Gut Pathogens

REVIEW

Open Access

In-silico computational approaches to study microbiota impacts on diseases and pharmacotherapy

Hassan Shokri Garjan[1], Yadollah Omidi[2], Mehdi Poursheikhali Asghari[3] and Reza Ferdousi[1*]

Abstract

Microorganisms have been linked to a variety of critical human disease, thanks to advances in sequencing technology and microbiology. The growing recognition of human microbe–disease relationships provides crucial insights into the underlying disease process from the perspective of pathogens, which is extremely useful for pathogenesis research, early diagnosis, and precision medicine and therapy. Microbe-based analysis in terms of diseases and related drug discovery can predict new connections/mechanisms and provide new concepts. These phenomena have been studied via various in-silico computational approaches. This review aims to elaborate on the computational works conducted on the microbe–disease and microbe–drug topics, discuss the computational model approaches used for predicting associations and provide comprehensive information on the related databases. Finally, we discussed potential prospects and obstacles in this field of study, while also outlining some recommendations for further enhancing predictive capabilities.

Keywords Microbiota, Microbiome, Microbe–disease prediction, Microbe–disease similarity, Similarity calculation method, Microbe–disease associations

Introduction

The gut microbiota is a collection of microorganisms that live in the mammalian gastrointestinal tract (GIT). This microbial population has a host-specific composition that changes with time and is susceptible to both exogenous and endogenous alterations [1]. Unlike the host genomic profile, the gut microbiome is a changeable environment that can be achieved with probiotics, prebiotics, nutrition, and community replacement techniques like fecal microbiota transplant [2]. The majority of these microbes live in the gastrointestinal tract, most abundant in the distal portion of the intestine. They synthesize essential amino acids, vitamins, and non-digestible components to aid in nutritional processes. Combined with host genetics, metabolic phenotypes can have a profound impact [3–7].

Environmentally, geography, diet, aging, the use of drugs and antibiotics, stress, and diseases can affect the gut microbiota [7–11]. A balance of microbiota is believed to protect the host body from physiological disorders. Based on a plethora of compelling evidence, there might be a correlation between the emergence of diseases and the changes in the composition and amount of microbiome in the body [12, 13]. Evidence suggests that changes in gut microbiota are linked to a variety of diseases and immune and metabolic dysfunctions, including hypertension, heart attack, myocardial

*Correspondence:
Reza Ferdousi
ferdousi.r@gmail.com
[1] Department of Health Information Technology, School of Management and Medical Informatics, Tabriz University of Medical Sciences, Tabriz, Iran
[2] Department of Pharmaceutical Sciences, Nova Southeastern University, College of Pharmacy, Fort Lauderdale, FL, USA
[3] Biotechnology Research Center, Tabriz University of Medical Sciences, Tabriz, Iran

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

infarction, stroke, coronary artery disease, diabetes, and chronic kidney disease [9, 14, 15]. Additionally, the human GIT microbiota, as a predictor of human health and the therapeutic response, can influence the reaction(s) of the body to a variety of treatments, from dietary and lifestyle modifications to drugs and surgical procedures [2]. It is deemed that the intestinal microbiota interacts with almost all human cells and is considered a major factor in host metabolism and also a new source of therapy [16].

A beneficial commensal or symbiotic relationship between the human and the microbiota of the body is proved by the advances in sequencing technology and recent bioinformatics discoveries [17, 18]. Meanwhile, some researchers develop and apply computer techniques to identify the effects of microbes on human disease. For instance, Coelho et al. have suggested a computational technique that takes into account the interaction between microbial and human proteins to anticipate the effect of microbial proteins on human biological processes [19]. The Human Microbiome Project started in 2007, is another well-known instance of a microbe project [20]. Finding Microbe–Disease Associations (MDA) might be extremely beneficial in areas that deal with diseases, such as medications and pathogenic genes [21]. The gut microbiota is now recognized as being responsible for adjusting many physiological functions of the host [25, 26]. Additionally, the identification of microbe–disease relationships offers several insights into the pathophysiology of disease. Notable computational techniques have recently been developed to investigate the influence of microbiota on human disease, and medications [30–35]. In this article, we strived to fully inspect the computational methods for predicting microbial disease associations, which can be divided into six categories:

I. Path-based methods: Path-based methods allow predictions in heterogeneous networks by calculating path-based scores between microbe nodes and disease nodes.
II. Methods based on Random Walks: A walker walks in a transfer likelihood network made up of microbe and disease nodes at random. These strategies look for a probable association by calculating the likelihood of a random walker completing a path that starts with a node on one side of the association and ends with a node on the other.
III. Bipartite Local Models (BLMs): Such methods compute Microbe–disease association (MDA) forecast scores from two viewpoints of diseases and microbes. The collective prediction scores on both sides are used to determine the final prediction ranking.
IV. Matrix factorization approaches: an interaction matrix is factorized into two low-dimensional matrices, one representing disease features and the other representing microbe ones. The final projected matrix is the sum of two low-dimensional matrices.
V. Machine learning-based: The machine learning-based method uses fewer parameters that can save time and achieve strong performance.
VI. Network-based methods: Network-based methods have used Graph Attention Network (GAT), MLP layers to predict new connections, automatic learning of a nonlinear function, and so on to predict new connections.
VII. Other methods: Certain methods may not be sorted into the groups above, but they are grouped as "other methods."

Additionally, various drugs can alter the structure and composition of the gut microbiome and thus change its biological function, such as the ability to metabolize. On the other hand, metabolism and drug outcomes may be influenced by microbial metabolic processes and their metabolites. Understanding the mutual relationship between drugs and microbiomes, as well as how it affects drug clinical outcomes, paves the way for next-generation interventions to reduce disease complications [22]. Little is known about the impact of the microbial gene pool on medications prescribed in various areas of the human body, as well as the impact of microbiome modifications on drug destiny, behavior, toxicity, and therefore a human reaction to care [23]. Remarkably, recent research has shown a solid connection between the microbiota and the pharmacological effects of chemotherapy [24] and immunotherapy [25, 26]. The microbial diversity in the body is intriguingly reduced, in large part due to the interaction of chemical drugs with the host immune system. As a result, the effect of such drug molecules might be decreased, and other consequences may occur too. The human microbiome, particularly the gut microbiome, improves the efficacy of chemo-drugs through digestion, enzyme degradation, ecological variations, and immunomodulatory. A recent study has taken advantage of the microbiome's role in shaping the effectiveness and toxicity of these chemotherapy agents [27]. The relationship between gut microbes and the currently used non-antibiotic drugs seems to be very complicated, in which drugs can affect the gut microbiome's makeup, and the gut microbiome can also enzymatically alter the drugs [28]. The individual's reaction to the medication may alter bioavailability, bioactivity, and/or toxicity, a phenomenon known as" pharmaco-microbiome" [29–31].

Variations in the human microbiome(i.e., the synthesis of human-associated microbial species and their genomes) might impact medication disposition, behavior, and toxicity, according to the concepts of pharmacy and toxico-microbiome [32].

The gut microbiota influences medication and cenobitic metabolism in both overt and indirect ways, which may affect effectiveness and toxicity [33]. Advances in gut microbiota modeling and research will expand our understanding of their function in health and disease, allowing for the customization of current and prospective medicinal and prophylactic modalities [34]. Moreover, various human infectious diseases are caused by an imbalance in microbial communities [20, 35]. GIT microbes also play an important role as a therapeutic target in precision medicine and modulation of drug activity or toxicity [36], while their diversity and function can be altered by drugs [37].

Besides, with the increasing emergence of drug-resistant microbes, it is necessary to identify microbial-pharmacological associations in very large sizes [36]. For this purpose, several models have been proposed and designed to identify the association of medicinal microbes, including the Ensembling graph attention networks for predicting human microbe–drug association [38]. Based on the heterogeneous network embedding representation, the association mining method was used to detect microbe–drug interactions [39]. To compute potent associations between microbe and drug, Zou et al. developed a method based on the KATZ measure [40]. In another study, Long et al. proposed a computational Method based on a novel Graph Convolutional Network (GCN) framework for predicting before-mentioned associations [41]. These approaches can be summarized as follows:

- Neural Network (GCN): A neural network is a collection of algorithms that attempts to understand underlying associations in a set of input data using a procedure that mimics how the human brain works.
- Assembling a graph with attention function: In various graphs, each node (e.g., microbes, and drugs) can contain a variety of semantic knowledge. The attention function at the diagram level is used to efficiently collect node embeds from input diagrams, merge information, and remove noise from various diagrams.
- Heterogeneous network embedding representation: In this method, by combining Metapath2Vec with the recommendation of a two-part network, a heterogeneous embedded network demonstration framework is used to predict the association of microbes and drugs. To improve the prediction accuracy, the proposed bias bipartite network Embedding (BiNE) forecasting algorithm has been created and used.
- KATZ measurements: In this method, most of the heterogeneous network of medicinal microbes is created based on two similar networks and known connections of medicinal microbes. Based on these networks and KATZ measures, the process of predicting the potential relationships between drugs and microbes was performed. The human gut microbiota, as a predictor of human health and therapeutic response, is shown in Fig. 1.

Prediction of microbiome association with drug and disease

In latest years, with the speedy development of strategies in bioinformatics and life science, a massive quantity of biomedical information has been amassed, based on which researchers have evolved numerous computational procedures to discover potential associations between human microbes, drugs and diseases. This article offers a thorough analysis of current developments in identifying possible relationships among microbes, drugs and diseases using biological data and computational models.

