RESEARCH ARTICLE

Open Access

Identification of microbe–disease signed associations via multi-scale variational graph autoencoder based on signed message propagation

Huan Zhu[1], Hongxia Hao[1*] and Liang Yu[1*]

Abstract

Background Plenty of clinical and biomedical research has unequivocally highlighted the tremendous significance of the human microbiome in relation to human health. Identifying microbes associated with diseases is crucial for early disease diagnosis and advancing precision medicine.

Results Considering that the information about changes in microbial quantities under fine-grained disease states helps to enhance a comprehensive understanding of the overall data distribution, this study introduces MSignVGAE, a framework for predicting microbe-disease sign associations using signed message propagation. MSignVGAE employs a graph variational autoencoder to model noisy signed association data and extends the multi-scale concept to enhance representation capabilities. A novel strategy for propagating signed message in signed networks addresses heterogeneity and consistency among nodes connected by signed edges. Additionally, we utilize the idea of denoising autoencoder to handle the noise in similarity feature information, which helps overcome biases in the fused similarity data. MSignVGAE represents microbe-disease associations as a heterogeneous graph using similarity information as node features. The multi-class classifier XGBoost is utilized to predict sign associations between diseases and microbes.

Conclusions MSignVGAE achieves AUROC and AUPR values of 0.9742 and 0.9601, respectively. Case studies on three diseases demonstrate that MSignVGAE can effectively capture a comprehensive distribution of associations by leveraging signed information.

Keywords Variational graph autoencoder, Microbe-disease association, Signed message propagation, XGBoost

Background

Microbes are a class of microorganisms that typically exist as single cells or cell colonies [1]. Accumulated research has shown that microbial communities primarily consist of viruses, archaea, bacteria, and protozoa, and they have close interactions with human hosts [2, 3]. The majority of commensal microorganisms in humans are harmless and even have mutually beneficial relationships with their human hosts. Human microbiota can resist pathogen invasion, promote nutrient absorption, and enhance metabolic capabilities [4]. For example,

*Correspondence:
Hongxia Hao
hxhao@xidian.edu.cn
Liang Yu
lyu@xidian.edu.cn
[1] School of Computer Science and Technology, Xidian University, Xi'an, China

probiotics can stimulate the host's immune system by producing immune-modulatory signals, enhancing protection against pathogens [5]. Therefore, ecological imbalance or dysbiosis of microbial communities can lead to human diseases [6–12]. Furthermore, studies have shown that microbial metabolism can significantly influence clinical responses to drugs, and drug administration can also have specific effects on microbial communities [3, 13, 14]. For example, Panebianco et al. reported interactions between gut microbiota and anticancer drugs, affecting drug efficacy and side effects [15]. Maier et al. discovered that approximately 24% of drugs designed for the human body have inhibitory effects on microorganisms, particularly antipsychotic drugs [16]. Despite the growing body of research that uncovers the role of microorganisms in the development of human diseases, our understanding of how microorganisms residing in the human body impact human health and contribute to diseases remains limited. The identification of microbes associated with diseases and the prediction of trends in microbial population changes can not only deepen our understanding of potential disease mechanisms but also facilitate early diagnosis and advancements in precision medicine.

In recent years, there has been a proliferation of computational methods for identifying disease-related microbes [17]. These methods can be broadly categorized into four groups: network-based methods, matrix factorization methods, regularization methods, and neural network methods, as discussed by Wang et al. [18] and Wen et al. [19]. (1) Network-based methods: This category includes methods that leverage topological information from networks constructed using multiple databases. For instance, Lei et al. designed LGRSH, which applies the node2vec [20] algorithm to obtain low-dimensional representations and employs an improved rule-based inference method to predict disease-related microbes [21]. (2) Matrix factorization methods: The core idea of these methods is to factorize the input matrix into two lower-dimensional matrices while preserving the reconstructive property. For example, Peng et al. introduced RNMFMDA, which employs random walk with restart for reliable negative sampling. They then employ a neighborhood regularized logistic matrix factorization method to predict disease-related microbes [22]. (3) Regularization methods: These methods involve applying different forms of regularization to least square classifications. Xu et al. proposed MDAKRLS, which combines Kronecker regularized least square with hamming interaction spectral similarity to predict the likelihood of microbe-disease associations [23]. (4) Neural network methods: This category of methods has gained considerable popularity in recent years. Long et al. introduced GATMDA, a framework that represents microbes and diseases and predicts associations using an optimized graph attention network with inductive matrix completion (Fig. 1) [24].

With the accumulation of microbe-disease association data, no research has yet utilized the information on microbial quantity changes under disease status to predict microbe-disease associations, which hampers the comprehensive capture of data and feature distribution between diseases and microbes. Furthermore, most existing models for signed graph representation learning are predominantly designed for social networks and struggle to effectively capture the signed structural characteristics of biological networks [25, 26]. In the realm of social networks, several notable methods for signed graph representation learning have been developed. Derr et al. pioneered the use of Signed Graph Convolutional Network (SGCN [27]), which builds upon the theory of structural balance to obtain signed graph representations. Huang et al. and Li et al. introduced two models, SiGAT [28] and SNEA [29], respectively, which leverage attention mechanisms to differentiate the importance of different neighboring nodes. More recently, Li et al. combined spectral graph theory with graph signal processing techniques and presented a powerful model called SLGNN [30] for capturing the structural information of signed graphs. Taking a spectral perspective, they effectively retained the similarity and dissimilarity between connected nodes by preserving the low-frequency and high-frequency information.

Although significant progress has been made in microbe-disease association prediction tasks, we still face some challenges [31]. First and foremost, the main challenge lies in how to effectively capture a more comprehensive and authentic data distribution using this signed message in microbe-disease association databases. Furthermore, there is a lack of consistency in the conditions of repeated biological experiment validations, and conflicting microbe-disease signed association information also exists in the signed association databases. Modeling the significant amount of noise in association data remains a key issue. Lastly, addressing biased similarity data solely through similarity fusion is insufficient to completely mitigate this bias. It is crucial to explore effective methods that can reduce bias in microbe-disease association studies.

In this study, based on signed message propagation, we propose a framework, Multi-scale Sign Variational Graph AutoEncoder (MSignVGAE), for microbe-disease signed association prediction. MSignVGAE utilizes a graph variational autoencoder to model noisy signed association data and extends the multi-scale concept from previous work [32] to enhance the representational power of the graph variational autoencoder. The

