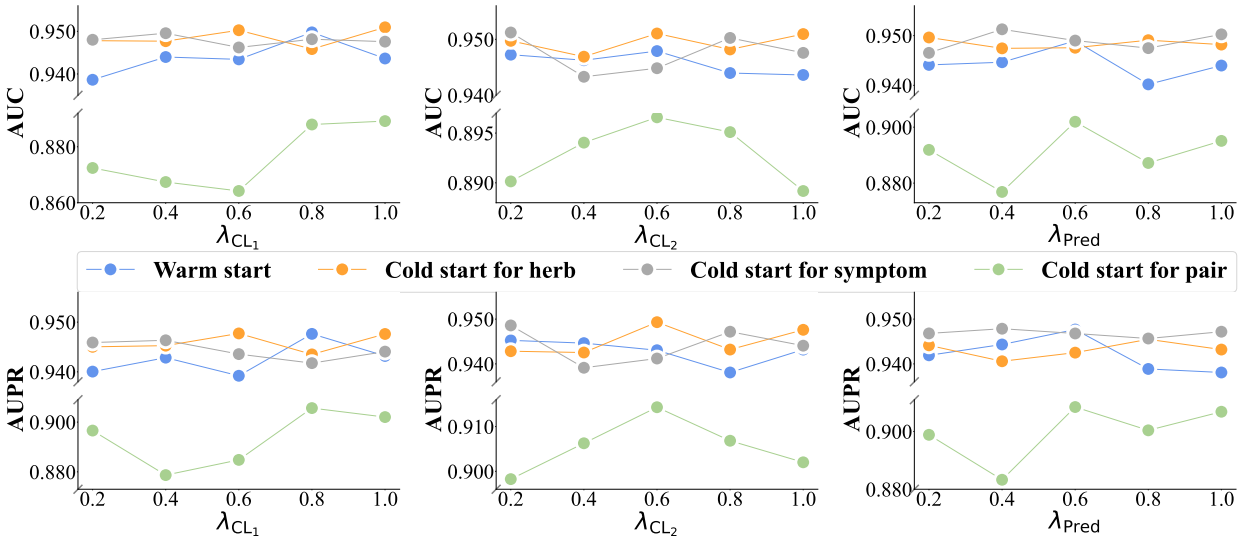

Fig. S3. Results of GraphHSA with different values of the coefficients λ_{CL_1}, λ_{CL_2}, and λ_{Pred} on the cross-validation set.

to learn more discriminative representations of herb-symptom pairs is helpful for HSA prediction. (3) Considering both PPI- and HSA-based classification tasks benefits HSA prediction.

Further, we assess the effect of the hop H by varying it in the range of {1, 2, 3, 4} (Fig. S4). The results show that H affects model performance differently under various settings. The AUC and AUPR remain stable as H increases, indicating that increasing the number of protein neighbors has a minor impact on performance in the scenario of the cold start for herb/symptom. The AUC and AUPR decline when H increases in the scenario of the warm start, while they increase as H increases and decrease after reaching 2 in the scenario of the cold start for pair. This suggests that expanding the neighborhood size may introduce noise and reduce the model's ability to capture meaningful information.

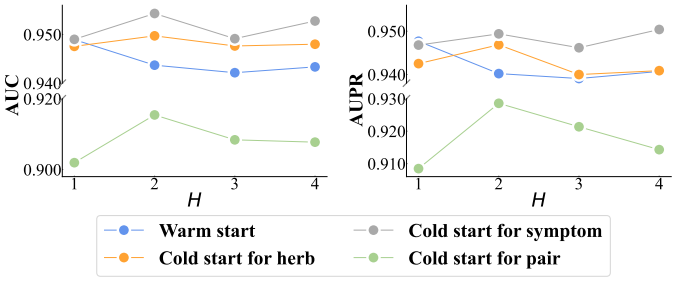

Fig. S4. Results of GraphHSA with different values of the hop H on the cross-validation set.