Drug–microbe association

It should be noted that drugs can change the species diversity and function of microbial communities [36], and the number of drug-resistant bacteria is growing. In this line,

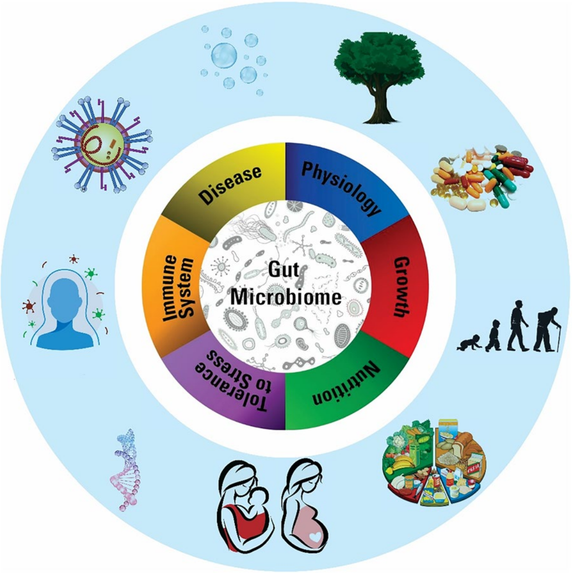


Fig. 1 The human gut microbiota is a predictor of human health and therapeutic response. The gut microbiota influences a person's reaction to a variety of treatments, from dietary and lifestyle modifications to drugs and surgical procedures

microorganisms play a vital role in lowering the medications' adverse reactions. Collectively, there is an urgent need to recognize the possible pharmaceutical-microbial associations [37]. In the rest of this section, the studies related to the prediction of microbe–drug relationships in the literature were reviewed.

Graph convolutional network (GCN)

Long, Y. et al. used various sources of biomedical information and created several networks (diagrams) for microbes and drugs. Then, they developed a novel ensemble framework of graph attention networks with a hierarchical attention mechanism for microbe–drug association prediction from the constructed multiple microbe–drug graphs, denoted as Ensembling graph attention networks for human microbe–drug association prediction (EGATMDA). Specifically, for each input graph, a graph convolution network is designed according to the node surface to learn to embed the nodes (e.g., microbes and drugs). To effectively integrate node embeds from multiple input diagrams, graph-level attention has been implemented to learn the importance of different input diagrams [38].

Graph attention networks

The proposed Graph Convolutional Network (GCN) based framework for predicting human Microbe–drug Associations (MDA), named GCNMDA is a convolutional neural network-based model for predicting drug-microbe interactions. Initially, a heterogeneous network is built to combine microbial gene information, drug chemical information, and microbe–drug interactions. Later, an RWR-based preprocessing mechanism is designed to extract effective properties. Finally, a CRF layer is generated in the GCN to enhance the learning of node representation for drugs and microbes so that similar nodes have similar representations. A layer of the CRF attention mechanism is designed to accurately collect representations from neighbors [42].

Heterogeneous network embedding representation

Adjacency matrix In this approach, the information obtained from the confirmed experimental results related to human microbe–disease (microbe–drug) is extracted from the corresponding databases for microbe–disease (microbe–drug) associations. Then, an adjacency matrix $A \in R^{nd*nm}$ is created (nd and nm show the number of diseases (drugs) and the number of microbes, respectively) as follows:

$$aij = \begin{cases} 1 & if\ association\ between\ disease(drug)\ di\ and\ microbe\ mi \\ 0, & else \end{cases}$$

Similarity calculation and heterogeneous network Various computational methods, that have been designed and proposed to predict microbe–disease (microbe–drug) data, are mentioned in the previous sections. The approaches can be classified into two groups: (i) those that use known disease-microbe relationships to calculate microbe–disease similarity, and (ii) those that use extra data.

In a method for determining similarity based on microbe–disease associations, the adjacent matrix $A \in R^{nd \times nm}$ is used as the input, and the similarity matrix between microbial $S_m \in R^{nm \times nm}$ and the similarity between $S_d \in R^{nd \times nd}$ disease is used as the output. The similarity calculation methods are the same for diseases-microbes (drugs) and the methods include Gaussian interaction profile kernel similarity [43]. The following approaches can be implemented:

Cosine similarity: In Euclidean space, the cosine similarity measures the cosine of the angle between two interaction profiles. Having capitalized on this approach, a few studies were able to obtain the microbe and disease similarity matrix [21, 44].

Spearman correlation similarity: Spearman correlation coefficients as similarity ratings are calculated using sequences of positions or time points of pairwise microbes [45].

In a recent study, Wang et al. advanced a gene-based disease association approach based on neighbor-dependent similarity estimation. In most studies, after creating similar networks for diseases and microbes, researchers have used known microbe–disease associations through databases to construct the proposed models [46].

Two researchers have proposed a biased two-way network algorithm to predict the most likely microbe–drug relationships and increase the accuracy of the proposed model. Heterogeneous Network Embedding Representation framework for Microbe Drug Association (HNERMDA) is based on the representation of an embedded heterogeneous network via metapath-2vec and the recommendation of a two-part network. To build heterogeneous networks, they capitalized on interactions between microbes and drugs, such as drug-microbe interactions [39].

KATZ measurements

Using known drug-microbe associations, a microbe similarity network is constructed by calculating the GIP core similarity of microbes. Due to the two similar networks and similar connections of known medicinal microbes, a heterogeneous network of medicinal microbes is created. An HMDAKATZ model is designed to predict drug–microbe communication [40].

Multi-modal variational graph embedding

A multi-modal variational graph embedding model for predicting microbe–drug associations (Graph2MDA) is a new technology that uses a graph autoencoder to predict microbe–drug interactions variational graph auto encoder (VGAE). Created multi-modal attributed graphs based on molecular structures, microbe genetic sequences, and function annotations of bacteria and pharmaceuticals. A deep neural network classifier was used to predict microbe–drug relationships [47]. Figure 2 represents the architecture of predicting microbe–drug relationships using a convolutional neural network model.

Recruited datasets and approaches for prediction of microbe–drug associations

Previous studies on the microbe–drug relationship have used a variety of data.

Table 1 lists the data used to predict microbe–drug based on the information we reviewed.

In addition, different approaches for predicting the relationship between microbe–drug are summarized in Table 2.

Comparison and application of models to each other

Since predicting the interactions between microbes and drugs is a new field of study, few computational approaches have been proposed for this critical task. The various approaches for link prediction problems in the field of bioinformatics and the existing techniques for microbe–drug interactions are compared [38, 48–50]. The Graph2MDA model had the highest AUC value, followed by LAGCN, while NTSHMDA had the lowest AUC value. Deep learning-based methods frequently outperform more traditional machine learning-based ones. The more effective method provides the following benefit over other models: Using multimodal feature graphs based on ontological information, multiple similarities between microbes and drugs, and their known relationships, methods may fully use many different sorts

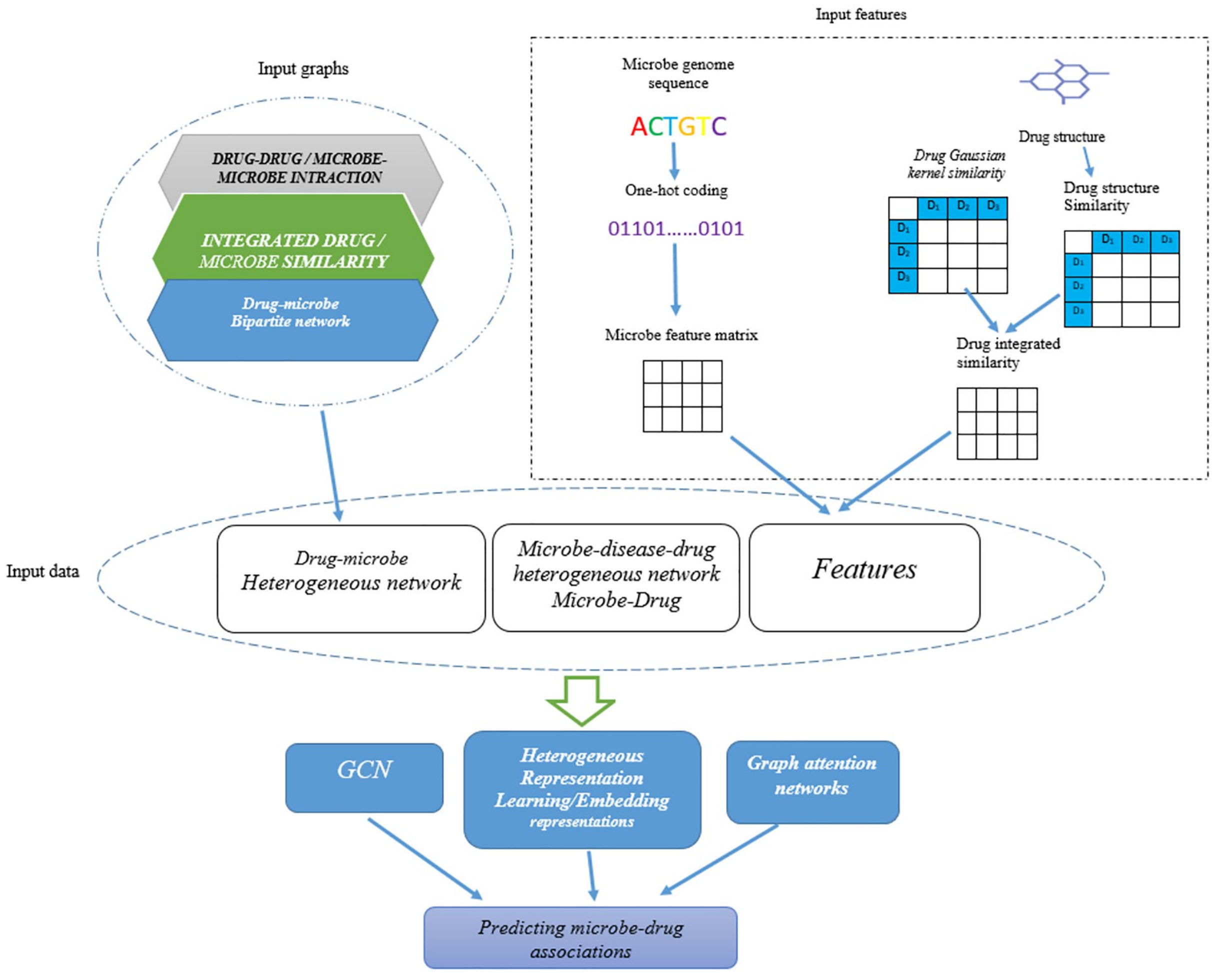

Input features
Input graphs
DRUG-DRUG / MICROBE-MICROBE INTRACTION
INTEGRATED DRUG / MICROBE SIMILARITY
Drug-microbe Bipartite network
Microbe genome sequence
ACTGTC
One-hot coding
01101……0101
Microbe feature matrix
Drug structure
Drug Gaussian kernel similarity
Drug structure Similarity
D₁
D₂
D₃
Drug integrated similarity
Input data
Drug-microbe Heterogeneous network
Microbe-disease-drug heterogeneous network Microbe-Drug
Features
GCN
Heterogeneous Representation Learning/Embedding representations
Graph attention networks
Predicting microbe-drug associations