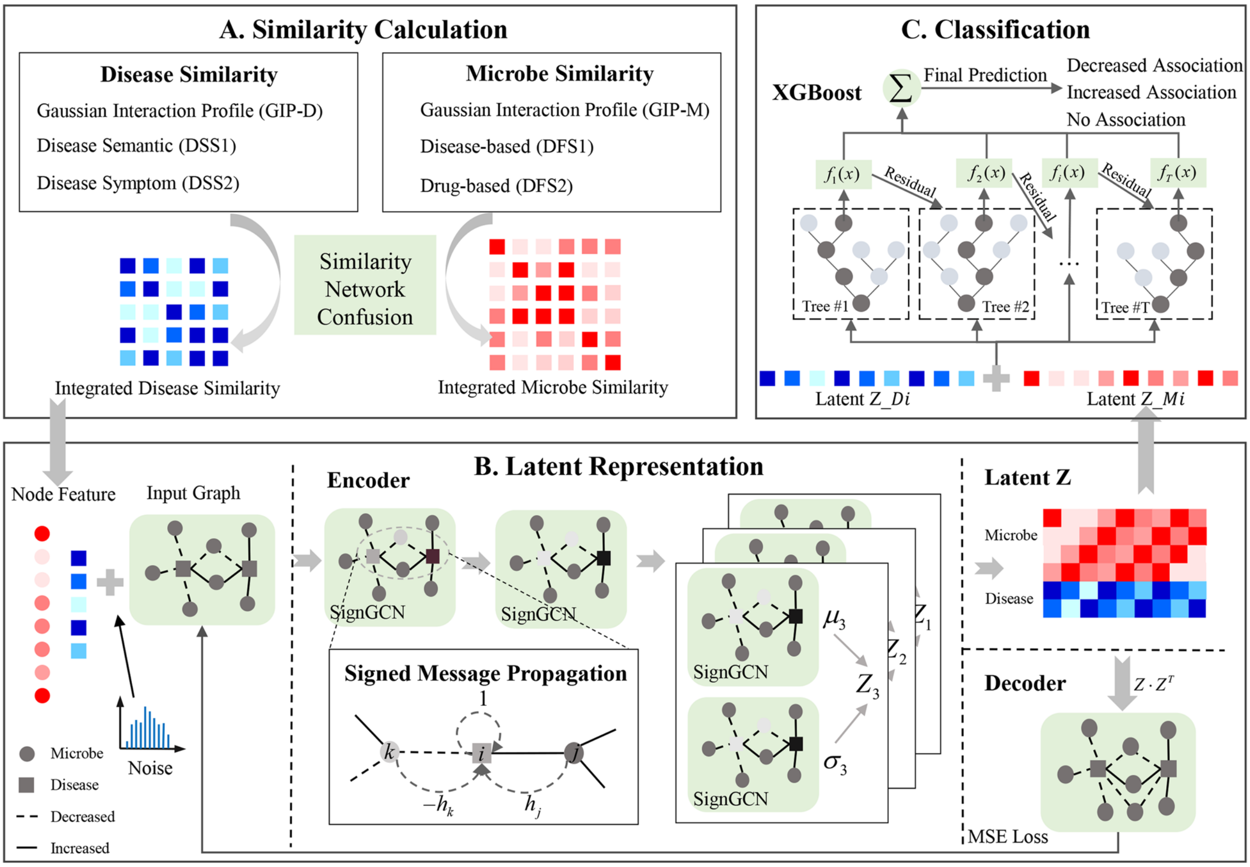

Fig. 1 Framework of MSignVGAE. **A** Calculate the similarities for diseases and microbes. **B** Adopt signed message propagation strategy and VGAE to obtain latent representation for microbes and diseases. **C** Leverage XGBoost for predicting potential disease-related microbes with signs

key contribution of our work lies in the development of a novel strategy for propagating signed message in signed networks. This strategy specifically addresses the propagation process between different nodes, effectively managing the heterogeneity and consistency among nodes connected by various signed edges. Additionally, building upon the similarity network fusion [33] method that combines multiple disease similarity matrixes and microbe similarity matrixes, we further employ the idea of denoising autoencoders to add noise to the similarity data and reconstruct signed associations through the graph variational autoencoder to overcome the bias issue present in the similarity data. MSignVGAE utilizes similarity information as node features to represent the heterogeneous graph of microbe-disease associations and then employs a multiple classifier XGBoost [34] for predicting the signed associations between diseases and microbes. Notably, MSignVGAE is the first method that utilizes signed message to predict microbe-disease signed associations. The AUROC value and AUPR value of MSignVGAE reached 0.9742 and 0.9601, respectively. Furthermore, case studies on three different diseases demonstrate that MSignVGAE, by leveraging the signed message, can capture a more comprehensive distribution of associations.

Results and discussion

Experiment settings

In this study, we employed tenfold cross-validation to ensure the accuracy and reliability of MSignVGAE. We utilized a range of commonly used metrics, including AUROC, AUPR, precision, recall, F1, and accuracy, to evaluate the performance of across all comparison experiments [35, 36]. Considering the sparsity and reliability of the microbe-disease signed association data, this work focuses primarily on experiments conducted using the Peryton database. In the SNF section, the number of neighbors for diseases and microbes in the KNN algorithm is set to 5 and 140, respectively. In the sign graph convolution encoder part, we employed three scales of multi-scale encoders for similarity networks. The scales used were 64, 32, and 16. Moreover, we set the parameters of the XGBoost classifier as default. To control the learning rate during training, we adopted the StepLR strategy, where the learning rate progressively updated until it reached the specified number of epochs. This strategy helps optimize the training process and enhance model convergence.

Ablation study

To analyze the contributions of each module in MSignVGAE, this section conducted ablation experiments based on the Peryton database. The results are shown in Table 1, where MSignVGAE refers to the complete model without removing any modules. Del_Noise represents the model with the similarity feature denoising module removed from MSignVGAE. Del_Multi represents the model with the multi-scale SignGCN removed from the sign graph convolutional encoder module. Del_SignGCN represents the model with a simple GCN module replacing the SignGCN module in MSignVGAE. This section aims to analyze the individual contributions of each component to the overall model accuracy and performance.

As shown in Table 1, it is evident that the whole MSignVGAE model, without removing any modules, achieves the highest performance across various metrics. Among the three ablated modules, the SignGCN module contributes the most. In fact, even when using the original GCN, which is not specifically designed for signed graph neural networks, the performance in the prediction task of microbe-disease signed associations is still considerable, with AUROC and AUPR values reaching 0.9418 and 0.9065, respectively. This can be attributed to the fact that the similarity features of diseases and microbes already possess certain representational capacity before undergoing graph representation learning. However, the original GCN fails to further integrate the similarity features with the structural information of the signed graph network, resulting in unsatisfactory performance in signed association prediction task. Furthermore, the improvement brought by the similarity feature denoising module is also significant. It enhances the overall model performance by 0.78% and 1.38% in terms of AUROC and AUPR, respectively. This indicates that the similarity feature denoising module helps further enhance the robustness of the model within the VGAE framework. The last ablated module is the multi-scale SignGCN module. From Table 1, it can be observed that although the performance improvement brought by the multi-scale SignGCN module is relatively small, this module allows the model to learn effective representations even when the AUROC reaches 0.9731. Considering the remaining potential for improvement, the multi-scale SignGCN module achieves a 4.09% improvement in the AUROC metric.

Performance comparison with SOTA methods

Considering that MSignVGAE is the first method to utilize sign information for predicting microbe-disease associations, we selected a subset of state-of-the-art (SOTA) graph representation learning methods from the fields of unsupervised graph representation learning, signed network embedding, and graph theory for comparison. After obtaining node representations using these SOTA graph representation learning methods, we uniformly input them into an XGBoost multi-classification model for microbe-disease signed association prediction. Additionally, considering the notable performance of the GATMDA model in microbe-disease association prediction task, we also compared it in the context of microbe-disease signed association task. Moreover, in line with the approach of the MVGAEW model, this section also compared MSignVGAE from the perspective of reconstructing similarity matrixes. All comparison experiments in this section were conducted on the Peryton database, and the results can be found in Fig. 2 and Table 2. The methods included in these comparisons are as follows:

- MSignVGAE-sign2: This method follows the approach of the MVGAEW model. It utilizes the known microbe-disease signed association matrix as node features and reconstructs the disease-disease similarity matrix and microbe-microbe similarity matrix separately. It employs a multi-classification XGBoost model to predict the presence of associations between diseases and microbes and to identify the associated signs in cases where associations exist.
- TrustSGCN [37]: TrustSGCN is a novel signed network embedding method based on GCN. This model introduces a strategy to measure the credibility of high-order associated sign edges inferred from the

Table 1 Performance of ablation experiments based on Peryton database

Method	AUROC	AUPR	F1	Precision	Recall	Accuracy
MSignVGAE	**0.9742**	**0.9611**	**0.8738**	**0.8744**	**0.8738**	**0.8841**
Del_Noise	0.9667	0.9480	0.8516	0.8534	0.8504	0.8651
Del_Multi	0.9731	0.9585	0.8719	0.8725	0.8719	0.8823
Del_SignGCN	0.9418	0.9065	0.7507	0.7508	0.7515	0.7975

The bold values denote the max value in columns

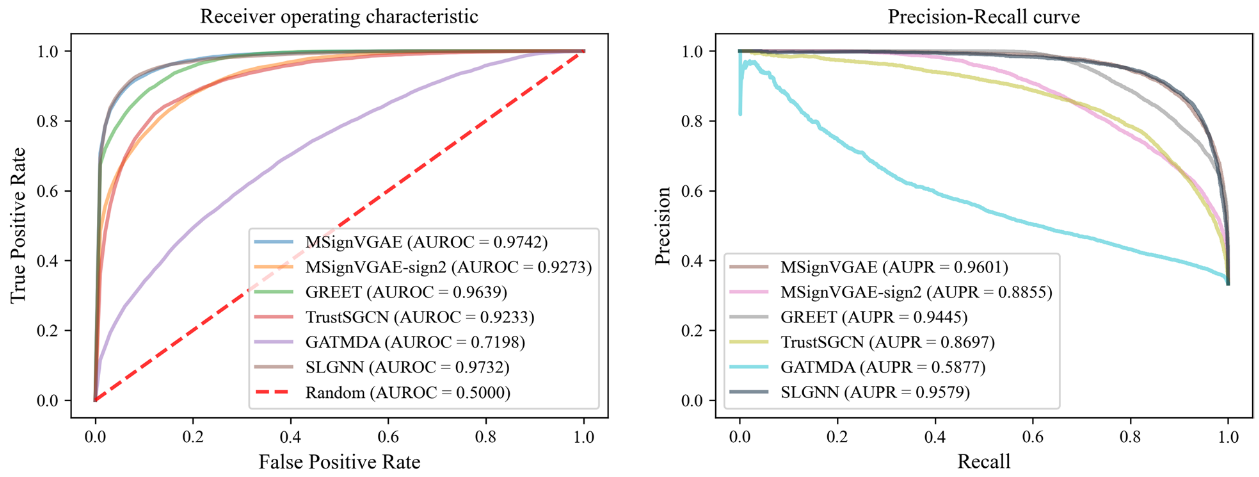

Fig. 2 The ROC curve and PR curve for signed association prediction under tenfold cross-validations on Peryton database