III. GraphHSA reveals HSAs and mechanisms of action

To further test the validity of predicted associations within the specified herb and symptom terms, we additionally select the well-known TCM compound prescription, 'Huo Xiang Zheng Qi Liquid' [2], as an example. This prescription is composed of Cangzhu (Atractylodis Rhizoma), Chengpi (Tangerine Peel), Houpu (Officinal Magnolia Bark), Baizhi (Dahurian Angelica Root), Fulin (Poria), Dafupi (Areca Peel), Banxia (Pinelliae Rhizoma), Gancao (Licorice), Guanghuoxiangyou (Patchouli Oil), and Zisuyeyou (Perilla Oil). Specifically, we predict symptom associations for each of these herbs, focusing on results with a prediction confidence exceeding 0.90. This analysis yields 16 high-confidence symptoms, including abdominal pain, cough, defecation pain, diarrhea, dizziness, eczema, edema, emesis, fever, hoarseness, hyperalgia, jaundice, pain, pneumonia, skin pruritus, and syncope [3]. Notably, three symptom types have been identified as correlating with the drug's reported clinical efficacy, which emphasizes the robustness and clinical validity of our method in predicting novel herb-symptom associations.

Herein, we select these herbs associated with the symptom of diarrhea as examples and filter genes with attention scores exceeding $1e-5$, resulting in the identification of 135 functional genes. Subsequently, we perform biological enrichment analysis using GO and KEGG for these genes. As illustrated in Fig. S5, we identify 10 enrichment terms with statistical significance for the biological process category. Notably, the terms 'defense response to bacterium', 'apoptotic process', and 'positive regulation of canonical NF-kappaB signal transduction' are closely associated with the therapeutic mechanisms underlying diarrhea treatment [4], [5]. Meanwhile, the key enriched pathways, including the 'Pathogenic Escherichia coli infection', 'Salmonella infection', and 'PI3K-Akt signaling pathway' [6], [7], along with core targets influencing these pathways (e.g., CASP8, FOS, and MAPK1) [8]–[10], have been extensively utilized in the clinical development of drugs for eczema.

References

[1] C. L. Kienbacher, W. Schreiber, H. Herkner, C. Holzhacker, C. C. Chwojka, K. Tscherny, A. Egger, V. Fuhrmann, M. Niederer, M. Neymayer *et al.*, "Drone-facilitated real-time video-guided feedback helps to improve the quality of lay bystander basic life support. a randomized