Fig. 2 The architecture of predicting microbe–drug relationships using a convolutional neural network model

Table 1 A list of all the data that fed into the microbe–drug association prediction

Data	Source	Original from	Size/coverage in MDAD	The process of similarity
MDAD	A database that gathers various types of data from different sources to discover the interaction between microbes and drugs and to promote the detection of antimicrobial drugs and the use of microbiomes in medicine (http://chengroup.cumt.edu.cn/MDAD)	Multiple prescription libraries and associated papers yielded 5,055 entries representing 1,388 medicines and 180 microbes	There are 5,055 entries in this collection of clinical or experimental support collaborations, including 180 microbes and 1,388 medications	After deleting duplicate information at the end of 2470, the relationship between 1373 drugs and 173 microbes has been used for the similarity process
Drug-Drug Interaction	DrugBank is one-of-a-kind bioinformatics and chemical informatics repository that integrates specific drug data with extensive drug target data. (https://www.drugbank.ca/releases/latest)	The latest version includes 14,522 drug imports, including 2,683 approved small-molecule drugs, 1,644 biology-approved drugs, 131 nutrients, and more than 6,654 experimental drugs	About 5249 extra protein sequences are associated with these drug inputs. Each input contains more than 200 pieces of data, half of which is dedicated to pharmaceuticals/Half of the data is related to chemical data, while the other half is related to target drug or protein data	Inputs such as drugs in MDAD Search have finally created a network of drug and drug interactions involving 5587 interactions with 181 drugs
Drug Structural Similarity Matrix a Biofilm Drug Virus	Chemical structure search servers are for network analysis and clarifying the relationship between the genomic and chemical importance of metabolic pathways: (http://www.genome.jp/tools/simcomp) The 'a Biofilm' (http://bioinfo.imtech.res.in/manojk/abiofilm/): There are three sub-categories in this category, including data visualization, a database, and a prediction module Lists the activities and stages of development of 118 drugs that target 83 human viruses (https://drugvirus.info/tech_doc/)	There are 5027 interactions between 1720 medications and 140 microorganisms in the database of anti-biofilm compounds, which includes gram-negative, gram-positive, and fungal microorganisms The database allows users to explore virus-BSAA (Broad-spectrum antiviral agents) interactions in real time. DrugVirus.info is a free tool that includes a feedback form on its website		SIMCOMP is a graph-based method for finding the maximum isomorphism of a common sub graph by finding the maximum number of clicks in the correlation diagram, and SUBCOMP is a broad method for solving the problem of sub graph isoforms
Human Drug Virus Database (HDVD)	• They devoted significant resources to collecting drug-virus interaction entries from the literature to create the Human Drug Virus Database (HDVD).SCPMF (similarity constrained probabilistic matrix factorization) is a novel technique for identifying new medication-virus interactions for therapeutic repurposing [68]			

Table 2 Different methods to predict microbe–drug association

Category	Method	Description	Refs.
Graph Convolutional Network (GCN)	GCNMDA	A conditional random context (CRF) and a CRF layer focus function in the hidden GCN layer are used to ensure that the same nodes have the same representations	[61]
Ensembling graph attention	EGATMDA	To learn embedded nodes for microbes and drugs, a convolutional graph network is built at the node level for each input diagram	[55]
Heterogeneous network embedding representation	HNERMDA	Metapath2vec has developed a heterogeneous network display learning approach for learning low-embedded microbe and drug displays in this context	[56]
Multi-modal variational graph embedding	Graph2MDA	A graph with variations A deep neural network classifier was used to predict microbe–drug relationships after an auto encoder was trained to learn the informative and interpretable latent representations of each node and the whole graph	[67]
Based on KATZ measurements	KATZ	By bringing the chemical structures together and the similarity of the nucleus, they have created the Gaussian interaction profile of the drug unification network	[57]

of features and links. Additionally, by incorporating the topological structure into multimodal feature networks, the impact of the cold start problem is lessened. Potentially mitigate the effect of similar noises [38].

Microbe–disease association

According to some new research [51], microbes are being increasingly linked to human pathogens. Disease-related microbe research aims to understand disease processes and the creation of novel diagnostic and therapeutic methods. Many theoretical models for predicting disease-causing microbes have been suggested. In the continuation of this section, we review the studies related to the prediction of microbe disease relationships that exist in the literature.

Path-based methods

Weighted meta-graph-based model on heterogeneous information network (WMGHMDA) have been presented to predict the relationship between diseases and microbes. Path-based approaches examine indirect pathways across networks, which often evaluate the weight of a prospective route as the score of unknown relation. The Meta-Graph search algorithm is run on the heterogeneous network to count the meta-weight patterns of each disease-microbe pair. Summing the contribution values of the related weighted Meta graphs yields the likelihood score for each pair of disease-Microbe [52].

BWNMHMDA (Bidirectional Weighted Network model Human Microbe–disease Association Prediction) is a new method for predicting the microbe–disease association based on the Bidirectional Weight Network. The main idea of this model is to produce a bidirectional disease-microbe communication network that converts them into matrices to compute the probability of correlation. It can be achieved by assigning weight to nodes and edges in the integrated network using the similarity of the Gaussian interaction profile kernel [53].

The PBHMDA (Path-Based Human Microbe–disease Association Prediction) proposes a new path-based prediction model for inferring potential microbe–disease associations. It is based on the main similarity of Gaussian interaction profiles for diseases and Gaussian interactions between microbes. A special depth-first search algorithm was designed in the model to ensure no duplicate nodes were found [54].

The KATZ measurement model was proposed to predict the Human Microbe–disease Association (KATZHMDA) Researchers combined the number of walks and their distances as an appropriate measure index for measuring the possible interaction likelihood between microbes and diseases. It is based on the graph constructed by the established microbe–disease association network, microbe similarity network, and disease similarity network [48].

By integrating several data sources and path-based HeteSim scores, Fan et al. developed a new method for predicting disease-microbe Multiple Data sources and Path-based HeteSim scores for Human Microbe–disease Associations (MDPH_HMDA) communication. The similarity of microbes was calculated by combining microbial functional scores and Gaussian core profile similarity. The similarity of the disease pairs was calculated using the similarity scores based on the symptoms. The HeteSim method has been used to obtain the relevance score and normalized measurement from each disease-microbe pair [55].

Random walk methods

For iterative walking, random walk methods use a graph-based transfer likelihood matrix. Niu et al. made a higher-order hyper graph sample to accurately determine the intrinsic association between microbes and human diseases. They develop a model based on the random walk on hypergraph for microbe–disease association prediction (RWHMDA). They ranked all-volunteer microbes for every perused human disease. Hypergraphs can efficiently mitigate data loss occurring in the normal graph methodology. For the generated hypergraph, the core similarity of the Gaussian interaction profile, random walk, and integration of known microbe–disease associations from the HMDAD database was performed [56].

A heterogeneous network by combining the Gaussian interaction profile microbial similarity network and the Gaussian interaction profile disease similarity network has been produced by known networks of microbe–disease associations. Then, a novel way for predicting the future microbial and disease relationships based on extensive optimized random walking was announced by introducing network topological similarity (NTSHMDA) [49].

Zou et al. have combined the microbial similarity network and the disease similarity network to generate a heterogeneous network. A two-random walk algorithm was implemented on the network generated by the Gaussian interaction profile's similarity and logistic transformation. A novel computational model to predict potential microbe–disease associations by bi random walk on the heterogeneous network. Developed a new computational model for predicting potential human microbe–disease associations by bi random walk in heterogeneous network (BiRWHMDA) [57].

Zhang et al. proposed the bi-direction similarity integration label propagation (BDSILP) method for predicting microbe–disease associations. Using the Mesh, the semantic similarity of the disease and the functional similarity of the microbes were calculated. With the help of integrated disease similarity and integrated microbial similarity, they have produced two graphs. And BDSILP does the label propagation on the graphs to score the pairs of disease-microbe. BDSILP accepts the weighted mean of their scores as final predictions [58].

The symptom-based likeness is calculated by the concurrence of diseases and the term symptoms. After calculating the similarity of the core of the Gaussian interaction profile of microbes based on known microbial disease associations, the similarity with the logistic function was obtained. Using the Similarity Network Fusion (SNF) method with similarity based on symptoms and the similarity of the core, the Gaussian interaction profile was calculated according to the known microbe–disease associations of the disease network. The two networks created for microbes and disease have been combined by well-known microbe–disease associations and used by BRWMDA (Bi-random walk microbe–disease associations) to predict potential new microbe–disease relationships through random walking with different stages in microbial and disease networks [59].