Table 2 Performance comparison for signed association prediction under 10-fold cross-validations on Peryton database

Method	AUROC	AUPR	F1	Precision	Recall	Accuracy
MSignVGAE	**0.9742**	**0.9601**	**0.8738**	**0.8744**	**0.8738**	**0.8841**
MSignVGAE-sign2	0.9273	0.8855	0.7398	0.7407	0.7395	0.7771
TrustSGCN	0.9233	0.8697	0.7710	0.7810	0.7635	0.7903
GREET	0.9639	0.9445	0.8133	0.8137	0.8149	0.8446
GATMDA	0.7198	0.5877	0.4500	0.4619	0.4510	0.5286
SLGNN	*0.9732*	*0.9579*	*0.8715*	*0.8726*	*0.8708*	*0.8793*

The bold values denote the maximum value in columns, while the italicized values represent the second-best value in columns

theory of structural balance. It further corrects the incorrect embedding propagation process in the structural balance theory based on the credibility strategy.

- GREET [38]: GREET addresses the tendency of existing unsupervised graph representation learning methods to perform smooth learning along all edges, thereby neglecting the heterogeneity of nodes with different attributes. It constructs a homogeneous/heterogeneous edge discriminator to infer the homogeneity/heterogeneity of edges based on both feature and structural information. By minimizing a carefully designed pivot-ranking loss, GREET utilizes a homogeneous/heterogeneous dual-channel encoder to learn node representations.
- SLGNN [30]: SLGNN is based on graph theory and graph signal processing. It designs different low-pass and high-pass graph convolution filters to extract low-frequency and high-frequency information from positive and negative links, respectively. It employs a "self-gating" mechanism to control the influence of low-frequency and high-frequency information during the message passing process, thereby combining them into a unified message propagation framework.
- GATMDA [24]: It incorporates the concept of "Talking Head" into an optimized graph attention network to learn latent representations of microbes and diseases.

Figure 2 displays the receiver operating characteristic curves and precision-recall curves of the comparative methods in signed association prediction task. Table 2 presents the performance of different methods across multiple metrics in signed association prediction task. Compared to other methods, the proposed MSignVGAE model demonstrates superior performance across all metrics, showcasing its excellent performance. It is worth noting that compared to MSignVGAE-sign2, MSignVGAE shows a 5.06% improvement in AUROC and an 8.42% improvement in AUPR. This suggests that directly reconstructing known signed associations and using similarity information as features is more effective than reconstructing disease-disease similarity matrixes and microbe-microbe similarity matrixes separately while using known signed association information as node features. The reason behind this improvement is that when separately reconstructing similarity matrixes, the fusion of heterogeneous graph structure information and similarity information is disconnected. In contrast, reconstructing known signed associations enables an additional cross-fusion of the two types of information.

Furthermore, both GREET and SLGNN exhibit high performance across various metrics. This can be attributed to the utilization of low-pass and high-pass filters in both methods, enabling the models to learn features at different hierarchical levels. In comparison to these two methods, the advantage of MSignVGAE lies in its multi-scale SignGCN module, which integrates different sign propagation processes into a unified whole. This allows the model to capture features at different hierarchical levels without the need to separate low-frequency and high-frequency information of the graph. Notably, TrustSGCN demonstrates superior performance on real-world sign networks but performs poorly in microbe-disease signed association prediction task. This is because there stands a significant difference between real-world sign networks and heterogeneous networks in the bioinformatics domain, making the structural balance theory inapplicable to heterogeneous networks in the bioinformatics field. An intuitive observation is that GATMDA exhibits a noticeable performance gap in signed association prediction task compared to other methods. This discrepancy may be attributed to GATMDA's failure to consider the influence of sign information. Based on this observation, it can be inferred that using only similarity information can still maintain a certain level of signed association prediction capability.

Performance comparison in unsigned association prediction

To further validate the effectiveness of the MSignVGAE method in efficiently integrating signed features, this section utilizes representations obtained from state-of-the-art graph representation learning methods in signed association prediction tasks. These representations are then inputted into an XGBoost binary classification model for unsigned association prediction. The results can be found in Fig. 3 and Table 3.

Figure 3 displays the receiver operating characteristic curves and precision-recall curves of the comparative methods in unsigned association prediction tasks. Table 3 presents the performance of different methods across multiple metrics in unsigned association prediction tasks. Compared to other methods, the proposed MSignVGAE model is not the optimal one in terms of performance. However, all its performance metrics are comparable to those of MVGAEW on the Peryton

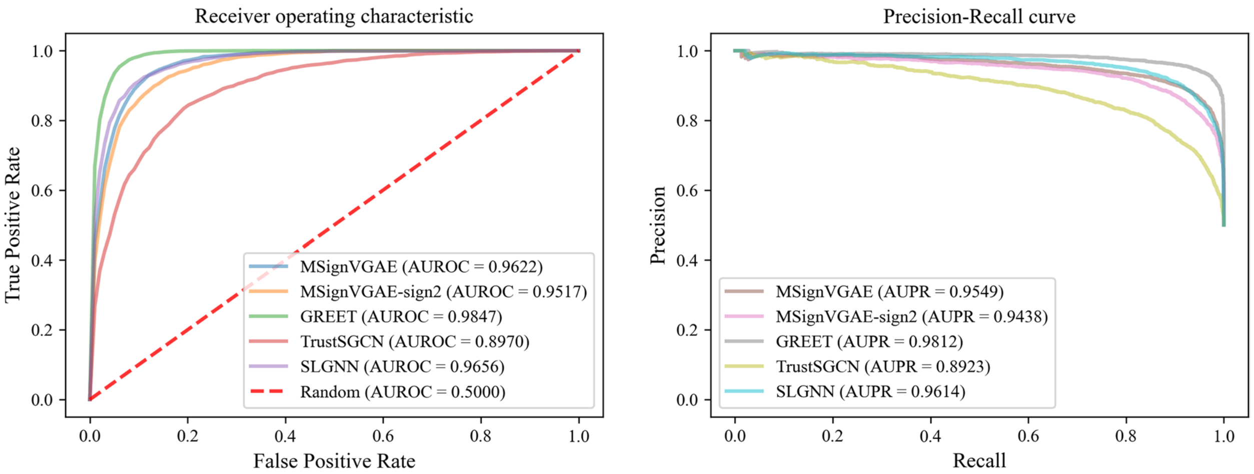

Fig. 3 The ROC curve and PR curve for unsigned association prediction under tenfold cross-validations on Peryton database

Table 3 Performance comparison for unsigned association prediction under tenfold cross-validations on Peryton database

Method	AUROC	AUPR	F1	Precision	Recall	Accuracy
MSignVGAE	0.9622	0.9549	0.9059	0.8887	*0.9241*	0.9041
MSignVGAE-sign2	0.9517	0.9438	0.8861	0.8694	0.9037	0.8839
TrustSGCN	0.8970	0.8923	0.8217	0.8143	0.8294	0.8201
GREET	**0.9847**	**0.9812**	**0.9489**	**0.9286**	**0.9703**	**0.9478**
SLGNN	*0.9656*	*0.9614*	*0.9071*	*0.8958*	0.9188	*0.9059*

The bold values denote the maximum value in columns, while the italicized values represent the second-best value in columns database [32]. This indicates that utilizing sign information does not have a significant impact on unsigned association prediction tasks and can improve the accuracy of the model. Notably, in Table 3, both the GREET and SLGNN methods exhibit superior performance compared to the MSignVGAE method. However, interestingly, their performance is relatively lower than the MSignVGAE method in Table 2. This observation suggests that the MSignVGAE method, specifically designed to incorporate sign information, effectively integrates sign features when predicting associations.