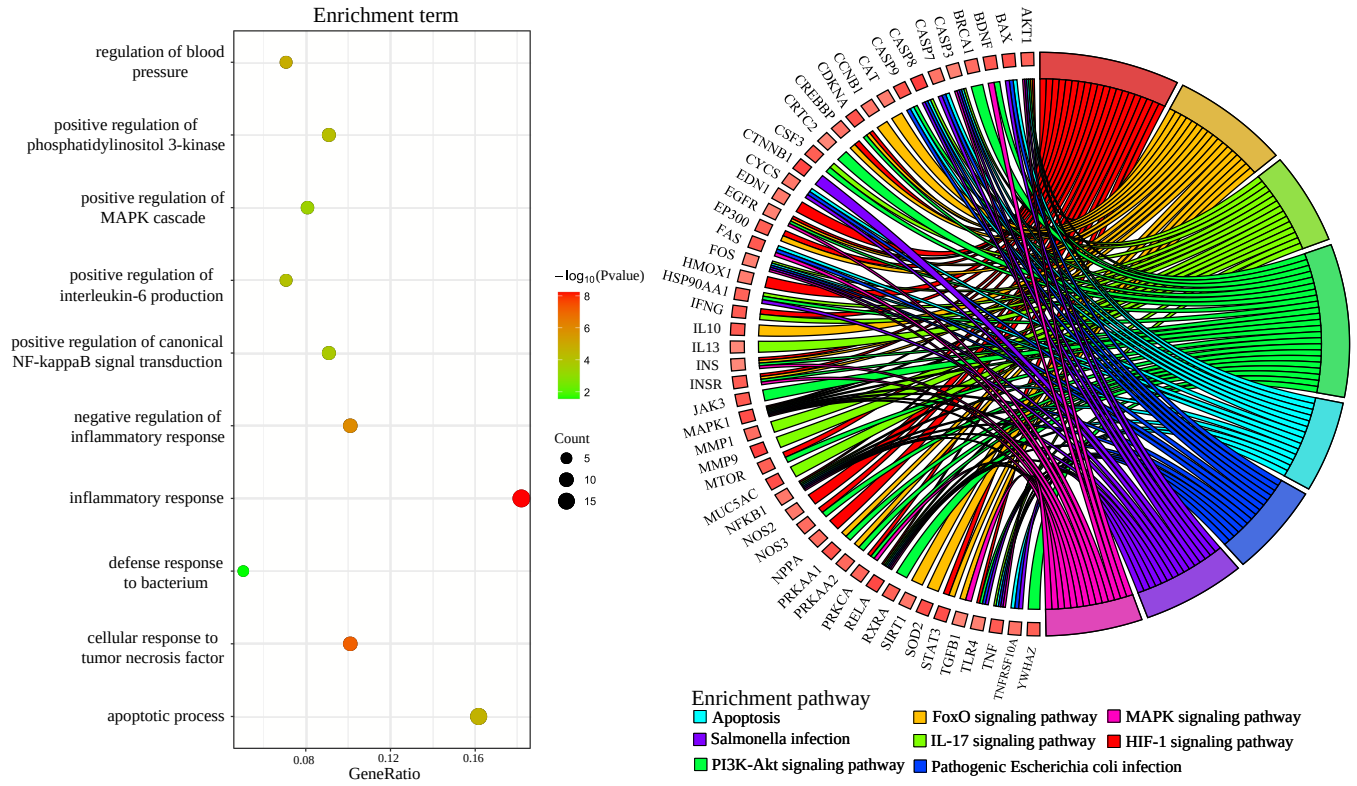

Fig. S5. Enrichment analysis for targets with high attention scores. Dot plot (left) of GO enrichment analysis (top 10 terms) for the biological processes. Bar plot (right) of the top 8 signaling pathways based on KEGG enrichment analysis.

TABLE S1
RESULTS OF GRAPHHSA AND BASELINES ON THE CROSS-VALIDATION SET

Methods	Warm start		Cold start for herb		Cold start for symptom		Cold start for pair	
	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC
Node2Vec	0.6647	0.6827	0.6665	0.6876	0.6733	0.6895	0.6999	0.7040
DeepWalk	0.6724	0.6886	0.6637	0.6878	0.6602	0.6864	0.7328	0.7212
LINE	0.6258	0.6237	0.6234	0.6253	0.6256	0.6271	0.6651	0.6448
SDNE	0.6780	0.7067	0.6673	0.7014	0.6635	0.7033	0.7064	0.7253
HIT_z	0.7083	0.7293	0.6989	0.7260	0.7022	0.7252	0.7392	0.7410
MLP-HSA	0.9100	0.9007	0.9180	0.9149	0.9156	0.9124	0.8329	0.8010
GraphSynergy	0.9362	0.9330	0.9373	0.9363	0.9382	0.9380	0.8786	0.8583
GraphHSA	**0.9477**	**0.9489**	**0.9425**	**0.9476**	**0.9468**	**0.9490**	**0.9085**	**0.9019**

Note: The bold in each column represents the highest score and the underlined denotes the second-best score.

controlled simulation trial," *Prehosp. Emerg. Care*, vol. 29, no. 1, pp. 46–52, 2025.

[2] Q.-X. Jiang, C. Yi-Meng, M. Jing-Jie, W. Yu-Ping, L. Ping, W. Xiao-Dong, and Y. Jie, "Effective fraction from simiao wan prevents hepatic insulin resistant by inhibition of lipolysis via ampk activation," *Chin. J. Nat. Med.*, vol. 20, no. 3, pp. 161–176, 2022.

[3] M. Xiao, J. Tian, Y. Zhou, X. Xu, X. Min, Y. Lv, M. Peng, Y. Zhang, D. Yan, S. Lang *et al.*, "Efficacy of huoxiang zhengqi dropping pills and lianhua qingwen granules in treatment of covid-19: a randomized controlled trial," *Pharmacol. Res.*, vol. 161, p. 105126, 2020.

[4] X. Shen, L. Yin, X. Pan, R. Zhao, and D. Zhang, "Porcine epidemic diarrhea virus infection blocks cell cycle and induces apoptosis in pig intestinal epithelial cells," *Microb. Pathog.*, vol. 147, p. 104378, 2020.