After extracting information about the disease and germs, microbial networks were built using Spearman, and the disease network was generated based on the symptoms. Then, by combining the networks formed, a heterogeneous network of disease microbes is formed. Shen et al. developed the random walk with a restart algorithm for the heterogeneous network, using the goal disease and corresponding microbes as seed nodes. They employed this algorithm to reveal the latent relationship between diseases and microbes [60].

A team of researchers has proposed a new model of extended random walking with restart optimized by Particle Swarm Optimization (PRWHMDA) based on human microbe–disease associations. Wu et al. used cosine to calculate the similarity of diseases and microbes. Then, by combining networks, they formed a heterogeneous interconnected network. They introduced the RWR method to obtain strong communications [44].

Wang et al. have proposed a novel computational model based on the bidirectional label propagation to predict potential human microbe–disease associations (NBLPIHMDA). The Gaussian interaction profile kernel similarity was applied to measure the disease similarity matrix along with the microbe similarity matrix. The edge weights of nodes in these two networks were determined. Bidirectional mark dissemination was used to achieve the association score matrix between diseases and microbes [61].

Using known connections from microbial network databases, disease networks and microbe–disease networks were created. A heterogeneous network was constructed using known microbe–disease associations from the database, the microbial network, and the disease network. Wang et al. then predicted novel microbe–disease associations by a new method called the double ended restart random walk human microbe–disease association model (DRWHMDA) implemented on the interconnection network [62].

Bipartite local models

Fundamentally, the bipartite local models work independently on both sides of a microbe–disease pair and can be combined to provide a conclusive prediction outcome. These approaches are capable of making independent observations on both the microbe and the disease fronts. The final scoring matrix is based on the combination of the probability scores related to user-based and case-based collaborative filtering [63].

Zou et al. proposed a model using a combination of a neighborhood-based prediction model and a graph-based recommendation model for human microbe–disease association (called NGRHMDA). The graph-based prediction model presents a two-step diffusion approach on the microbe–disease bipartite graph. Two new integrated adjacent matrices have been developed based on the similarity of symptom-based diseases and on the similarity of Gaussian-based microbes to consider microbial and disease similarities [64].

The core similarity of the Gaussian interaction profile for germs and disease was extracted from the microbe–disease linkage network. Then, constructing and minimizing the cost function for optimal classifiers in microbe and disease space turned it into an integrated classification. A semi-supervised computational model_Laplacian Regularized Least Squares for Human Microbe–Disease Association (called LRLSHMDA) was proposed by Wang et al. to predict disease-microbe relationships [65]. Based on known microbe–disease communication networks, a heterogeneous network was created from the HMDAD database for the main similarity of disease Gaussian interaction profiles and microbe Gaussian interaction profiles. Then, Bao et al. planned the Network Consistency Projection for Human Microbe–disease Association prediction model (called NCPHMDA) to discover potential disease-microbe associations [66]. The KATZBNRA model, like the KATZHMDA, was designed by Li et al. using the KATZ criterion and the core similarity of the Gaussian interaction profile for diseases and microbes based on the known associations. In addition, they utilized a bipartite (two-way) Network Recommendation (BNR) algorithm to increase the prediction accuracy more than KATZHMDA [67].

Matrix factorization methods

The theory behind matrix factorization is that the input matrix decomposes into two low-dimensional matrices and the product of the two low-dimensional matrices approximates the input matrix [68, 69]. Wu et al. discovered disease characteristics by combining two similarities based on the Gaussian kernel and one based on symptoms. The microbial properties have also been calculated using the similarity of the Gaussian kernel. They presented a computational model using matrix completion to predict the association of the human microbe–disease profile (called MHMDA) [70]. Chen et al. introduced a method for predicting microbe–disease associations based on the Kernelized Bayesian Matrix Factorization (KBMF), which is dependent on the Gaussian interaction profile kernel similarity for microbes and diseases [71].To compute the microbial similarity and similarity of the disease, Liu et al. used the similarity of the core of the Gaussian interaction profile and applied logical functions to adjust the similarity of the disease. Based on the known microbe–disease associations, they suggested a model for predicting microbial disease associations using the regular non-negative matrix factorization chart (NMFMDA) [72].

By merging the known disease-microbe associations and the similarity of the core of the Gaussian interaction profile, Shen and his colleagues offered a Collaborative Matrix Factorization for Human Microbe–disease Association Prediction (CMFHMDA) model [73].

For the prediction of human microbe–disease associations, a novel predictive model of graph regularized non-negative matrix factorization (called GRNMFHMDA) was developed by He et al. Microbe and disease similarity were initially calculated using symptom-based disease similarity and Gaussian interaction profile kernel similarity for microbes and diseases, respectively. To prevent a negative effect on prediction results, a preprocessing phase was used in which unknown microbe–disease pairs were given associated probability scores. Finally, a graph-regularized non-negative matrix factorization method was employed to concurrently determine the possible correlations with all diseases [74]. Qu et al. introduced a statistical model of matrix decomposition and label propagation for the Human Microbe–disease Association prediction (so-called MDLPHMDA) by integrating proven microbe–disease associations obtained from the HMDAD database, disease symptom similarity, and Gaussian interaction profile kernel similarity for microbes and diseases. Using the spare learning method (SLM) on the original association details derived from HMDAD, a new adjacency matrix of microbe–disease associations was developed, and possible microbe–disease associations were further predicted using the label propagation algorithm (LPA) [75]. A Deep Matrix Factorization Prediction (DMFMDA) model has been proposed by Liu et al. to predict the associations between microbes and diseases that do not require microbial and disease-like networks and is based on deep neural networks, which combine the linear modeling advantages of matrix factorization with the non-linear modeling advantages of a multi-layer perceptron [76].

Network based methods

Graph attention networks Long et al. present a new graph-attention network-based model for microbe–disease association prediction (called GATMDA) in a bipartite network, combining inductive matrix completion (IMC). Researchers used functional similarities of microbes, functional similarities of diseases, and Gaussian kernel similarities to obtain comprehensive specifications for microbes and diseases. Graphic Attention Networks (GAT) then introduced a GAT criterion for learning to display nodes using talking heads, which helps maintain a more informative display model [77].

Liu et al. proposed a multi-component Graph Attention Network based system to predict microbe -disease association (MGATMDA). By using a node-level attention mechanism, the decomposer first decomposes the edges in a bipartite network to discover the latent components. The combiner then automatically reassembles these hidden parts to provide a coherent embedding for component-level attention prediction. Finally, a fully linked network is employed to forecast known and unknown connections between bacteria and diseases [78].

Models based on neural networks Using the similarity of microbial classification, the similarity of microbial interaction characteristics and disease interaction, semantic similarities and disease symptoms, and known disease and microbial associations, Ma et al. have developed a new method (NinimHMDA) based on neural integration of neighborhood information in a multiplex heterogeneous network (MHEN)for different types of human microbe–disease association prediction [79]. Li et al. proposed a new back-propagation neural network model to predict microbial-disease association (BPNNHMDA). The model input is a matrix of known microbe–disease associations, and its output is a matrix of potential microbe–disease association probabilities. An activation function is built based on the hyperbolic tangent function to activate the hidden and output layers. The Gaussian interaction profile core for microbes has been employed to improve binding weights and increase training speed [80].

Network consistency projection and multi-data integration

Then Fan et al. combined the matrix created for microbes and diseases with the linear network integration method. Get an integrated similarity matrix for diseases and microbes, and by integrating this matrix, network cohesion prediction was created. Disease-microbe associations were detected by predicting network cohesion and analyzing privileges extracted from them. Human Microbe–Disease Associations Prediction (HMDA-Pred) is a network-based computational method that connects multiple similarity networks to an integrated linear network method and predicts the association of disease-related microbes based on the Network Consistency Projection (NCP) algorithm [81].

Link propagation based on node information

PENG et al. have proposed a computational model of node information-based link Propagation for human microbe–disease association prediction (LPHMDA) to prioritize disease-associated microbes. Using well-known associations between disease-causing microbes and similarities between them, the Gaussian interaction profile of the matrix has created a likeness for microbes. They have formed a disease similarity matrix by combining the symptoms of the disease [82].