Furthermore, on the Peryton database, the GREET method demonstrates a significant improvement across various performance metrics compared to the MVGAEW method. This finding further emphasizes the benefits of utilizing sign information, as it leads to enhanced model performance. Consistent with the results presented in Table 2, the MSignVGAE method also outperforms the MSignVGAE-sign2 method in unsigned association prediction tasks. However, it is noteworthy that TrustSGCN, which demonstrates outstanding performance in real-world signed networks, does not perform well in microbe-disease association prediction tasks. This finding highlights the inherent differences between real-world signed networks and the heterogeneous networks present in the field of bioinformatics.

Performance comparison with widely used databases

With the accumulation of data, databases have become more mature and now contain an increasing number of effective signed associations between microbes and diseases. To verify the generalization ability of MSignVGAE on databases of different scales, this section conducts several experiments on three additional databases (HMDAD, Disbiome, and MicroPhenDB), all of which also contain sign information. Considering the sparse matching of microbes between the microbe-disease database and the microbe-drug database, this section calculates the microbial similarity in the latter database without relying on drug-based functional similarity. Table 4 presents the performance comparison of different microbe-disease signed association databases under tenfold cross-validation. It can be observed that the performance of MSignVGAE is lowest on the HMDAD database. This is because the HMDAD data contains fewer signed association samples. Even in the case of a small dataset, MSignVGAE still maintains good signed association prediction ability. Consistent with the trend observed in previous work [32], as the quality and quantity of signed associations between diseases and microbes in the database increase, the performance of the MSignVGAE model also improves.

Interpretation of latent representation

In order to further explore the interpretability of latent representations from the distributional perspective, this section visualizes the feature distributions of microbe representations. Specifically, we achieved this by employing the t-SNE [39] method, which is a dimensionality reduction technique, to project high-dimensional data into a lower-dimensional space for visualization. The visualization results for the Peryton database are illustrated in Fig. 4. Figure 4a displays the distribution of microbe representations obtained after applying MSignVGAE to the microbe-disease sign association matrix, while Fig. 4b shows the distribution of the original microbe-microbe similarity matrix. In Fig. 4, the points labeled as "Increased" and "Decreased" represent different types of changes (increase or decrease) in microbe quantity under disease states. The points labeled as "Non_association" represent microbes that are not associated with Alzheimer's disease [40] in the Peryton database. The points labeled as "Increased_pred" and "Decreased_pred" represent the top 50 microbes predicted by the MSignVGAE model to have the highest probability of increasing or decreasing in association with Alzheimer's disease.

From Fig. 4b, it can be observed that the two types of microbes in the original Peryton database are roughly distributed together. This indicates that the original microbe feature distribution is difficult to distinguish between the two types of microbes. In Fig. 4a, the dark green and light green points tend to be biased towards the left. This phenomenon is primarily due to the introduction of sign information propagation strategy, which causes the feature distributions of different types of microbes to be pulled apart from each other. One notable

Table 4 The comparison of different microbe-disease signed association databases under tenfold cross-validation

Database	AUROC	AUPR	F1	Precision	Recall	Accuracy
HMDAD	0.9028	0.8313	0.7112	0.7232	0.7099	0.7556
Disbiome	0.9367	0.8954	0.7696	0.7718	0.7680	0.7995
MicroPhenDB	0.9673	0.9463	0.8418	0.8427	0.8417	0.8666
Peryton	**0.9742**	**0.9601**	**0.8738**	**0.8744**	**0.8738**	**0.8841**

The bold values denote the max value in columns

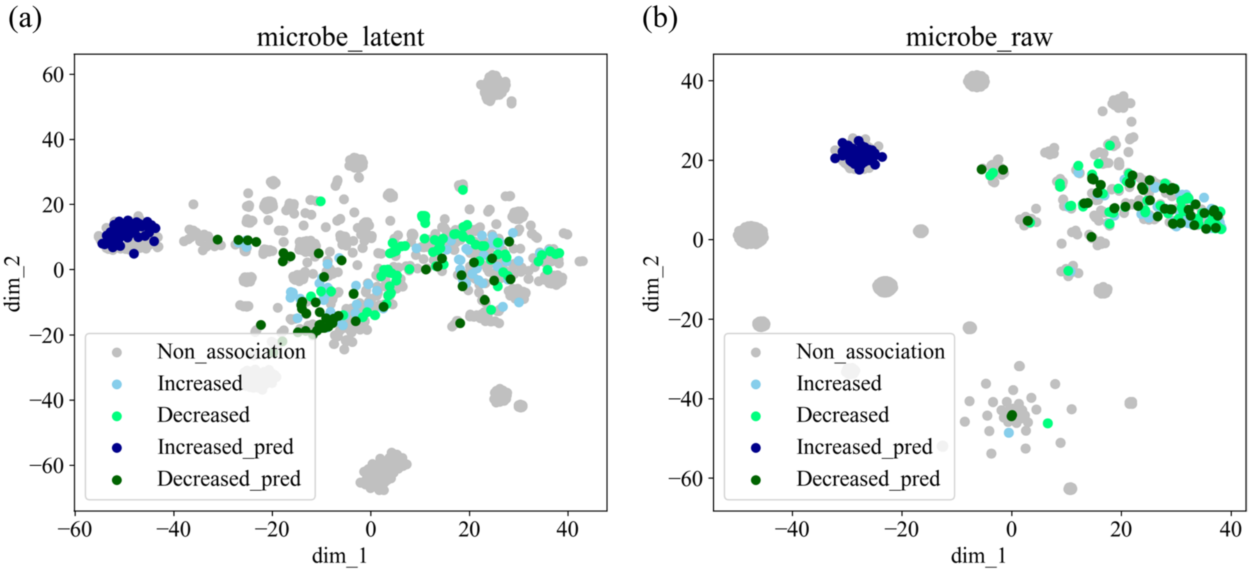

Fig. 4 Visualizations of distribution whether adopt MSignVGAE for microbes related to Alzheimer's disease. **a** The latent distribution by adopting MSignVGAE. **b** The raw distribution of integrated similarity network

observation is the presence of outlier clusters in deep blue in the original feature distribution. These clusters exhibit significant differences in distribution compared to the known signed associations. Most proposed methods for predicting microbe-disease associations tend to predict potential disease-related microbes that fall near the known association distribution, rather than exhibiting outlier clusters similar to the deep blue cluster in Fig. 4a. Regarding the deep blue cluster in Fig. 4a, the prediction of these microbe nodes is likely to rely on the dark green nodes (predicted to be "Decreased" type) as bridges to establish connections between the light blue nodes (predicted to have "Increased" type) and the deep blue nodes (predicted to have "Increased" type).

Case studies

Consistent with previous work [32], this section presents case studies focusing on specific diseases to showcase the predictive capability for disease-related microbes. The diseases examined in this section are colorectal neoplasms [41], Alzheimer's disease [40], and Crohn's disease [42]. The analysis is based on the Peryton database, where known microbe-disease signed associations were excluded. The top 20 "Increased" and "Decreased" microbes with the highest predicted probabilities for each respective disease were identified. Furthermore, relevant literature from PubMed was provided to substantiate the presence of these signed associations. The specific microbes associated with colorectal neoplasms, Alzheimer's disease, and Crohn's disease can be found in Tables 5, 6, and 7, respectively.

By cross-referencing the prediction results from this section with the corresponding results from previous work [32], we identified common microbes that were predicted as relevant. These include (1) Epsilonproteobacteria, Schwartzia, and Bacillaceae associated with Crohn's disease; (2) Erysipelotrichia, Erysipelatoclostridium, and Flavonifractor associated with colorectal neoplasms; and (3) Klebsiella and Oscillospira associated with Alzheimer's disease. Furthermore, it is noteworthy that all the commonly predicted relevant microbes in both studies belong to the "Decreased" type, indicating that the omission of signed message may introduce a certain bias in the model. Among them, only Erysipelotrichia lacks literature reporting its association with colorectal neoplasms, while all other commonly predicted microbes are supported by literature. This suggests a high likelihood of a "Decreased" association between Erysipelotrichia predicted by the MSignVGAE model and colorectal cancer, despite the lack of specific literature evidence.

Additionally, this section visualizes the distribution of known signed associations and predicted signed associations related to specific diseases, as depicted in Fig. 5. The distribution of "Decreased" associations predicted by MSignVGAE also reveals a pattern where central microbes influence multiple diseases. Furthermore, the distribution of "Increased" associations predicted by MSignVGAE tends to be associated with a single disease, indicating that this pattern arises from the transmission of signed message and was not observed in the case studies of previous work.