[5] Z. Yan, K. Zhang, K. Zhang, G. Wang, L. Wang, J. Zhang, Z. Qiu, Z. Guo, Y. Kang, X. Song *et al.*, "Huang bai jian pi decoction alleviates diarrhea and represses inflammatory injury via pi3k/akt/nf-κb pathway: In vivo and in vitro studies," *J. Ethnopharmacol.*, vol. 292, p. 115212, 2022.

[6] J. Meng, Y. Li, M. J. Fischer, M. Steinhoff, W. Chen, and J. Wang, "Th2 modulation of transient receptor potential channels: an unmet therapeutic intervention for atopic dermatitis," *Front. Immunol.*, vol. 12, p. 696784, 2021.

[7] T. Liu, S. Li, S. Ying, S. Tang, Y. Ding, Y. Li, J. Qiao, and H. Fang, "The il-23/il-17 pathway in inflammatory skin diseases: from bench to bedside," *Front. Immunol.*, vol. 11, p. 594735, 2020.

[8] S. E. Ledwaba, D. V. Costa, D. T. Bolick, N. Giallourou, P. H. Medeiros, J. R. Swann, A. N. Traore, N. Potgieter, J. P. Nataro, and R. L. Guerrant, "Enteropathogenic escherichia coli infection induces diarrhea, intestinal damage, metabolic alterations, and increased intestinal permeability in a murine model," *Front. Cell. Infect. Microbiol.*, vol. 10, p. 595266, 2020.

[9] G. L. Popa and M. I. Papa, "Salmonella spp. infection-a continuous threat worldwide," *Germs*, vol. 11, no. 1, p. 88, 2021.

[10] H. Lin, B. Li, M. Liu, H. Zhou, K. He, and H. Fan, "Nonstructural protein 6 of porcine epidemic diarrhea virus induces autophagy to promote viral replication via the pi3k/akt/mtor axis," *Vet. Microbiol.*, vol. 244, p. 108684, 2020.

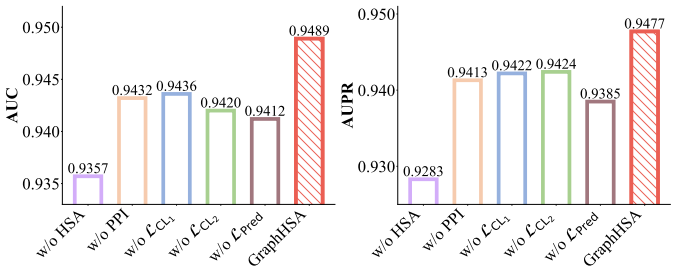

Fig. S6. Results of GraphHSA and its variants in the scenario of the warm start.

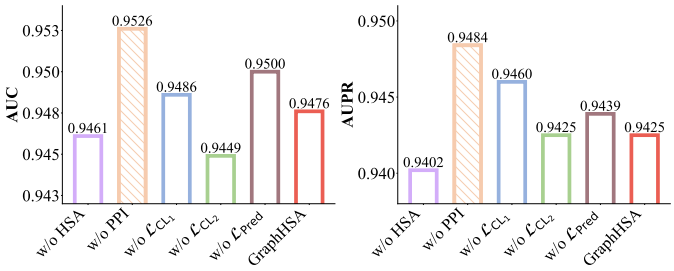

Fig. S7. Results of GraphHSA and its variants in the scenario of the cold start for herb.

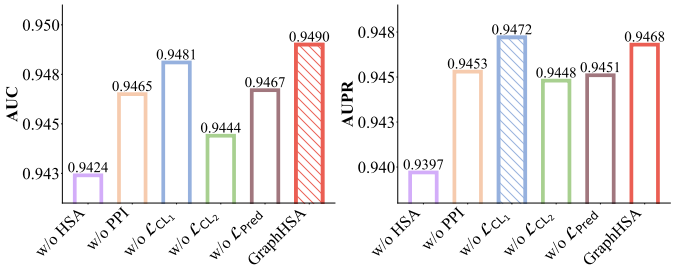

Fig. S8. Results of GraphHSA and its variants in the scenario of the cold start for symptom.