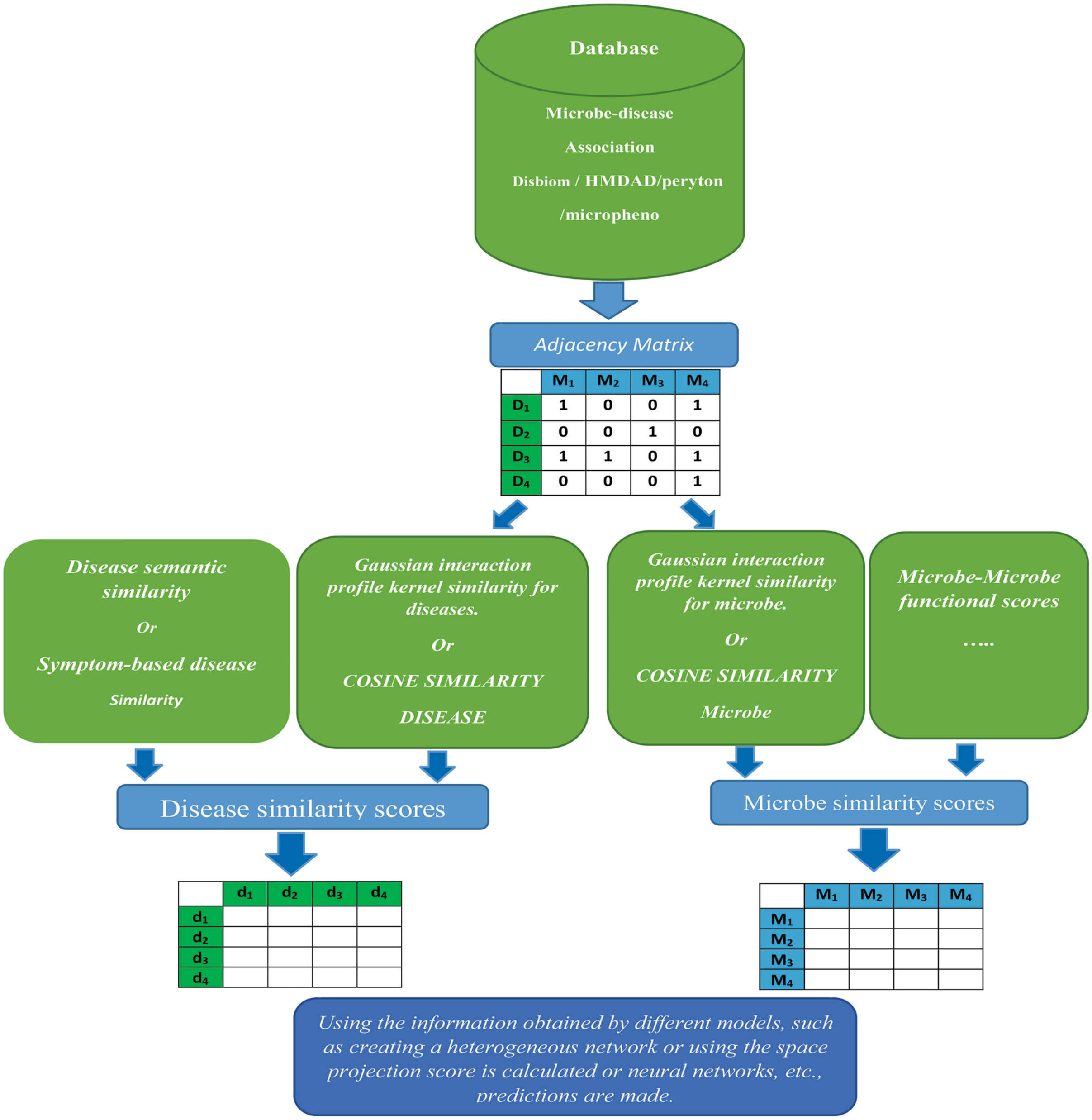
Machine learning-based

Xu et al. proposed a new computational method based on the Kronecker regularized least squares (MDAKRLS) method, which is a machine learning approach, to identify potential associations of microbe–disease communication. To measure the microbial similarity of diseases, they introduced the similarity of the Hamming interaction characteristics. To construct two types of Kronecker similarities between pairs of microbes. Based on the well-known associations, they have calculated the similarity of Kronecker and the similarity of Hamming to disease-microbe pairs. To obtain prediction scores, Kronecker has designed at least four regular squares with different Kronecker similarities. They attained the ultimate forecast outcome by integrating the contributions of distinct similarities [83]. The architecture of predicting the microbe–disease relationship is shown in Fig. 3.

Other methods

There are some methods in the literature that do not fit into any of the above-mentioned groups. As a result, these approaches are discussed in this section.

The microbe similarity was calculated from the Gaussian Interaction Profile (GIP) kernel similarity, which is based on the well-known microbe–disease associations. Disease similarity was calculated using the mean of GIP similarity, symptom-based similarity, and functional similarity of the disease. The matrix completion method was used by the Singular Value Threshold (SVT algorithm) to compute the scores of unknown communication between disease-causing microbes. Finally, a low-rank matrix completion(called the MCHMDA) model was proposed [84]. Shi et al. suggest a new predictive method based on the Binary Matrix Completion (called BMCMDA) to forecast possible microbe-noninfectious disease associations (MDAs) by grouping a series of microbe–disease associations into a binary Microbe–disease association matrix. The suggested method suggests that the observed incomplete microbe–disease association matrix is the


Database
Microbe-disease
Association
Disbiom / HMDAD/peryton
/micropheno
Adjacency Matrix
Disease semantic similarity
Or
Symptom-based disease
Similarity
Gaussian interaction profile kernel similarity for diseases.
Or
COSINE SIMILARITY
DISEASE
Gaussian interaction profile kernel similarity for microbe.
Or
COSINE SIMILARITY
Microbe
Microbe-Microbe functional scores
…..
Disease similarity scores
Microbe similarity scores
Using the information obtained by different models, such as creating a heterogeneous network or using the space projection score is calculated or neural networks, etc., predictions are made.

	M_1	M_2	M_3	M_4
D_1	1	0	0	1
D_2	0	0	1	0
D_3	1	1	0	1
D_4	0	0	0	1

	d_1	d_2	d_3	d_4
d_1				
d_2				
d_3				
d_4				

	M_1	M_2	M_3	M_4
M_1				
M_2				
M_3				
M_4				

Fig. 3 The overall architecture of predicting microbe–disease relationship

sum of a latent parameterizing matrix and a noise matrix. It also provides a binomial model for sharing observations that occur independently of the microbial-disease association matrix [85].

The adaptive boosting for human microbe–disease association prediction (ABHMDA) was developed to explore the relationship between diseases and microbes. Due to the lack of sufficient information, the combination of microbial similarity of the GIP kernel and the similarity of symptomatic disease has been considered a feature of the experimental sample. Unknown associations have been used as negative examples as well as positive examples to maintain the balance between the samples during the decision tree training [86]. Lei et al. have proposed a model of microbial disease association with learning graph representations and a modified scoring mechanism on the heterogeneous network (called LGRSH). A heterogeneous network was shaped by combining microbial similarity networks, disease similarity networks, and known microbe–disease associations[87].

Recruited datasets and approaches for prediction of microbe–disease associations

Previous studies on the microbial-disease relationship have used a variety of data sources. Table 3 summarises the recruited datasets to predict microbial disease based on the information we reviewed.

In addition, different approaches for predicting the relationship between microbes and disease are summarized in Table 4.

Advantages and disadvantages

The KATZ measure might rebuild probable links concurrently in a vast network, but the computation of GIP kernel similarity will always lead to a bias towards those known relationships. Although the label propagation and random walk algorithms are effective and simple to use, the majority of prediction techniques built on them tend to have less detail. However, when more data is added to the network, training the embeddings will become more challenging. The weighted network-based and heteSim-based methods are excellent at capturing potential subtle semantic associations, but they cannot predict a microbe (drug, disease) in the absence of any known associations. The methods based on matrix factorization can mine deeper potential connections. Matrix factorization has a relatively low spatial complexity because it saves storage space, but selecting the optimal parameters is more challenging. GCN improves the applicability of translation invariance to non-matrix-structured data but it has poor flexibility and scalability. GAT can effectively enhance the aggregation effect of graph neural networks, but it is difficult to aggregate higher-order neighbors. The pooling layer will lose a lot of valuable information and ignore the correlation between the local and the whole.

Challenges and prospects

Based on the existing studies, some valuable suggestions are provided for further improving predictive performances.

Integrating multiple types of data for a single task

In this review, we briefly summarized the advanced and widely used dataset of computational methods related to the problems of microbe–disease and microbe–drug prediction, respectively. To improve prediction performance, the most basic idea is to combine all of these commonly represented databases as a whole to predict any single problem, because they are all closely related In addition, other types of datasets were also introduced, for example, chemical structure-based and phenotype-based data widely used in predictions [88–90], symptom-based disease similarity, and disease semantic similarity in predictions [48, 55]. Certainly, it is a challenge to improve the performance of the prediction model to rationally integrate different types of bioinformatics data to target a prediction task.

Introducing new mechanisms

The majority of currently available computational methods improved their performance by enriching more entity similarities than the previous algorithm. In addition to this strategy, many other approaches, such as heterogeneous graph neural network (GCN) and attention mechanisms [91–93], also work for this problem. For example, the attention mechanism can learn the importance of different neighboring nodes and the importance of different node (information) types to a current node. Many GNN models, such as the Spatial Convolution concept [94], can be introduced in link prediction problems. Moreover, most of the existing computational methods are supervised. The limited known associations' dataset is used as both training and testing sets, which will significantly hinder the utility and performance of the prediction model.

Benchmark evaluation

LOOCV and K-fold CV have been widely used as benchmark evaluation frameworks for link predictions. AUROC and ROC plots provide an overview of a predictor's performance and are commonly used to assess the prediction results. The computational approaches developed for the prediction problems of microbe-borne diseases and drugs always use strongly imbalanced datasets. The ROC plots could be misleading when applied in imbalanced prediction scenarios. [95].

Handling negative samples

The loss of negative samples significantly affects the prediction performance of the proposed model, and it is crucial to collect negative samples from biomedical databases and literature. To our knowledge, no actual negative samples have been collected and utilized in these predictive tasks presented in this survey. Developing computational methods to generate high-quality negative samples is an alternative to solving this problem [96].

Available microbiome databases

To the best of our knowledge, three databases have been developed on the subject of microbe–disease interaction, including HMDAD [21], Peryton [97], and Disbiome [98]. Several databases for empirically proven microbe–drug relationships are freely available in the field, such as MDAD [99], abiofilm [100], and Drug Virus.