Table 5 Top-20 "Increased" and "Decreased" microbes associated with colorectal neoplasms

Increased			Decreased		
Rank	**Microbes**	**PMID**	**Rank**	**Microbes**	**PMID**
1	Prevotella Scopos	31791356	1	Abiotrophia	30112040
2	Actinomyces sp. oral taxon 877	31171880	2	Erysipelotrichia	Unconfirmed
3	Bacteroides-like sp. oral clone AU126	Unconfirmed	3	Sphingomonas Azotifigens	28914591
4	Thermales	37317301	4	Pseudomonas Aeruginosa	36018829
5	Sporosarcina	Unconfirmed	5	Erysipelatoclostridium	35269806
6	Thermaerobacter	Unconfirmed	6	Cutibacterium Acnes	38027096
7	Eubacterium Ramulus	Unconfirmed	7	Flavonifractor	34799562
8	Oribacterium sp. oral taxon 108	31609493	8	Limosilactobacillus fermentum	31729242
9	Arthrobacter	30101351	9	Raoultella	Unconfirmed
10	Anaerotignum Lactatifermentans	Unconfirmed	10	Schlegelella	Unconfirmed
11	Shigella Boydii	Unconfirmed	11	Negativicutes	31619268
12	Blautia sp.	36539569	12	Erysipelotrichales	Unconfirmed
13	Porphyromonas Bennonis	31450675	13	Brevibacillus	Unconfirmed
14	Shigella Flexneri	30418409	14	Candidatus Saccharibacteria	30574173
15	Entylomataceae	Unconfirmed	15	Lachnobacterium	Unconfirmed
16	Tremellales	Unconfirmed	16	Anaerotruncus Colihominis	Unconfirmed
17	Eggerthellaceae Bacterium AT8	36313092	17	Methanobrevibacter Smithii	15963794
18	Streptococcus Gallolyticus subsp. Gallolyticus	29666615	18	Mycoplasma	37772998
19	Ruminococcus Bicirculans	37548332	19	Methanobacteria	35420474
20	Neisseria Mucosa	32517306	20	Barnesiellaceae	Unconfirmed

Table 6 Top-20 "Increased" and "Decreased" microbes associated with Alzheimer's disease

Increased			Decreased		
Rank	**Microbes**	**PMID**	**Rank**	**Microbes**	**PMID**
1	Pseudogymnoascus sp. VKM F-4518 (FW-2643)	36861650	1	Limosilactobacillus Fermentum	33536656
2	Neurospora Crassa	32946564	2	Tissierellaceae	Unconfirmed
3	Pisolithus	Unconfirmed	3	Prevotella Copri	36093695
4	Victivallales	35275534	4	Streptococcus Sanguinis	Unconfirmed
5	Fusobacterium Naviforme	35364661	5	Shigella	27776263
6	Cetobacterium Somerae	Unconfirmed	6	[Ruminococcus] Gnavus	37254223
7	Anaerolineae	Unconfirmed	7	Streptococcus Mutans	35139675
8	Actinomyces Radicidentis	Unconfirmed	8	Burkholderiaceae	36286029
9	Salmonella Enterica	30723884	9	Klebsiella	36068280
10	Schaalia Cardiffensis	Unconfirmed	10	Oscillospira	36185477
11	Prevotella sp. oral taxon 300	35364661	11	Micrococcus	2560791
12	Aspergillus Versicolor	Unconfirmed	12	Fusobacteriaceae	Unconfirmed
13	Treponema sp. oral taxon 250	35364661	13	Roseburia	36430144
14	Olsenella Profusa	Unconfirmed	14	Erysipelatoclostridium	36615777
15	Phascolarctobacterium Succinatutens	Unconfirmed	15	Porphyromonas Endodontalis	Unconfirmed
16	Cardiobacteriales	Unconfirmed	16	Capnocytophaga	35950713
17	Tannerella Forsythia	26063967	17	Megasphaera	Unconfirmed
18	Thielaviopsis	Unconfirmed	18	Fusobacterium Nucleatum	25576662
19	Peptoniphilaceae	32533776	19	Centipeda	27846826
20	Neisseria Oralis	Unconfirmed	20	Escherichia Coli	29472250

Table 7 Top-20 "Increased" and "Decreased" microbes associated with Crohn's disease

Increased			Decreased		
Rank	**Microbes**	**PMID**	**Rank**	**Microbes**	**PMID**
1	Uncultured Selenomonas sp.	Unconfirmed	1	Epsilonproteobacteria	32040665
2	Bordetella	27557706	2	Oceanospirillales	Unconfirmed
3	Orthomyxoviridae	24374880	3	Prevotella Nanceiensis	Unconfirmed
4	Poxviridae	23624886	4	Fusobacterium Varium	29216329
5	Cladosporium	34850076	5	Tissierellaceae	Unconfirmed
6	Polyomaviridae	20298966	6	Schwartzia	3318407
7	Geotrichum	Unconfirmed	7	Bacillaceae	35967326
8	Spirochaeta	4235262	8	Bifidobacterium Bifidum	37240476
9	Uncultured Succinivibrionaceae Bacterium	33125440	9	Bradyrhizobium	Unconfirmed
10	Hymenolepis	20044996	10	Streptococcus Parasanguinis	34427649
11	Edwardsiella	31016054	11	Corynebacteriales	Unconfirmed
12	Toxocara	26878617	12	Raoultella	37337895
13	Pleistophora	Unconfirmed	13	Acidobacteria	26922889
14	Arcanobacterium	Unconfirmed	14	Corynebacteriaceae	31155731
15	Uncultured Veillonellaceae Bacterium	24629344	15	Filifactor	Unconfirmed
16	Cardiobacteriales	Unconfirmed	16	Capnocytophaga	35950713
17	Tannerella Forsythia	26063967	17	Megasphaera	Unconfirmed
18	Thielaviopsis	Unconfirmed	18	Fusobacterium Nucleatum	25576662
19	Peptoniphilaceae	32533776	19	Centipeda	27846826
20	Neisseria Oralis	Unconfirmed	20	Escherichia Coli	29472250

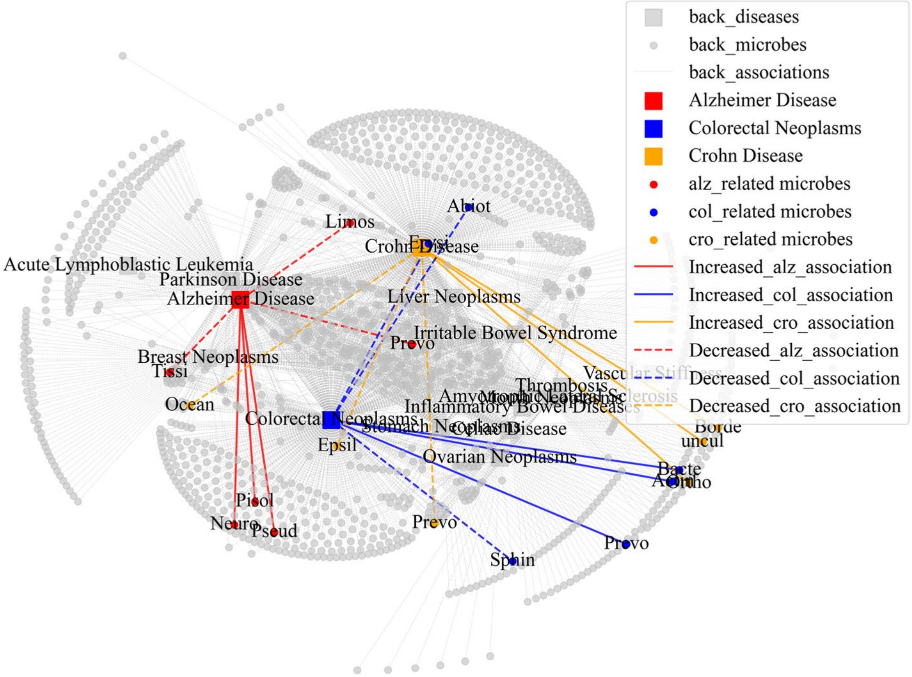

Fig. 5 The distribution of existing and predicted Increased/Decreased association related to case diseases

Methods

Data sources

Microbe-disease association databases

Until now, researchers have developed several widely used microbe-disease association prediction databases, summarized in Table 8. Ma et al. [43] developed the first Human Microbe–Disease Association Database (HMDAD). By eliminating redundancy, HMDAD gathered 450 confirmed microbe-disease associations between 292 microbes and 39 diseases from published literature. Among these associations, there were 205 "Decreased" type associations and 245 "Increased" type associations. In 2018, Janssens et al. [44] established Disbiome, a database documenting 8731 known associations between 1622 microbes and 374 diseases. The content was selected from 1,191 published academic papers without redundancy, and the numbers for "Decreased" and "Increased" types were 4157 and 4574, respectively. Subsequently, MicroPhenDB was constructed using the same methodology as HMDAD and Disbiome by Yao et al. [45]. It included 5511 non-redundant associations between 1774 microbes and 500 diseases in 22 newly collected human body sites. Among these associations, there were 1819 "Decreased" type associations and 3692 "Increased" type associations. The last one, Peryton, proposed by Skoufos et al. [46], encompasses 4172 associations that are supported by experimental evidence, linking 1396 microbes with 43 diseases. Specifically, there were 2130 associations categorized as "Decreased" and 2042 associations categorized as "Increased." To facilitate usability, we transformed the information regarding known microbe-disease signed associations into a matrix $A \in \mathbb{R}^{nm \times nd}$. In this matrix, a value of 1 indicates the presence of an increased microbe-disease association in the database, while a value of -1 indicates the presence of a decreased microbe-disease association. Conversely, a value of 0 signifies the absence of the corresponding item. Let us denote the variables *nd* and *nm* to represent the number of diseases and microbes, respectively.

Similarity calculation

Based on previous related work [32, 47–53], This study extends the similarity calculation methods within the MVGAEW model framework. The key distinction lies in the utilization of disease-disease similarity and microbe-microbe similarity compared to the known microbe-disease association matrix. In the MVGAEW model framework, the association matrix elements are binary, taking values of either 0 or 1 to indicate the absence or presence of an association, respectively. However, in this study, the known microbe-disease signed association matrix is used. In this matrix, elements representing edges connecting decreased associations are assigned a value of -1, whereas edges connecting increased associations are assigned a value of 1. Elements corresponding to no association retain a value of 0. The disease-disease similarity measures employed in this study encompass disease Gaussian interaction profile kernel similarity (GIP-D), disease semantic similarity (DSS1), and disease symptom similarity (DSS2). For microbe-microbe similarity, the measures include microbe Gaussian interaction profile kernel similarity (GIP-M), disease-based functional similarity (DFS1), and drug-based functional similarity (DFS2). Finally, a similarity network fusion approach [33] is employed to separately integrate the similarities of diseases and microbes, enabling a comprehensive analysis and understanding of the relationships within the system.

MSignVGAE

The overall framework of MSignVGAE is depicted in Fig. 1. Firstly, MSignVGAE employs a similarity network fusion approach independently integrate multiple disease similarities and microbe similarities. Furthermore, MSignVGAE utilizes a graph variational autoencoder with a signed message propagation strategy to reconstruct the known microbe-disease signed association matrix. The noisy similarity data, which has undergone a denoising process, is employed as the initial feature input for the variational autoencoder component. Notably, signed graph structural features are leveraged to characterize diseases and microbes. Lastly, based on the representations of diseases and microbes, a multi-class XGBoost classifier is applied to determine the presence of associations between given microbe-disease pairs and identify the corresponding signs for the associations.

Table 8 Databases for microbe-disease association prediction

Database	Associations	Microbes	Diseases	Decreased	Increased	Year
HMDAD	450	292	39	205	245	2016
Disbiome	8731	1622	374	4157	4574	2018
MicroPhenDB	5511	1774	500	1819	3692	2020
Peryton	4172	1396	43	2130	2042	2021

Figure 1 illustrates the framework of MSignVGAE, and subsequent sections of this paper will elaborate on each component in the framework in detail.

Similarity feature noising

Similarity feature noising refers to the introduction of Gaussian noise to the similarity data during the processing. Building upon the utilization of a similarity network fusion approach to integrate multiple disease similarity matrixes and microbe similarity matrixes, further advancements are achieved by incorporating the concept of a denoising autoencoder. This involves the addition of Gaussian noise to the similarity data and utilizing a graph variational autoencoder to reconstruct signed associations, thereby overcoming biases present in the similarity data. For convenience, the node similarity features were represented as F, as shown below:

$$F = \begin{pmatrix} SM, O \\ O, SD \end{pmatrix}^{(nd+nm)}, \tag{1}$$

where SD denotes the integrated disease similarity matrix, SM denotes the integrated microbe similarity matrix, O denotes the zero matrix, and F is a $(nd + nm)$ -dimensional square matrix. After applying Gaussian noise, the node similarity features F' can be expressed as:

$$F' = F + \varepsilon^{(nd+nm)},\ where\ \varepsilon^{(nd+nm)} \in N(0,1), \tag{2}$$

where $\varepsilon^{(nd+nm)}$ represents Gaussian noise following a standard normal distribution, with dimensions matching F.

Sign graph convolution encoder

For convenience, in this section, the initial graph node features X denotes the node similarity features F' after adding Gaussian noise. This module consists of two shared SignGCN layers and a multi-scale variational inference layer. Each scale of the variational inference layer has two SignGCN modules, which calculate the mean μ and variance σ of the latent variable Z, respectively. Additionally, W_0 represents the model parameters that need to be learned in the first SignGCN layer. The first shared SignGCN layer can be represented by the following equation:

$$\overline{X_1} = SignGCN(X, A) = \mathrm{ReLU}\left(\overline{A_{norm}} \cdot X \cdot W_0\right), \tag{3}$$

$$\overline{A_{norm}} = \widetilde{D}^{-\frac{1}{2}} \cdot \overline{A} \cdot \widetilde{D}^{-\frac{1}{2}}, \widetilde{D} = \overline{D} + I, \tag{4}$$

$$\overline{D} = diag\left\{\overline{d}_1, \cdots, \overline{d}_{nd+nm}\right\}, \overline{d}_i = \sum\nolimits_j |A_{ij}|, \tag{5}$$

where $\overline{A}$ represents the matrix A with self-loops, which can be denoted as $\overline{A} = A + I$. $\overline{A_{norm}}$ represents the matrix after symmetrically normalized Laplacian matrix processing. Compared with unsigned GCN, in SignGCN, the used $\widetilde{D}$ is no longer the degree matrix of the input graph structure matrix with self-loops but the absolute degree matrix of the signed association matrix in Eqs. (4) and (5). In essence, the matrix in this section corresponds to the low-pass feature aggregation filter [54]. The propagation of sign information in this filter is illustrated in the sign information propagation module in Fig. 1 and can be represented by the following equation:

$$h_i^l = \frac{1}{\overline{d}_i} h_i^{l-1} - \sum_{k \in \mathcal{N}_i^-} \frac{1}{\sqrt{\left(\overline{d}_i + 1\right)\left(\overline{d}_k + 1\right)}} h_k^{l-1} + \sum_{j \in \mathcal{N}_i^+} \frac{1}{\sqrt{\left(\overline{d}_i + 1\right)\left(\overline{d}_j + 1\right)}} h_j^{l-1} \tag{6}$$

In details, h_i^l represents the feature vector of the i-th node in the l-th layer of SignGCN. $\mathcal{N}_i^-$ represents the neighboring nodes that have "Decreased" associations with node i, while $\mathcal{N}_i^+$ represents the neighboring nodes that have "Increased" associations with node i. By utilizing the absolute degree matrix as weights for aggregating information from nodes connected by different signed edges, the model can effectively control the diversity and consistency among nodes with different signed associations. The equation for the second shared SignGCN layer can be expressed as follows:

$$\overline{X_2} = SignGCN\left(\overline{X_1}, A\right) = \mathrm{ReLU}\left(\overline{A_{norm}} \cdot \overline{X_1} \cdot W_1\right), \tag{7}$$

where W_1 represents the model parameters that need to be learned in the second shared SignGCN layer. Similarly, the third multi-scale SignGCN layer represents the data distribution using the logarithm of the mean μ and the logarithm of the variance σ, as follows:

$$\mu_i = SignGCN_\mu\left(\overline{X_2}, A\right) = \overline{A_{norm}} \cdot \overline{X_2} \cdot W_\mu^i,\ i \in \{1, 2, 3\}, \tag{8}$$

$$\log \sigma_i = SignGCN_\sigma\left(\overline{X_2}, A\right) = \overline{A_{norm}} \cdot \overline{X_2} \cdot W_\sigma^i,\ i \in \{1, 2, 3\}, \tag{9}$$

Considering that the concatenation and reparameterization technique in previous work, the resulting latent variables are shown below:

$$Z = Z_1 | Z_2 | Z_3,\ Z_i = \mu_i + \sigma_i * \varepsilon,\ \varepsilon \in N(0,1), \tag{10}$$

where   denotes concatenation procedure.