MDAD (http://chengroup.cumt.edu.cn/MDAD) gathers 5,055 entries containing 1,388 drugs, 180 microbes,

Table 3 List of all the data that was utilized in the microbe–disease prediction

Data	Source	Original state	Similarity process	URL
HMDAD	The HMDAD database provides documentation of population disorders of disease-related microorganisms in PubMed	HMDAD integrated 483 disease-microbe entries which include 39 diseases and 292 microbes	They're reduced to 450 known MDAs that are then utilized to calculate GIP kernel, Cosine, and Spearman correlation similarity	https://www.cuilab.cn/hmdad
PERYTON	The content of Peryton is entirely supported by the manual curation of biomedical journals. Using reference tools to construct database dictionaries, diseases and Microbiota are supplied in a well-structured, well-organized format	There are currently over 7,900 entries in the database, which link 43 diseases and 1,396 microorganisms	Peryton also provides interactive visualizations, and the data may be downloaded straight to your computer for local storage and analysis	https://dianalab.e-ce.uth.gr/peryton/
GEN-BASED	On DisGeNET, you may find GDAs from UNIPROT, CGI, ClinGen, Genomics England, CTD (human subset), PsyGeNET, Orphanet, and those produced from text mining MEDLINE abstracts	Between 17 549 genes and 24 166 diseases, there are 628 685 GDAs covered. There are 37 diseases mapped, 1850 chromosomes, and 2715 GDAs Size/coverage in HMDAD	The neighbor-based similarity approach calculates GDA scores which were used to find further commonalities among a selection of disorders	https://www.disgenet.org
SYMPTOM-BASED disease data	HSDN pulls data from PubMed's large-scale medical bibliographic records of disease–symptom correlations	Simultaneous counting and TF-IDF weight values for 322 symptoms and 4442 disorders, with 147 97 connections and 22 mapped diseases, 269 symptoms, and 1858 associations of disease symptoms	The symptom-based illness similarity is calculated using Co-occurrence TF-IDFs between one illness and other symptoms	https://www.nature.com/articles/ncomms5212
Semantics-based disease data	MeSH trees are in the National Library of Medicine for a hierarchical definition of disease	Hierarchical trees systematically describe a variety of diseases 33 diseases of size/coverage mapped in HMDAD	The DAG-based semantic similarity of two disease trees made up of hierarchical descriptors is calculated	https://meshb.nlm.nih.gov/search
PROTEIN	STRING is a database that collects protein–protein interactions and data on proteins from several sources	At the species level, 1391 microbes were mapped, with gene neighbor scores of 932 370 pairs of COGs	The neighborhood score is used to determine if there is an edge between two COGs. Also provides interactive visualizations	https://string-db.org
Comprehensive Antibiotic Resistance Database (CARD)	A carefully curated resource offering high-quality reference material on the molecular basis of antimicrobial resistance (AMR), with a focus on the genes, proteins, and mutations implicated in AMR	CARD found 2441 model reference sequences, 853 single nucleotide alterations, as well as an increasing number of indels, frame shift, and nonsense mutations linked to antimicrobial resistance	Additional search criteria include mutations conferring AMR (if relevant) and curated BLAST(P/N) bit score cut-offs are included in the ontology	https://card.mcmaster.ca/
Disbiome	Created in 2018, is a more comprehensive database that is constantly updated every three months	As of December 2019, the Disbiome database includes 322 diseases, 1,470 microbiome organisms, and 9,102 experiments published in 1,018 scholarly articles	The human annotation guarantees a clear and organized presentation of the material that is accessible	https://disbiome.ugent.be/home/
MicroPhenoDB	There are 5677 non-redundant correlations between 1781 microorganisms and 542 human illness phenotypes across more than 22 human body locations in this study	In addition, MicroPhenoDB has 696,934 connections between 27,277 clade-specific core genes and 685 microorganisms	The software allows scientists to search DNA and RNA sequences for potential pathogens without running the usual meta-genomic data processing and assembly steps	http://www.liwzlab.cn/microphenodb

and 824 various strains (not including the microbes without defined strains) related to 993 references. All the references were from the 1970s to 2018. We can get all of the codes that correlate here (https://github.com/SunYazhou/MDAD).

The HMDAD (http://www.cuilab.cn/hmdad) is a database that compiles and organizes data on human microbe–disease associations from microbiota investigations. The database-integrated 39 diseases and 292 microbes among the 483 disease-microbe entries.

Peryton (https://dianalab.e-ce.uth.gr/peryton/) is a new database and resource containing empirically supported microbial disease associations. It hosts more than 7900 inputs related to 43 diseases and 1396 microorganisms.

Disbiome (https://disbiome.ugent.be/home/): Created in 2018, is a more comprehensive database that is constantly updated every three months. As of December 2019, the Disbiome database includes 322 diseases, 1,470 microbiome organisms, and 9,102 experiments published in 1,018 scholarly articles.

The 'a Biofilm' (http://bioinfo.imtech.res.in/manojk/abiofilm/): There are three sub-categories in this category, including data visualization, a database, and a prediction module. There are 5027 interactions between 1720 medications and 140 microorganisms in the database of anti-biofilm compounds, which includes gram-negative, gram-positive, and fungal microorganisms. Most studies from 1988 to 2017 reported experimental anti-biofilm agents against various microorganisms.

Drug Virus (https://drugvirus.info/tech_doc/) lists the activities and stages of development of 118 drugs that target 83 human viruses. The database allows users to explore virus-BSAA (Broad-spectrum antiviral agents) interactions in real time. DrugVirus.info is a free tool that includes a feedback form on its website. The website will be updated upon request or if a new save-in-man BSAA is discovered or a novel activity of an existing BSAA is discovered.

Available data based on similarity calculation

This section first discusses the computational method for germ-disease similarities (microbe–drug). It then lists databases and web servers that provide more information about the various diseases, drugs, and microbial components used there.

Based on disease similarity

Disease semantic similarity: Medical records of a particular disease are presented hierarchically in the National Library of Medicine (Mesh). Therefore, to measure the significance of a disease pair, we can use the overlap between the descriptors of the parents. Using both DAGs (directed acyclic graphs), the severity of the disease can be computed [101, 102].

- *Disease symptom similarity*: The similarities of symptomatic diseases based on HSDN were collected by Wheeler et al. [103]. With 849,103 PubMed records, they constructed 147,978 connections between 322 symptoms and 4219 diseases. Based on this data, they derived the similarities based on the symptoms of common diseases from HMDAD [104].
- *Gene-based disease data:* DisGeNET is the largest database on human gene-Disease Association (GDA) and disease types that combines all data in expert-curated repositories, GWAS catalogs, animal models, and scientific articles [105]. MEDLINE is the primary bibliographic database at the National Library of Medicine that holds the number of GDAs [106]. Bravo et al. used HMDAD and GDA databases to calculate the similarity of the recorded diseases.
- *Gene interactions:* The HumanNet v2.0 database (https://www.inetbio.org/humannet/download.php) is now available for efficiently accessing gene interactions, with each interaction having a log-likelihood score (LLS) that assesses the likelihood of a practical linkage between genes [107].

Based on microbe similarity

Microbe–microbe interactions: The MIND database curates the microbe-microbe interaction network data (http://www.microbialnet.org/mind_home.html/) Obtained[39].

- *Microbe data based on protein families*: The STRING database (https://string-db.org) includes protein–protein interactions and protein-related information from a variety of sources. The resource consists of the interactions obtained from computer prediction, information transmission across species, and interactions gathered from other (primary) databases. [52]. The purpose of this database is to achieve a global network of direct and indirect interactions. Collecting, integrating, scoring, and interacting protein-to-protein information, and completing these with computational predictions. Utilizing the proposed method, Kamneva calculated the functional similarity of the microbes [108, 109].
- *Microbe taxonomic similarity*: It contains more than 160,000 species with molecular data in the NCBI database, along with phylogenetic names and lineages and if two microbes have a common progenitor

Table 4 Various approaches for predicting the relationship between microbes and diseases

Category	Method	Description
Path-based methods	KATZHMDA, PBHMDA, MDPH_HMDA, BWNMHMDA, WMGHMDA	Numbers and weighted scores of various sorts of pathways between two nodes are often taken into consideration by path-based approaches
Random walk methods	RWRHMDA, BiRWHMDA, PRWHMDA, NTSHMDA, BDSILP, BiRWMP, BRWMDA, NBLPIHMDA, RWHMDA	For iterative walking, random walk algorithms provide a graph-based transition probability matrix
Bipartite local models	LRLSHMDA, NGRHMDA, NCPHMDA, KATZBNRA	BLMs are capable of making independent predictions on both the microbial and disease fronts
Matrix factorization methods	CMFHMDA, GRNMFHMDA, NMFMDA, KBMF, MDLPHMDA, mHMDA	Matrix factorization methods maximize two latent informative matrices, whose multiplication approximates the association matrix with distinct constraint terms, using different constraint terms
Network-based methods	MGATMDA, GATMDA, NINIMHMDA, BPNNHMDA,HMDA-PRED, LPHMDA	Because neural networks can adapt to changing input, they can produce the best possible outcome without requiring the output criteria to be redesigned
Machine learning-based	MDAKRLS	It is a machine learning-based strategy that employs fewer model parameters, saving time and ensuring reliable results
Other methods	ABHMDA, BMCMDA, MCHMDA	Ensemble learning and matrix completion are two of the most common strategies used

in a certain rank, they have a sequencing likeness to some extent [79].

Based on drug similarity

- *Drug microbe associations*: Drug Bank (http://www.drugbank.ca/) is a web-based database that contains detailed molecular information about medications, their mechanisms, interactions, and targets. The most recent update was in 2018 [110].
- *Genome sequences:* The NCBI database (https://www.ncbi.nlm.nih.gov/genome/) is used to obtain genome sequences. It contains a wealth of information about the disease that can be used to create DAG charts for disease expression.
- *Heterogeneous networks:* heterogeneous networks, namely microbe–drug heterogeneous networks and microbe–disease–drug networks, from a variety of sources including DrugBank [110], HMDAD [21], and CTD [111].

Drug structural similarity matrix: SIMCOMP search service (http://www.genome.jp/tools/simcomp/), is a chart-based solution for finding the most uniformity with the most click-through search on the chart. This server is used to find chemical similarities. The second search server is SUB COMP (http://www.genome.jp/tools/sumcomp/), which is a suggested method for solving the problem of uniformity under graphs. Both of these provide a basis for the study of chemical and physical properties [112].

Web-based tools

There are several web-based tools available to customize the prediction of microbe–disease associations. Among the web-based tools, there is a Micro Pattern for calculating similarities. For comparison, it divides microorganisms into disease-related classes. Currently, there are no tools available for enrichment analysis of a list of microorganisms. MicroPattern (http://www.cuilab.cn/micropattern) is a web-based tool for microbe set enrichment analysis. [113]. For other areas of expertise MicroPro predicts phenotypes using the complete case and controls frequency profiles and can estimate unknown microbial abundance profiles based on the unplanned readings of metagenomics results (for example, meta PPISP [114], DINIES [115],and DIANA-microT [116] are advanced forecasting methods) [117].Net Cooperate, an online instrument, can measure a host's capacity to provide nutritive support for a parasitic or commensal cell, as well as the (in addition to) complementarity of two microorganisms depending on their metabolic networks [118]. Using web-based operating systems and existing software, a methodological and biomedical study of microorganisms' reactions and humans becomes easier. (http://pharmacomicrobiomics.org) is a research-based online website dedicated to learning about how microbes modulate drug action [119].

Future directions and conclusion

Machine learning is a very useful technique, and similar algorithms like least squares, matrix factorization, and completion have been commonly applied to problems. Feature-based machine learning algorithms have been hampered by a lack of effective functionality and hence have gained little recognition. In comparison to machine learning, deep learning, which is considered a worthwhile effort, has yet to be implemented in MDA prediction. Many studies have proposed methods based on deep learning that target a complex topological network and catch its node embeddings in response to the aforementioned dilemma. The proposed models in the literature used many types of neural networks to predict drug-microbial communication. Since deep learning methods are a kind of machine learning method, it should be pointed out that their methods could be put to work for further studies to achieve better and more accurate predictions. Considering the ongoing trends of sources, databases, and experimental and laboratory articles in the field of microbiome, medicine, and diseases, more and stronger links between drugs, diseases and microbes could be considered, and forecasting these relationships with the help of computational approaches could pave the way for new microbe-based research discoveries [120–125].

In recent years, significant computational work has been done in the fields of microbes–disease and microbes–drugs. The work done in the field of microbes-disease has been used in different ways than microbes-medicine. Predicting drug and diseases associations with microbiome is very important in revealing the relationship between human diseases and drugs with microbiota. This article provides a thorough examination of forecasting microbial associations. Advanced omics technology and sequencing technologies enable a variety of methods to detect changes in the microbial composition of the patient. Data from existing trials and clinical results are problematic with information loss, non-uniform dispersion, lack of an integrated classification standard, and ambiguity of disease and drugs. To solve these problems using computational methods, machine learning algorithms and especially neural networks are recommended as an inimitable strategy. Machine learning methods are continually evolving, it is believed that the integration and development of these computational algorithms will improve the speed and accuracy of predicting interaction and structure. Most studies and work related to disease and medicine in the field of the microbiome could be conducted with the existence of databases in this field. The specific database for drugs and diseases in the field of the microbiome is very limited and needs to be developed with more experimental entries and accurate computationally predicted entries. Known associations in the field of microbes-disease and microbe–drug are relatively low and this leads to less prediction accuracy. If more links between them are identified and checked experimentally, other interactions will be predicted by computational methods accurately.

Key points

- Human health is influenced by the microbes that reside within and on human bodies. The microbe–disease association prediction is a computer-based pre-screening tool for clinical trials investigating microorganism-related pathogenic processes.
- Quantitative records of microbial population fluctuation in experimental instances enable the models to conduct fine-grained prediction tasks, and network analysis might become used to infer microbiological pathogenesis with annotated networks of biological events in the future.
- Predicting microbe–drug interactions can assist humans by making medication research and customized therapy efficiently.
- Exploring intricate mechanisms of microorganisms in clinical therapy, drug development, interactions, and repurposing will be considerably aided by prospective microbe–drug relationships prediction.

Abbreviations

LOOCV	Leave-one-out cross validation
AUC	Area under ROC curve
DAG	Directed acyclic graph
HIN	Heterogeneous information network
HMDAD	Human microbe–disease association database
RWR	Random walk with restart

Acknowledgements

Tabriz University of Medical Sciences supported this study as part of an MSc thesis.

Author contributions

HSG, YO, MPA, RF planned, coordinated, drafted and revised the manuscript. All the authors equally contributed to the preparation of this manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 9 January 2023 Accepted: 21 February 2023
Published online: 07 March 2023

References

1. Sekirov I, et al. Gut microbiota in health and disease. Physiolog Rev. 2010;90:859.
2. Kashyap PC, et al. Microbiome at the frontier of personalized medicine. Mayo Clin Proc. 2017;92:1855.
3. Goodrich JK, et al. Human genetics shape the gut microbiome. Cell. 2014;159(4):789–99.
4. Gill SR, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.
5. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. Cell. 2014;157(1):121–41.
6. Li N, et al. The commensal microbiota and viral infection: a comprehensive review. Front Immunol. 2019;10:1551.
7. Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J. 2017;474(11):1823–36.
8. Wang B, et al. The human microbiota in health and disease. Engineering. 2017;3(1):71–82.
9. Fernandes R, et al. Diabetic gut microbiota dysbiosis as an inflammaging and immunosenescence condition that fosters progression of retinopathy and nephropathy. Biochim Biophys Acta BBA Mol Basis Dis. 2019;1865(7):1876–97.
10. Hasan N, Yang H. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ. 2019;7: e7502.
11. Moran-Ramos S, et al. Environmental and intrinsic factors shaping gut microbiota composition and diversity and its relation to metabolic health in children and early adolescents: a population-based study. Gut Microbes. 2020;11(4):900–17.
12. Jesmok EM, Hopkins JM, Foran DR. Next-generation sequencing of the bacterial 16S rRNA gene for forensic soil comparison: a feasibility study. J Forensic Sci. 2016;61(3):607–17.
13. Ranjan R, et al. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469(4):967–77.
14. Busnelli M, Manzini S, Chiesa G. The gut microbiota affects host pathophysiology as an endocrine organ: a focus on cardiovascular disease. Nutrients. 2020;12(1):79.
15. Sanchez-Rodriguez E, et al. The gut microbiota and its implication in the development of atherosclerosis and related cardiovascular diseases. Nutrients. 2020;12(3):605.
16. Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018;67(9):1716–25.
17. Methé BA, et al. A framework for human microbiome research. Nature. 2012;486(7402):215.
18. Huttenhower C, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
19. Coelho ED, et al. Computational methodology for predicting the landscape of the human–microbial interactome region level influence. J Bioinform Comput Biol. 2015;13(05):1550023.
20. Turnbaugh PJ, et al. The human microbiome project. Nature. 2007;449(7164):804–10.
21. Ma W, et al. An analysis of human microbe–disease associations. Brief Bioinform. 2017;18(1):85–97.
22. Woting A, Blaut M. The intestinal microbiota in metabolic disease. Nutrients. 2016;8(4):202.
23. Rizkallah RM, et al. The pharmacomicrobiomics portal: a database for drug–microbiome interactions. Curr Pharmacogenom Personal Med. 2012;10(3):195–203.
24. Montassier E, et al. Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. Genome Med. 2016;8:1–11.
25. Gopalakrishnan V, et al. The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. Cancer Cell. 2018;33(4):570–80.
26. Li W, et al. Gut microbiome and cancer immunotherapy. Cancer Lett. 2019;447:41–7.
27. Alexander LJW. Gutmicrobiot a modulation of chemotherapy efficacy and toxicity. Nat Rev Gastroenterol He Patol. 2017;14(6):356.
28. Maier L, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 2018;555(7698):623–8.
29. Singh A, et al. Microbiome and host crosstalk: a new paradigm to cancer therapy. Semin Cancer Biol. 2021;70:71–84.
30. Liu F, et al. Dysbiosis of the gut microbiome is associated with tumor biomarkers in lung cancer. Int J Biol Sci. 2019;15(11):2381.
31. Panebianco C, Andriulli A, Pazienza V. Pharmacomicrobiomics: exploiting the drug–microbiota interactions in anticancer therapies. Microbiome. 2018;6(1):1–13.
32. Aziz RK, et al. Drug pharmacomicrobiomics and toxicomicrobiomics: from scattered reports to systematic studies of drug–microbiome interactions. Expert Opin Drug Metab Toxicol. 2018;14(10):1043–55.
33. Wilson ID, Nicholson JK. Gut microbiome interactions with drug metabolism, efficacy, and toxicity. Transl Res. 2017;179:204–22.
34. Yamashiro Y. Gut microbiota in health and disease. Ann Nutr Metab. 2017;71(3–4):242–6.
35. Sommer F, Bäckhed F. The gut microbiota–masters of host development and physiology. Nat Rev Microbiol. 2013;11(4):227–38.
36. Zimmermann M, et al. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature. 2019;570(7762):462–7.
37. Haiser HJ, et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. Science. 2013;341(6143):295–8.
38. Long Y, et al. Ensembling graph attention networks for human microbe–drug association prediction. Bioinformatics. 2020;36(Suppl_2):i779–86.
39. Long Y, Luo J. Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. IEEE J Biomed Health Informat. 2020;25(1):266–75.
40. Zhu L, et al. Prediction of microbe–drug associations based on Katz measure. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. IEEE.
41. Yahui L, et al. Predicting human microbe–drug associations via graph convolutional network with conditional random field. 2020.
42. Long Y, et al. Predicting human microbe–drug associations via graph convolutional network with conditional random field. Bioinformatics. 2020;36(19):4918–27.
43. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics. 2011;27(21):3036–43.
44. Wu C, et al. PRWHMDA: human microbe–disease association prediction by random walk on the heterogeneous network with PSO. Int J Biol Sci. 2018;14(8):849.
45. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
46. Wang L, et al. A novel approach based on bipartite network to predict human microbe–disease associations. Curr Bioinform. 2018;13(2):141–8.
47. Deng L, et al. Graph2MDA: a multi-modal variational graph embedding model for predicting microbe–drug associations. Bioinformatics. 2022;38(4):1118–25.
48. Chen X, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017;33(5):733–9.
49. Luo J, Long Y. NTSHMDA: prediction of human microbe–disease association based on random walk by integrating network topological similarity. IEEE/ACM Trans Comput Biol Bioinf. 2018;17(4):1341–51.
50. Yu Z, et al. Predicting drug–disease associations through layer attention graph convolutional network. Brief Bioinformat. 2021;22(4):bbaa243.
51. Petrova MI, et al. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. Front Physiol. 2015;6:81.