Dot decoder

After obtaining low-dimensional representations Z through a multi-scale encoder, in this section, a simple and efficient dot product decoder is utilized to reconstruct the signed association matrix, denoted as $\widehat{A}$. The matrix $\widehat{A}$ is used to reconstruct the input matrix A, as shown below [32]:

$$\widehat{A} = Z \cdot Z^{T}, \tag{11}$$

In fact, the dot product decoder module alone can achieve satisfactory outcomes for microbe-disease signed association prediction task. However, considering that the objective of the decoder in the graph variational autoencoder framework is to reconstruct the original input matrix as accurately as possible, its core lies in the fusion of similarity features with heterogeneous network structure information. Therefore, relying solely on the predictions of the simple dot product decoder tends to favor known associations. To overcome this limitation, an efficient ensemble learning method, XGBoost, is employed to fully leverage the strengths of the graph variational autoencoder in effectively integrating similarity features with heterogeneous network structure information. This approach enhances the overall performance of the MSignVGAE model framework.

Loss function

The loss function can be formulated as below [55]:

$$\begin{aligned} L = & \frac{1}{nd \cdot nm} \sum_{i}^{nd} \sum_{j}^{nm} \left(\widehat{A}_{ij} - A_{ij}\right)^{2} \\ & + \frac{1}{M} \sum_{m=1}^{M} \left(KL[q(Z_m|A,X)|p(Z_m)]\right) \end{aligned} \tag{12}$$

In details, the first part $\sum_{i}^{nd} \sum_{j}^{nm} \left(\widehat{A}'_{ij} - A'_{ij}\right)^{2} / (nd \cdot nm)$ represents the mean square error loss between the input signed association matrix A and the reconstructed signed association matrix $\widehat{A}$. The second part represents the Kullback–Leibler divergence loss between the latent representation distributions $q(Z_m|SM, X)$ at all scales and the prior standard normal distribution $p(Z_m) \sim N(0, I)$. Additionally, similar to MVGAEW, each iteration of MSignVGAE involves training on the entire graph and utilizes the Adam optimizer [56] to optimize the learnable parameters of the MSignVGAE model. To ensure model convergence, a stepLR learning rate decay strategy is employed during the training phase of MSignVGAE to control the learning rate.

XGBoost classifier

In this work, similar to MVGAEW, MSignVGAE also utilizes the concatenation of disease representations and microbe representations to train an XGBoost [34] multiclass classification model. The objective is to predict the existence of associations between pairs of microbes and diseases as well as the specific type of association (e.g., an edge indicating an increase in microbe abundance or a decrease in microbe abundance).

XGBoost is known for its excellent scalability [57–60] and can be easily extended from binary classification to multiclass task. In the multiclass XGBoost setting, the One-vs-All strategy is employed for classification. This means that a separate binary classification XGBoost model is trained for each class, treating the target class as the positive class and the other classes as the negative class. The goal of each binary classification XGBoost model is to differentiate whether a sample belongs to the current class or not. The models are then optimized using gradient boosting algorithms. The multiclass algorithm in XGBoost uses class scores to indicate the degree of membership for each class. It employs the soft-max Loss function for optimization. By normalizing the class scores, it yields the probability distribution of a sample belonging to each class.

Conclusions

In this work, we propose a novel model framework called MSignVGAE, which can effectively identify disease-associated microbes and predict trends in microbial quantity changes. Firstly, we start with fine-grained signed message and design a new strategy for signed message propagation that defines the information dissemination process between different nodes while controlling the heterogeneity and consistency among nodes connected by different signed edges. Secondly, we employ a graph variational autoencoder framework with a multi-scale perspective to model the signed association data and address the issue of inconsistent signed associations. Additionally, we utilize the denoising autoencoder approach to handle the noise in similarity feature information, which helps overcome biases in the fused similarity data. Notably, MSignVGAE is the first method that utilizes signed message to predict microbe-disease signed associations. The AUROC value and AUPR value of MSignVGAE reached 0.9742 and 0.9601, respectively. Furthermore, case studies on three different diseases demonstrate that MSignVGAE, by leveraging the signed message, can effectively capture distinct feature distribution patterns in signed networks.

It is worth noting that the signed message propagation strategy designed in MSignVGAE only controls the information propagation process among nodes connected by different signed edges, without considering the differences among neighbors of nodes connected by the same type of edges. Thus, the utilization of signed message is still not fully optimized. In reality, the introduction of signed message can improve the performance ceiling of the microbe-disease association prediction task. Further exploration of information related to diseases and microbes can help complete the global distribution of microbe-disease associations. The relationship between diseases and microbes is highly complex, and both are intricately connected to the bridge of medications. Solely focusing on processing microbe-disease association data may overlook this information. The next focus should be on constructing various bridges that connect diseases and microbes, considering factors such as polysaccharide information that can simultaneously affect the states of both diseases and microbes.

Abbreviations

SGCN	Signed Graph Convolutional Network
MSignVGAE	Multi-scale Sign Variational Graph AutoEncoder
HMDAD	Human Microbe–Disease Association Database
GIP-D	Disease Gaussian interaction profile kernel similarity
DSS1	Disease semantic similarity
DSS2	Disease symptom similarity
GIP-M	Microbe Gaussian interaction profile kernel similarity
DFS1	Disease-based functional similarity
DFS2	Drug-based functional similarity
PMID	PubMed IDs

Acknowledgements

Not applicable.

Authors' contributions

All authors contributed to the article. HZ and LY conceived and designed this paper. HZ collected and analyzed the data. HZ, HH, and LY designed the experiments and analyzed the results. HZ drafted the paper. HZ, HH, and LY revised and edited the paper. All authors read and approved the final manuscript.

Funding

This research was funded by National Natural Science Foundation of China, grant number 62072353 and 62272065.

Availability of data and materials

The code of the model and datasets can be downloaded from GitHub (https://github.com/LiangYu-Xidian/MSignVGAE, https://doi.org/10.5281/zenodo.12789669). All data generated or analyzed during this study are included in this published article, its supplementary information files. and publicly available repositories.

For previously published datasets:

Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, Yang J, Kong W, Zhou X, Cui Q. An analysis of human microbe–disease associations. https://academic.oup.com/bib/-article/18/1/85/2562737?login=false#supplementary-data. (2016);

Janssens Y, Nielandt J, Bronselaer A, Debunne N, Verbeke F, Wynendaele E, Van Immerseel F, Vandewynckel Y-P, De Tré G, De Spiegeleer B. Disbiome database: linking the microbiome to disease. https://www.bmcmicrobiol.biomedcentral.com/-articles/10.1186/s12866-018-1197-5#Sec10. (2018);

Yao G, Zhang W, Yang M, Yang H, Wang J, Zhang H, Wei L, Xie Z, Li W. Microphenodb associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. http://www.liwzlab.cn/microphenodb/-#/download. (2020);

Skoufos G, Kardaras FS, Alexiou A, Kavakiotis I, Lambropoulou A, Kotsira V, Tastsoglou S, Hatzigeorgiou AG. Peryton: a manual collection of experimentally supported microbe-disease associations. https://dianalab.e-ce.uth.gr/peryton/-#/associations. (2021).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 23 April 2024 Accepted: 1 August 2024
Published online: 15 August 2024

References

1. Cénit M, Matzaraki V, Tigchelaar E, Zhernakova A. Rapidly expanding knowledge on the role of the gut microbiome in health and disease. Biochimica et Biophysica Acta -Molecular Basis of Disease. 2014;1842(10):1981–92.
2. Sommer F, Bäckhed F. The gut microbiota—masters of host development and physiology. Nat Rev Microbiol. 2013;11(4):227–38.
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
4. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.
5. Cross ML. Microbes versus microbes: immune signals generated by probiotic lactobacilli and their role in protection against microbial pathogens. FEMS Immunol Med Microbiol. 2002;34(4):245–53.
6. Henao-Mejia J, Elinav E, Thaiss CA, Licona-Limon P, Flavell RA. Role of the intestinal microbiome in liver disease. J Autoimmun. 2013;46:66–73.
7. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nat Methods. 2008;455(7216):1109–13.
8. Huang YJ, Boushey HA. The microbiome in asthma. J Allergy Clin Immunol. 2015;135(1):25–30.
9. Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer. 2013;13(11):800–12.
10. Feng J, Wu S, Yang H, Ai C, Qiao J, Xu J, Guo F. Microbe-bridged disease-metabolite associations identification by heterogeneous graph fusion. Brief Bioinform. 2022;23(6):bbac423.
11. Wang L, Yang X, Kuang L, Zhang Z, Zeng B, Chen Z. Graph convolutional neural network with multi-layer attention mechanism for predicting potential microbe-disease associations. Curr Bioinform. 2023;18(6):497–508.
12. Wang L, Li H, Wang Y, Tan Y, Chen Z, Pei T, Zou Q. MDADP: a webserver integrating database and prediction tools for microbe-disease associations. IEEE J Biomed Health Inform. 2022;26(7):3427–34.
13. McCoubrey LE, Gaisford S, Orlu M, Basit AW. Predicting drug-microbiome interactions with machine learning. Biotechnol Adv. 2022;54: 107797.
14. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature. 2019;570(7762):462–7.
15. Panebianco C, Andriulli A, Pazienza V. Pharmacomicrobiomics: exploiting the drug-microbiota interactions in anticancer therapies. Microbiome. 2018;6:1–13.
16. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, Brochado AR, Fernandez KC, Dose H, Mori H. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 2018;555(7698):623–8.
17. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, Feng J, Su R, Nakai K, Zou Q. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res. 2023;51(7):3017–29.

18. Wang L, Tan Y, Yang X, Kuang L, Ping P. Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. Brief Bioinform. 2022;23(3):bbac080.
19. Wen Z, Yan C, Duan G, Li S, Wu F-X, Wang J. A survey on predicting microbe-disease associations: biological data and computational methods. Brief Bioinform. 2021;22(3):bbaa157.
20. Grover, Aditya, and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 2016. 2016. p. 855–64.
21. Lei X, Wang Y. Predicting microbe-disease association by learning graph representations and rule-based inference on the heterogeneous network. Front Microbiol. 2020;11:579.
22. Peng L, Shen L, Liao L, Liu G, Zhou L. RNMFMDA: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. Front Microbiol. 2020;11:592430.
23. Xu D, Xu H, Zhang Y, Wang M, Chen W, Gao R. MDAKRLS: Predicting human microbe-disease association based on Kronecker regularized least squares and similarities. J Transl Med. 2021;19:1–12.
24. Long Y, Luo J, Zhang Y, Xia Y. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. Brief Bioinform. 2021;22(3):bbaa146.
25. Tao W, Liu Y, Lin X, Song B. Zeng XJBiB: prediction of multi-relational drug–gene interaction via dynamic hypergraph contrastive learning. Brief Bioinform. 2023;24(6):371.
26. Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018;34(23):4007–16.
27. Derr T, Ma Y, Tang J: Signed graph convolutional networks. In: 2018 IEEE International Conference on Data Mining (ICDM): 2018. IEEE: 929–934.
28. Huang J, Shen H, Hou L, Cheng X. Signed graph attention networks. In: Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28. Berlin, Heidelberg: Springer-Verlag; 2019. p. 566–77.
29. Li Y, Tian Y, Zhang J, Chang Y. Learning signed network embedding via graph attention. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34. 2020. p. 4772–9.
30. Li Y, Qu M, Tang J, Chang Y. Signed laplacian graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37. 2023. p. 4444–52.
31. Zeng X, Wang F, Luo Y, Kang S-G, Tang J, Lightstone FC, Fang EF, Cornell W, Nussinov R, Cheng F. Deep generative molecular design reshapes drug discovery. Cell Rep Med. 2022;4:100794.
32. Zhu H, Hao H, Yu L. Identifying disease-related microbes based on multiscale variational graph autoencoder embedding Wasserstein distance. BMC Biol. 2023;21(1):294.
33. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
34. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, vol. 2016. 2016. p. 785–94.
35. Zulfiqar H, Guo Z, Ahmad RM, Ahmed Z, Cai P, Chen X, Zhang Y, Lin H, Shi Z. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front Med. 2024;10:1291352.
36. Zou X, Ren L, Cai P, Zhang Y, Ding H, Deng K, Yu X, Lin H, Huang C. Accurately identifying hemagglutinin using sequence information and machine learning methods. Front Med (Lausanne). 2023;10:1281880.
37. Kim M-J, Lee Y-C, Kim S-W. TrustSGCN: learning trustworthiness on edge signs for effective signed graph convolutional networks. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, vol. 2023. 2023. p. 2451–5.
38. Liu Y, Zheng Y, Zhang D, Lee VC, Pan S. Beyond smoothing: unsupervised graph representation learning with edge heterophily discriminating. In: Proceedings of the AAAI conference on artificial intelligence, vol. 37. 2023. p. 4516–44.
39. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579.
40. Mancuso C, Santangelo R. Alzheimer's disease and gut microbiota modifications: the long way between preclinical studies and clinical evidence. Pharmacol Res. 2018;129:329–36.
41. Amitay EL, Krilaviciute A, Brenner H. Systematic review: gut microbiota in fecal samples and detection of colorectal neoplasms. Gut microbes. 2018;9(4):293–307.
42. Eckburg PB, Relman DA. The role of microbes in Crohn's disease. Clin Infect Dis. 2007;44(2):256–62.
43. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, Yang J, Kong W, Zhou X, Cui Q. An analysis of human microbe–disease associations. Brief Bioinform. 2017;18(1):85–97.
44. Janssens Y, Nielandt J, Bronselaer A, Debunne N, Verbeke F, Wynendaele E, Van Immerseel F, Vandewynckel Y-P, De Tré G, De Spiegeleer B. Disbiome database: linking the microbiome to disease. BMC Microbiol. 2018;18(1):1–6.
45. Yao G, Zhang W, Yang M, Yang H, Wang J, Zhang H, Wei L, Xie Z, Li W. Microphenodb associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. Genomics, Proteomics Bioinform. 2020;18(6):760–72.
46. Skoufos G, Kardaras FS, Alexiou A, Kavakiotis I, Lambropoulou A, Kotsira V, Tastsoglou S, Hatzigeorgiou AG. Peryton: a manual collection of experimentally supported microbe-disease associations. Nucleic Acids Res. 2021;49(D1):D1328–33.
47. Zhou X, Menche J, Barabási A-L, Sharma A. Human symptoms–disease network. Nat Commun. 2014;5(1):4212.
48. Chen X, Huang Y-A, You Z-H, Yan G-Y, Wang X-S. A novel approach based on KATZ measure to predict associations of human microbiota with noninfectious diseases. Bioinformatics. 2017;33(5):733–9.
49. Deng L, Huang Y, Liu X, Liu H. Graph 2MDA: a multi-modal variational graph embedding model for predicting microbe–drug associations. Bioinformatics. 2022;38(4):1118–25.
50. Ding Y, Lei X, Liao B, Wu F-X. Predicting mirna-disease associations based on multi-view variational graph auto-encoder with matrix factorization. IEEE J Biomed Health Inform. 2021;26(1):446–57.
51. Li H, Liu B. BioSeq-Diabolo: biological sequence similarity analysis using Diabolo. PLoS Comput Biol. 2023;19(6):e1011214.
52. Li H, Pang Y, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. Nucleic Acids Res. 2021;49(22):e129.
53. Ai C, Yang H, Ding Y, Tang J, Guo F. Low rank matrix factorization algorithm based on multi-graph regularization for detecting drug-disease association. Ieee-Acm Transact Comput Biol Bioinform. 2023;20(5):3033–43.
54. Singh R, Chen Y. Signed graph neural networks: a frequency perspective. arXiv preprint 2022, arXiv:2208.07323.
55. Guo Z, Wang F, Yao K, Liang J, Wang Z. Multi-scale variational graph autoencoder for link prediction. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, vol. 2022. 2022. p. 334–42.
56. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint 2014, arXiv:1412.6980.
57. Yang H, Luo YM, Ma CY, Zhang TY, Zhou T, Ren XL, He XL, Deng KJ, Yan D, Tang H, et al. A gender specific risk assessment of coronary heart disease based on physical examination data. NPJ Digit Med. 2023;6(1):136.
58. Yang H, Luo Y, Ren X, Wu M, He X, Peng B, Deng K, Yan D, Tang H, Lin H. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. Inform Fusion. 2021;75:140–9.
59. Abbas Z, Rehman MU, Tayara H, Zou Q, Chong KT. XGBoost framework with feature selection for the prediction of RNA N5-methylcytosine sites. Mol Ther. 2023;31(8):2543–51.
60. Wang Y, Zhai, Y., Ding, Y., Zou, Q: SBSM-Pro: support bio-sequence machine for proteins. *arXiv preprint* 2023:arXiv:2308.10275.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.