52. Long Y, Luo J. WMGHMDA: a novel weighted meta-graph-based model for predicting human microbe–disease association on heterogeneous information network. BMC Bioinformat. 2019;20(1):1–18.
53. Li H, et al. A novel human microbe–disease association prediction method based on the bidirectional weighted network. Front Microbiol. 2019;10:676.
54. Huang Z-A, et al. PBHMDA: path-based human microbe–disease association prediction. Front Microbiol. 2017;8:233.
55. Fan C, et al. Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. Neurocomputing. 2019;323:76–85.
56. Niu Y-W, et al. RWHMDA: random walk on hypergraph for microbe–disease association prediction. Front Microbiol. 2019;10:1578.
57. Zou S, Zhang J, Zhang Z. A novel approach for predicting microbe–disease associations by bi-random walk on the heterogeneous network. PLoS ONE. 2017;12(9): e0184394.
58. Zhang W, et al. The bi-direction similarity integration method for predicting microbe–disease associations. IEEE Access. 2018;6:38052–61.
59. Yan C, et al. Brwmda: predicting microbe–disease associations based on similarities and bi-random walk on disease and microbe networks. IEEE/ACM Trans Comput Biol Bioinf. 2019;17(5):1595–604.
60. Shen X, et al. Predicting disease-microbe association by random walking on the heterogeneous network. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016. IEEE.
61. Wang L, et al. A bidirectional label propagation based computational model for potential microbe–disease association prediction. Front Microbiol. 2019;10:684.
62. Wang D, et al. Human microbe–disease association prediction by a novel double-ended random walk with restart. BioMed Res Int. 2020.
63. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–403.
64. Huang Y-A, et al. Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. J Transl Med. 2017;15(1):1–11.
65. Wang F, et al. LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. Sci Rep. 2017;7(1):1–11.
66. Bao W, Jiang Z, Huang D-S. Novel human microbe–disease association prediction using network consistency projection. BMC Bioinformat. 2017;18(16):173–81.
67. Li S, Xie M, Liu X. A novel approach based on bipartite network recommendation and KATZ model to predict potential micro-disease associations. Front Genet. 2019;10:1147.
68. Li L-X, et al. A fast algorithm for nonnegative matrix factorization and its convergence. IEEE Trans Neural Netw Learn Syst. 2014;25(10):1855–63.
69. Tian L-P, et al. CASNMF: a converged algorithm for symmetrical nonnegative matrix factorization. Neurocomputing. 2018;275:2031–40.
70. Wu C, Gao R, Zhang Y. mHMDA: human microbe–disease association prediction by matrix completion and multi-source information. IEEE Access. 2019;7:106687–93.
71. Chen S, et al. Predicting microbe–disease association by kernelized Bayesian matrix factorization. In: International Conference on Intelligent Computing. 2018. Springer.
72. Liu Y, Wang S-L, Zhang J-F. Prediction of microbe–disease associations by graph regularized non-negative matrix factorization. J Comput Biol. 2018;25(12):1385–94.
73. Shen Z, Jiang Z, Bao w. CMFHMDA: Collaborative matrix factorization for human microbe–disease association prediction. In: International Conference on Intelligent Computing. 2017. Springer.
74. He B-S, Peng L-H, Li Z. Human microbe–disease association prediction with graph regularized non-negative matrix factorization. Front Microbiol. 2018;9:2560.
75. Qu J, Zhao Y, Yin J. Identification and analysis of human microbe–disease associations by matrix decomposition and label propagation. Front Microbiol. 2019;10:291.
76. Liu Y, et al. DMFMDA: prediction of microbe–disease associations based on deep matrix factorization using Bayesian Personalized Ranking. IEEE/ACM Trans Comput Biol Bioinformat. 2020.
77. Long Y, et al. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. Brief Bioinform. 2021;22(3):bbaa146.
78. Dayun L, et al., MGATMDA: predicting microbe–disease associations via multi-component graph attention network. In: IEEE/ACM Transactions on computational biology and bioinformatics. 2021.
79. Ma Y, Jiang H. NinimHMDA: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe–disease association. Bioinformatics. 2020;36(24):5665–71.
80. Li H, et al. Identifying microbe–disease association based on a novel back-propagation neural network model. In: IEEE/ACM transactions on computational biology and bioinformatics, 2020.
81. Fan Y, et al. Inferring disease-associated microbes based on multi-data integration and network consistency projection. Front Bioeng Biotechnol. 2020; 8.
82. Peng L, et al. Prioritizing human microbe–disease associations utilizing a node-information-based link propagation method. IEEE Access. 2020;8:31341–9.
83. Xu D, et al. MDAKRLS: predicting human microbe–disease association based on Kronecker regularized least squares and similarities. J Transl Med. 2021;19(1):1–12.
84. Yan C, et al. Mchmda: predicting microbe–disease associations based on similarities and low-rank matrix completion. IEEE/ACM Trans Comput Biol Bioinf. 2021;18(02):611–20.
85. Shi J-Y, et al. BMCMDA: a novel model for predicting human microbe–disease associations via binary matrix completion. BMC Bioinformat. 2018;19(9):85–92.
86. Peng L-H, et al. Human microbe–disease association prediction based on adaptive boosting. Front Microbiol. 2018;9:2440.
87. 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.
88. Yang M, et al. Computational drug repositioning based on multi-similarities bilinear matrix factorization. Brief Bioinformat. 2021;22(4):bbaa267.
89. Cai L, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinformat. 2021;22(6):319.
90. Lu L, Yu H. DR2DI: a powerful computational tool for predicting novel drug–disease associations. J Comput Aided Mol Des. 2018;32(5):633–42.
91. Jing Y, et al. Amalgamating knowledge from heterogeneous graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
92. Wang X., et al. Heterogeneous graph attention network. In: The world wide web conference. 2019.
93. Li X, et al. Type-aware anchor link prediction across heterogeneous networks based on graph attention network. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020.
94. Kanehisa M, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34(Suppl_1):D354–7.
95. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3): e0118432.
96. Cheng Z, et al. Selecting high-quality negative samples for effectively predicting protein–RNA interactions. BMC Syst Biol. 2017;11(2):1–11.
97. Skoufos G, et al. Peryton: a manual collection of experimentally supported microbe–disease associations. Nucleic Acids Res. 2021;49(D1):D1328–33.
98. Janssens Y, et al. Disbiome database: linking the microbiome to disease. BMC Microbiol. 2018;18(1):1–6.
99. Sun Y-Z, et al. MDAD: a special resource for microbe–drug associations. Front Cell Infect Microbiol. 2018;8:424.
100. Rajput A, et al. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. Nucleic Acids Res. 2018;46(D1):D894–900.
101. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA. 1994;271(14):1103–8.
102. Schriml LM, et al. Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019;47(D1):D955–62.
103. Wheeler DL, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003;31(1):28–33.

104. Zhou X, et al. Human symptoms–disease network. Nat Commun. 2014;5(1):1–10.
105. Piñero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2016;45:833.
106. Bravo À, et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformat. 2015;16(1):1–17.
107. Hwang S, et al. HumanNet v2: human gene networks for disease research. Nucleic Acids Res. 2019;47(D1):D573–80.
108. Szklarczyk D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.
109. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. PLoS Comput Biol. 2017;13(2): e1005366.
110. Wishart DS, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
111. Davis AP, et al. The comparative toxicogenomics database: update 2019. Nucleic Acids Res. 2019;47(D1):D948–54.
112. Hattori M, et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res. 2010;38(Suppl_2):W652–6.
113. Ma W, et al. MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. Sci Rep. 2017;7(1):1–6.
114. Qin S, Zhou H-X. meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics. 2007;23(24):3386–7.
115. Yamanishi Y, et al. DINIES: drug–target interaction network inference engine based on supervised analysis. Nucleic Acids Res. 2014;42(W1):W39–45.
116. Maragkakis M, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res. 2009;37(Suppl_2):W273–6.
117. Zhu Z, et al. MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biol. 2019;20(1):1–13.
118. Levy R, et al. NetCooperate: a network-based tool for inferring host-microbe and microbe–microbe cooperation. BMC Bioinformat. 2015;16(1):1–6.
119. Emami N, Ferdousi R. AptaNet as a deep learning approach for aptamer–protein interaction prediction. Sci Rep. 2021;11(1):1–19.
120. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.
121. Huang Z, Mamoulis N. Heterogeneous information network embedding for meta path based proximity. arXiv preprint arXiv:1701.05291, 2017.
122. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
123. Shang J, et al. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. arXiv preprint arXiv:1610.09769, 2016.
124. Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015.
125. Tang J, et al. Line: large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web. 2015.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC