OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–25
https://doi.org/10.1093/bib/bbac080
Review

Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models

Lei Wang, Yaqin Tan, Xiaoyu Yang, Linai Kuang and Pengyao Ping

Corresponding authors. Lei Wang, College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, 410022, Hunan, China and Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, 411105, Hunan, China. Email: wanglei@xtu.edu.cn; Pengyao Ping, College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, 410022, Hunan, China. Email: ping.pengyao@gmail.com

Abstract

In recent years, with the rapid development of techniques in bioinformatics and life science, a considerable quantity of biomedical data has been accumulated, based on which researchers have developed various computational approaches to discover potential associations between human microbes, drugs and diseases. This paper provides a comprehensive overview of recent advances in prediction of potential correlations between microbes, drugs and diseases from biological data to computational models. Firstly, we introduced the widely used datasets relevant to the identification of potential relationships between microbes, drugs and diseases in detail. And then, we divided a series of a lot of representative computing models into five major categories including network, matrix factorization, matrix completion, regularization and artificial neural network for in-depth discussion and comparison. Finally, we analysed possible challenges and opportunities in this research area, and at the same time we outlined some suggestions for further improvement of predictive performances as well.

Keywords: microbe–drug interactions, microbe–disease associations, drug–disease relationships, computational models, predictive methods

Introduction

Microorganisms, including bacteria, viruses, archaea, fungi and protozoa, are dynamic, diverse and complex genetic reservoirs that exist in interactive flux, colonize human cells and play significant roles in human beings [1]. The scale of microbiome is enormous, and the symbiotic bacteria alone outnumbers human cells and encodes 150 times more unique genes than their human hosts [2]. The relationships between microbiome and human hosts can be regarded as a continuum that ranges from symbiosis and commensalism (mutualism) to pathogenesis. On the one hand, microorganisms can facilitate metabolism [3], produce essential vitamins and gene products [4] and protect against invasion from pathogens [5]. On the other hand, it has been demonstrated that unusual growth or decline of microorganisms will influence human health and cause diseases including obesity [6]and inflammatory bowel disease [7] etc. For instance, some microbes, including pathogens, germs and bugs, may cause infectious diseases such as the coronavirus disease 2019 [8]. However, studies have shown that microbial metabolism can significantly affect the clinical response to medication, and drug administration can also have a specific impact on the microbiome [1, 9, 10]. For example, it has been proved that there exist various interactions between drugs and the gut microbiome [13, 185]. Javdan *et al*. developed a quantitative experimental framework, Microbiome-Derived Metabolism (MDM)-Screen [186], to study the variable ability of the human gut microbe to metabolize multiple drugs. They found differences in drug–microbial interactions between individuals, which means that the differentiation of microorganisms needs to be taken into account in personalized drug development [187]. Maier *et al*. compared the effects of 1079 marketed drugs on representative intestinal symbiotic microorganisms. They found that 24% of medicines targeted at the human body had inhibitory effects on microorganisms, especially antipsychotic drugs [188]. Concetta *et al*. reported that gut microbiota could interact with anticancer drugs, thus affecting the therapeutic efficiency and toxic side effects of drugs. They considered the use of probiotics, prebiotics, synbiotics, biologics and antibiotics as emerging strategies for microbiota control, which might improve treatment outcomes or ensure that patients have a better quality of life during anticancer treatment [189]. The human microbiome regulates many important physiological functions, drug administration

微生物可以促进代谢，产生必需的维生素和基因产物，防止病原体的入侵

微生物和药物之间存在相互影响

Lei Wang is a full professor in the College of Computer Engineering and Applied Mathematics at Changsha University, Changsha,Hunan,China.
Yaqin Tan is a graduate student in the School of Computer Science at Xiangtan University,Xiangtan,Hunan,China.
Xiaoyu Yang is a graduate student in the School of Computer Science at Xiangtan University,Xiangtan,Hunan,China.
Linai Kuang is an associate professor in the School of Computer Science at Xiangtan University,Xiangtan,Hunan,China.
Pengyao Ping is a research assistant in the College of Computer Engineering and Applied Mathematics at Changsha University,Changsha,Hunan,China.

Received: December 14, 2021. **Revised:** February 14, 2022. **Accepted:** February 15, 2022

has specific effects on the microbiome and microbial metabolism can significantly affect clinical drug response, which indicate that there is an inseparable and complex relationship between microorganisms, drugs and diseases. Therefore, discovering potential pairwise associations between microbes, drugs and diseases can offer essential insights into the understanding of underlying disease mechanisms from the perspective of human microbes and drugs, which may be very helpful for investigating pathogenesis, promoting early diagnosis and improving precision medicines.

However, it is time-consuming and quite expensive to adopt traditional clinical trials to identify relationships between microbes, drugs and diseases. For example, it was reported that a new drug might need at least 10 years and cost as much as $1 billion from laboratory research to a flourishing market [11, 12]. Hence, in the past few years, with the rapid development of techniques in genomics, proteomics, life sciences and pharmaceutical researches, a large amount of biomedical data has been accumulated, based on which many researchers have designed numerous calculative methods to infer potential Microbe–Disease Associations (MDsAs), Drug–Disease Associations (DgDsAs) and Microbe–Drug Associations (MDgAs).

Regarding the collection of biomedical data, a large number of public databases associated with MDsAs, DgDsAs and MDgAs have been established separately. For instance, in terms of MDsAs, there are five typical databases, including HMDAD [13], Disbiome [14], MicroPhenDB [15], MDIDB [16] and Peryton [17], having been constructed from 2016 to 2021. Additionally, Zhao *et al.* briefly introduced six databases and five web servers relevant to microbes in a study on microbes and complex diseases from experimental results to computational models in 2020 [183]. In terms of DgDsAs, representative databases include CTD [18–24], DrugBank [25], TTD [26], OMIM [27], DailyMed [28] and PubChem [29], etc. In terms of MDgAs, MDAD [30], aBiofilm [31] and DrugVirus [32] are the three most widely used databases. Meanwhile, with the rapid development and expansion of public databases, over the past few years, a series of computational approaches have been designed in succession to identify potential associations between microbes, drugs and diseases as well. For example, Zhao *et al.* [183] reviewed four types of prediction algorithms developed based on score function, network algorithm, machine learning and experimental analysis, while Wen *et al.* provided a general overview of predictive studies of MDsAs in terms of biological data and predictive methods, and also experimentally evaluated the prediction performances of the reviewed methods based on different similarity data and calculation methods [33]. However, all these existing studies and related investigations only focus on a certain kind of forecasting problem, and there is no overarching survey so far putting all these predictive issues of MDgAs, MDsAs and DgDsAs together from biological data to computational methods. Hence, in this paper, we aim to provide an all-sided review about prediction of pairwise relationships between human microbes, drugs and diseases from biological data to computational models. First of all, as illustrated in Table 1, we would provide an in-depth review of 17 state-of-the-art datasets relevant to identification of potential MDsAs, DgDsAs and MDgAs, respectively. And then, as shown in Table 2, all kinds of carefully selected representative computing models would be classified into four main categories, such as network, matrix factorization, machine learning and artificial neural network etc., for a comprehensive analysis. After that, all methods in each category would be further divided into different subclasses according to the difference of core techniques adopted by them for detailed discussion and comparison.

介绍数据集

两篇综述可以看

它们的综述是分开讨论，这篇文章把 Microbes , Drugs , Diseases 放在一起讨论

In the rest of the sections, we would begin by detailing the widely used databases in Table 1 that are relevant to identifying possible relationships between human microbes, drugs and diseases. And then, we would present a brief analysis for the data collected from all these databases. Thereafter, we would carefully select state-of-the-art calculation methods for each predictive task and compare them according to their frameworks from main categories to subclasses in detail. Besides, we would further compare the predictive performances of typical methods based on massive experiments and discuss the advantages and disadvantages of them as well. Finally, we would propose some suggestions for improving performances of predictive models and outline some challenges and opportunities in the research area of predicting pairwise relationships between human microbes, drugs and diseases.

1. 首先从已往数据库挖掘三者之间的联系
2. 选择SOTA的预测模型进行比较
3. 进一步比较在大规模数据集上
4. 提出改进的意见和还有那些挑战

Data resources

The growing number of publicly available databases promotes the development of predictive tasks in computational biology. In this section, we presented a detailed review on those widely used datasets related to identification of potential MDsAs, DgDsAs and MDgAs, respectively. At the beginning, we selected 17 representative datasets (from DS1 to DS17) and briefly summarized them in Table 1. And then, in order to explore the logical relationships between these selected datasets, three new datasets including DS18, DS19 and DS20 (please see these bold rows in Table 1) were obtained by analysing intersections and unions of these selected datasets.

Data related to MDsAs

In order to study potential relationships between microorganisms and diseases, various databases relevant to MDsAs have been constructed in recent years. For example, in 2016, Ma *et al.* created a Human Microbe–Disease Association Database (HMDAD) by collecting confirmed MDsAs from published literatures, in which 483 known microbe–disease associations between 39 diseases and 292 microbes (marked as DS1) were

Table 1. An overview of widely used databases that are related to identification of potential MDsAs, DgDsAs and MDgAs

		Number of							
Data type	ID	Associations	Diseases	Drugs	Microbes	Ref.	URL	Data Source	Used in the Ref
MDsAs	DS1	483	39	\	292	[13]	http://www.cuilab.cn/hmdad	HMDAD	[34–41] [42–49] [50–57] [58–65] [66–70]
	DS2	10 922	372	\	1622	[14]	https://disbiome.ugent.be/	Disbiome	[42, 46]
	DS3	5677	542	\	1781	[15]	http://www.liwzlab.cn/microphenodb	MicroPhenDB	[71]
	DS4	44 900	1198	\	1065	[16]	http://dbmdi.com/index/	MDIDB	\
	DS5	7977	43	\	1396	[17]	https://dianalab.e-ce.uth.gr/peryton/	Peryton	[71]
	DS18	**15 633**	**787**	\	**2916**	\	\	\	\
DgDsAs	DS6	1933	593	313	\	[72]	https://www.embopress.org/doi/full/10.1038/msb.2011.26	OMIM[27],DrugBank[25]	[73–80] [81–88] [89–96] [97–99]
	DS7	2794	654	698	\	[100]	https://github.com/huayu1111/DR2DI	DrugBank[25],UMLS[101]	\
	DS8	1008	4516	1490	\	[102]	http://genome2.ugr.es/drugnet/	DrugBank[25],DO[103]	[80, 81, 98]
	DS9	2352	409	663	\	[80]	http://github.com//bioinfomaticsCSU/MBiRW	OMIM[27],DrugBank[25],DO[103]	[74, 80, 81, 84, 86, 88, 92] [95, 97]
	DS10	3051	681	763	\	[104]	https://github.com/LiangXujun/LRSSL	DrugBank[25],UMLS[101]	[76, 77, 95, 105–108] [109–111]
	DS11	6677	1299	1519	\	[112]	https://github.com/stjin-XMU/HeTDR	DrugBank[25],repoDB[113]	[114]
	DS12	18 416	598	269	\	[76]	http://www.bioinfotech.cn/SCMFDD/	CTD[115]	[77, 95, 116–118]
	DS13	49 217	2834	1323	\	[76]	http://www.bioinfotech.cn/SCMFDD/	CTD[115]	\
	DS14	26 521	2093	4501	\	[119]	https://www.frontiersin.org/articles/10.3389/fchem.2019.00924/full&#x2216;#supplementary-material	CTD[115]	\
	DS19	**70 315**	**4369**	**5312**	\	\	\	\	\
MDgAs	DS15	5505	\	1388	180	[30]	http://www.chengroup.cumt.edu.cn/MDAD/	MDAD	[120–125]
	DS16	5027	\	1720	140	[31]	http://bioinfo.imtech.res.in/manojk/abiofilm/	aBiofilm	[122, 124]
	DS17	1281	\	118	83	[32]	https://drugvirus.info	DrugVirus	[122]
	DS20	**4491**	\	**2051**	**315**	\	\	\	\

Note: These data types listed on this table do not mean that the given database only contains these data types since this review only focuses on the related datasets instead of the whole database.

Table 2. A summary of the computational methods targeting predictions of potential MDsAs, MDgAs and DgDsAs. The abbreviations in this table are as follows: NE: Network Embedding, LP: Label Propagation, RW: Random Walk, BiRW: Bi-Random Walk, WN: Weighted Network, NCP: Network Consistency Projection, CF: Collaborative Filtering; GRLNN: Graph Regularized Non-negative, SC: Similarity Constrained, KB: Kernelized Bayesian; CNN: Convolution Neural Network, GCN: Graph Convolution Neural Network, GAT: Graph Attention Network, CVAE: Collective Variational Autoencoder, SAE: Sparse Autoencoder, DNN: Deep Neural Network, KAE: Kernel-based Autoencoder, BPNN: Back Propagation Neural Network.

Methods category	Topic category	Methods sub-category	Methods	Data and code open source	Publication year	Ref.
network-based	MDsAs	KATZ	KATZHMDA	Yes	2016	[62]
		NE	MDKG	No	2020	[66]
			LGRSH	No	2020	[40]
			MSLINE	No	2021	[65]
		LP	BDSILP	No	2018	[35]
			NBLPIHMDA	No	2019	[64]
			MDLPHMDA	No	2019	[53]
			NCPLP	No	2020	[47]
		RW	RWRH	No	2016	[43]
			BiRWMP	No	2018	[63]
			PRWHMDA	No	2018	[37]
			RWHMDA	No	2019	[36]
			NTSHMDA	No	2019	[45]
			DRWHMDA	No	2020	[68]
			BiRWHMDA	No	2017	[61]
			BRWMDA	No	2019	[58]
		HeteSim score	MDPH_HMDA	No	2019	[128]
		WN	PBHMDA	No	2017	[44]
			BWNMHMDA	No	2019	[60]
			WMGHMDA	Yes	2019	[34]
		NCP	HMDA-Pred	Yes	2020	[52]
		CF	NGRHMDA	Yes	2017	[67]
		NE	Wang *et al.*	No	2014	[78]
			Chen *et al.*	No	2015	[129]
			HED	No	2019	[79]
			EMP-SVD	No	2019	[82]
			TS-SVD	No	2020	[89]
			NEDD	No	2020	[85]
			HeTDR	Yes	2021	[114]
	DgDsAs	LP	Huang *et al.*	No	2013	[130]
			DrugNet	data: Yes code: No	2015	[102]
			Heter-LP	No	2017	[131]
			NTSIM	No	2018	[77]
		RW	TP-NRWRH	Yes	2016	[97]
			miRDDCR	No	2017	[132]
		BiRW	MBiRW	Yes	2016	[80]
			DR-IBRW	No	2019	[133]
			BiRWDDA	No	2019	[93]
		HeteSim Score	HSDD	No	2018	[94]
		WN	Min *et al.*	No	2014	[134]
			Lee *et al.*	No	2018	[135]
	MDgAs	KATZ	HMDAKATZ	No	2019	[123]
		NE	HNERMDA	No	2020	[124]
Matrix factorization	MDsAs	GRLNN-MF	HNGRNMF	No	2018	[136]
			GRNMFHMDA	No	2018	[54]
			NMFMDA	No	2018	[38]
Matrix factorization	MDsAs	GRLNN-MF	MDNMF	Yes	2020	[137]
		Logistic-MF	RNMFMDA	No	2020	[69]
		KB-MF	KBMF	No	2018	[41]
			DMFMDA	No	2021	[56]
		Collaborative-MF	CMFHMDA	No	2017	[57]
	DgDsAs	GRLNN-MF	DivePred	No	2019	[105]
			DisDrugPred	No	2019	[108]
		Logistic-MF	CI-PMF	No	2014	[138]
			DDAPRED	Yes	2020	[109]

(Continued)

Table 2. Continued

Methods category	Topic category	Methods sub-category	Methods	Data and code open source	Publication year	Ref.
		SC-MF	DDR	No	2014	[139]
			MSBMF	Yes	2020	[92]
			SCMFDD	data: Yes code: No	2018	[76]
		Collaborative-MF	CMFMTL	Yes	2018	[116]
Matrix completion	MDsAs	\	BMCMDA	No	2018	[59]
			MCHMDA	No	2019	[50]
			mHMDA	No	2019	[48]
	DgDsAs	\	OMC	No	2019	[87]
			BNNR	Yes	2019	[91]
			HGIMC	Yes	2020	[140]
Regularization	MDsAs	\	LRLSHMDA	No	2017	[51]
			MDAKRLS	No	2021	[49]
	DgDsAs	\	LRSSL	Yes	2017	[104]
			RLSDR	No	2018	[98]
			DR2DI	Yes	2018	[100]
	MDgAs	\	LRLSMDA	No	2021	[125]
Neural Network	MDsAs	GCN-based AE	NinimHMDA	Yes	2021	[46]
		GAT-based AE	GATMDA	Yes 没有开源，是个无效链接	2021	[42]
			MGATMDA	Yes 没有开源，没有提供核心代码	2021	[71]
		BPNN	BPNNHMDA	No	2020	[70]
	DgDsAs	CNN	HeteroDualNet	No	2019	[106]
			CBPred	No	2019	[141]
			SKCNN	data: Yes code: No	2019	[86]
			CGARDP	No	2019	[107]
			Zhan *et al.*	No	2020	[119]
		GCN-based AE	LAGCN	Yes	2020	[117]
			GFPred	No	2020	[111]
			DRHGCN	Yes	2021	[95]
			ANPred	No	2021	[127]
		CVAE	deepDR	Yes	2019	[112]
			SNF-CVAE	No	2020	[142]
		SAE	SAEROF	No	2020	[84]
		KAE	GIPAE	data: Yes code: No	2019	[74]
		GCN	MGRL	No	2021	[118]
		DNN	HNet-DNN	No	2020	[88]
	MDgAs	GCN-based AE	GCNMDA	Yes	2020	[122]
			Graph2MDA	Yes	2021	[120]
		GAT-based AE	EGATMDA	Yes	2020	[121]

selected from 61 previous research works [13]. In 2018, Janssens *et al.* built an MDsAs related database named Disbiome, in which there are 10 922 known associations between 372 diseases and 1622 microbiome organisms (marked as DS2) having been screened from 1191 published academic papers and included in Disbiome [14]. In addition, based on HMDAD and Disbiome, Yao *et al.* constructed a new MDsAs related database called MicroPhenoDB, including 5677 non-redundant associations between 1781 microorganisms and 542 human disease phenotypes (marked as DS3) in 22 newly collected human parts, and 696 934 associations between 27 277 branch-specific core genes and 685 microorganisms [15]. Moreover, in 2021, Wu *et al.* excavated known MDsAs from literatures via a text mining method based on the transfer learning framework and created a new database named MDIDB, comprising 1198 diseases, 1065 microbes and 44 900 associations (marked as DS4) [16]. Subsequently, Skoufos *et al.* structured a novel database called Peryton consisting of experimentally supported MDsAs, in which 7977 associations linking 43 diseases and 1396 microbes (marked as DS5) were included [17].

Data related to DgDsAs

In 2009, Allan *et al.* released the first version of CTD, a robust and open database that provides hand-curated chemical, gene, protein, disease information and their relationships for latent DgDsAs prediction. Apart from CTD, the databases such as DrugBank, TTD and OMIM include datasets related to DgDsAs as well. Based on these databases, in 2011, Gottlieb *et al.* assembled 1933 associations between 593 diseases presented in the OMIM and 313 drugs registered in DrugBank (marked as DS6) [72]. In addition, in 2012, Li *et al.* collected 3250 treatment relationships between 799 drugs and 719 diseases [127], based on which Lu *et al.* used only the FDA-approved small molecule drugs with the target protein, therapeutic indication and PubChem ID recorded in DrugBank to select 2794 associations between 698 drugs and 654 diseases (marked as DS7) [100]. In 2015, Martínez *et al.* collected 1008 known DgDsAs and built a database named DNdataset (marked as DS8), in which 4516 diseases annotated by Disease Ontology (DO) terms and 1490 drugs registered in DrugBank were contained [102]. Additionally, through combining DS6 and DS8, Luo *et al.* established a new dataset named Cdataset, in which 2352 known drug–disease associations between 663 drugs registered in DrugBank and 409 diseases listed in OMIM (marked as DS9) were included [80]. In 2017, Liang *et al.* built a database in which 3051 associations between 763 drugs and 681 diseases (marked as DS10) were contained [104]. Besides, Jin *et al.* contructed a database consisting of 6677 reported drug–disease associations between 1519 drugs and 1299 diseases (marked as DS11) [114]. In 2018, Zhang *et al.* provided two datasets including SCMFDD-S and SCMFDD-L for DgDsAs prediction [76], in which SCMFDD-S contains 18 416 associations between 269 drugs and 598 diseases (marked as DS12), while SCMFDD-L consists of 49 217 associations between 1323 drugs and 2834 diseases (marked as DS13). In 2020, Li *et al.* sorted out the drugs that directly affect diseases in CTD and built a new database including 26 521 associations between 4501 drugs and 2093 diseases (marked as DS14) [119].

Data related to MDgAs

In 2018, the Microbe–Drug Association Database (MDAD), a publicly accessible database containing 5505 associations between 180 microbes and 1388 drugs (marked as DS15) collected from 993 *pieces of* literature, was built first and utilized to predict potential microbe–drug associations [30]. And in the same year, Rajput *et al.* developed another database called aBiofilm as well, in which biological, chemical and structural details of 5027 anti-biofilm agents (1720 unique) reported from 1988–2017 were contained, and these 5027 agents target over 140 microorganisms (marked as DS16), including Gram-negative, Gram-positive bacteria and fungus [31]. Two years later, Andersen *et al.* summarized the activities and developmental statuses of 118 compounds/drugs that target 83 human viruses to construct a dataset named DrugVirus, in which 1281 associations (marked as DS17) were included [32]. All these above-mentioned datasets lay a solid foundation for researchers to study prediction of latent MDgAs. For instance, Long *et al.* integrated DS15, DS16 and DS17 into one new dataset for human microbe–drug association prediction [121].

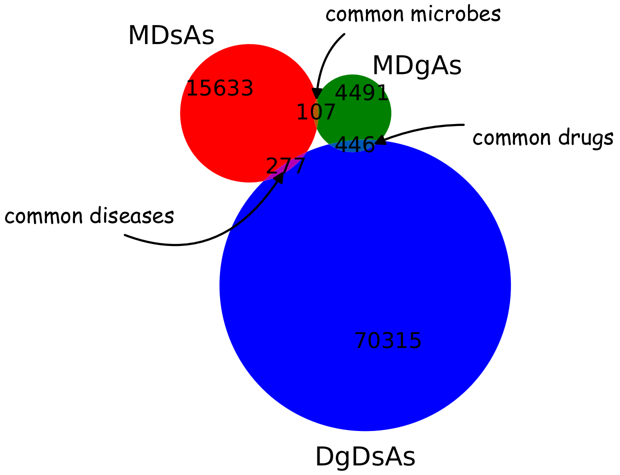

Figure 1. The relationships between datasets of MDsAs (DS18), DgDsAs (DS19) and MDgAs (DS20).

Data analysis

The biomedical data of MDsAs, DgDsAs and MDgAs are closely related due to their pairwise interactions; it is obvious that each of the above-mentioned database can be applied to different forecasting problems. Hence, in order to demonstrate that these 17 selected well-known datasets are closely related to each other, in this section, we constructed three new datasets DS18, DS19 and DS20 by analysing the intersection and union of above selected databases related to MDsAs, DgDsAs and MDgAs in Table 1, respectively. And after merging and removing redundant data, we finally obtained 15 633 associations between 2916 microbes and 787 diseases in DS18 based on DS1, DS2, DS3 and DS5, 70 315 associations between 5312 drugs and 4369 diseases in DS19 based on DS6, DS7, DS8, DS9, DS10, DS11, DS12, DS13 and DS14 and 4491 associations between 315 microbes and 2051 drugs in DS20 based on DS5, DS16 and DS17. Moreover, based on these three kinds of newly constructed datasets, we further obtained 107 identical microbes between DS18 and DS20, 277 diseases between DS18 and DS19 and 446 drugs between DS19 and DS20, separately. To better illustrate the relationships between DS18, DS19 and DS20, the Venn diagram of these three datasets is given in Figure 1.

Computational models

In recent years, a good deal of computational approaches have been proposed to predict potential relationships between microbes, drugs and diseases based on above-mentioned databases. In this section, we selected 93 representative state-of-the-art methods from existing mainstream literatures for discussion and comparison, and provided a brief summary of these methods in Table 2. For convenience, we firstly divided all these state-of-the-art methods into five major categories based on network, matrix factorization, matrix completion, regularization and neural network. And then, according to core ideas employed in these methods, we further divided them into

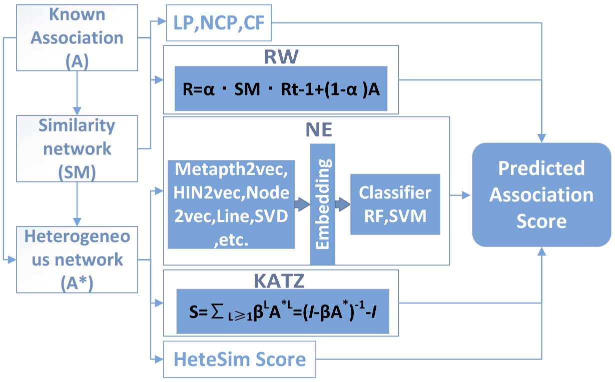

Figure 2. The main framework of network-based methods.

different subclasses and introduced each representative method related to different prediction problem in detail.

Network-based methods 基于网络的方法

As illustrated in Figure 2, the network-based methods mainly focus on adopting the topological information of different networks constructed based on different databases to infer latent correlations. According to different core techniques adopted in these predictive models, these network-based approaches can be further roughly classified into eight subcategories: KATZ Measure, Network Embedding (NE), Label Propagation (LP), Random Walk (RW) or the Bi-random Walk (BiRW), HeteSim Score, Weighted Network (WN), Network Consistency Projection (NCP) and Collaborative Filtering (CF), etc.

KATZ measure

The KATZ measure [143] computes the relative influence of a node within a network by measuring the numbers of both immediate neighbours and all other nodes in the network that connects to the node through these immediate neighbours, while connections established through distant neighbours are penalized by an attenuation factor.

Inspired by the KATZ measure, in terms of potential MDsAs prediction, Chen *et al.* developed a calculation model called KATZHMDA by adopting the Gaussian interaction profile (GIP) kernel similarities of both microbes and diseases [62].

Besides, in terms of MDgAs prediction, Zhu *et al.* proposed a computing model named HMDAKATZ through integrating the GIP kernel similarities of microbes with the chemical structure similarities of drugs [123].

Network embedding (NE)

The NE-based methods aim to learn low-dimensional embedding representations (vectors) for nodes in a network and then adopt these vectors as inputs of predictive models. In terms of MDsAs prediction, Lei *et al.* proposed a method called LGRSH by implementing Node2vec [144] to learn low-dimensional representations of microbes and drugs and calculated the correlations between microbes and diseases by adopting a modified rule-based inference method [40].

Besides, in terms of DgDsAs prediction, Yang *et al.* designed a method named HED through combining metapth2vec with a support vector machine classifier [79]. Zhou *et al.* introduced a method called NEDD by adopting HIN2vec [145]to learn network embedding vectors and trained a Random Forest (RF) classifier to infer potential DgDsAs [85]. Moreover, TS-SVD [89] and EMP-SVD [82] were designed to learn low-dimensional embedding representations of drug–disease pairs by using Singular Value Decomposition (SVD) based on a new drug–protein–disease heterogeneous network.

Moreover, in terms of MDgAs prediction, Long *et al.* proposed a method called HNERMDA by combining a network embedding approach named metapath2vec [146] with a bipartite network recommendation algorithm [124].

Label propagation (LP)

LP is a strategy that updates the labels of nodes by integrating the information of neighbours with the probability α and retaining the initial labels with the probability 1- α in a directed graph, in which the probability matrix P at step t will be updated as follows:

$$P^t = \alpha W P^{t-1} + (1-\alpha)A, \tag{1}$$

where W and A denote similarity matrix and association matrix, respectively. According to above formula (1), after convergence of iterations, the final probability matrix P can be obtained as follows:

$$P = (1-\alpha)(I-\alpha W)^{-1}A. \tag{2}$$

Based on the above LP strategy, in terms of MDsAs prediction, Yin *et al.* applied LP on the disease semantic similarity network and the microbe GIP kernel similarity network to infer potential MDsAs [47], in which the NCP [147] was adopted to calculate the projection scores from these two similarity networks. In addition, Zhang *et al.* proposed a prediction model called BDSILP by implementing LP on the integrated microbe similarity network and the integrated disease similarity network separately [35].

Besides, in terms of DgDsAs prediction, Zhang *et al.* proposed a model named NTSIM by applying the LP strategy on both the drug similarity network and the disease similarity network based on linear neighbourhoods similarity [77]. Lotfi *et al.* designed a heterogeneous LP method called Heter-LP and applied it on the drug–target–disease heterogeneous network to predict new DgDsAs [131].

Random walk (RW) or bi-random walk (BiRW)

The main idea of RW is to obtain the walking probability of each node by traversing a network starting with one node. The traverser will walk to the neighbour node with the probability of 1-α, and jump randomly to any other nodes in the network with the probability of α. Compared with RW, in BiRW, node starts walking at both ends of the network simultaneously. Luo *et al*. designed a model called NTSHMDA through employing the RW with restart algorithm to detect potential MDsAs [45]. Zou *et al*. developed a computational model named BiRWHMDA for latent MDsAs prediction by implementing the BiRW strategy on the microbe–disease heterogeneous network [61].

Besides, in terms of DgDgAs prediction, Chen *et al*. developed a predictive model called miRDDCR through applying the RW technique on both drug-miRNA network and miRNA-disease network simultaneously to infer potential DgDsAs [132]. Luo *et al*. proposed a method named MBiRW for latent DgDsAs detection by implementing RW on a newly constructed drug–disease heterogeneous network [80]. Moreover, Yan *et al*. applied BiRW on a newly established network by fusing multiple similarity measures to develop a computational approach called BiRWDDA to discover potential associations between drugs and diseases [93].

HeteSim score

HeteSim is a general framework for relevance measures in heterogeneous networks, which can effectively capture the subtle semantics of search paths [148]. The HeteSim Score between two nodes *s* and *t* in a heterogeneous network can be depicted as follows:

$$
\begin{aligned}
&HeteSim(s,t|R) = HeteSim(s,t|R1 \circ R2 \circ \cdots Rl)\\
&= \frac{1}{|O(s|R1)|I(t|Rl)|} \times \sum_{i=1}^{|O(s|R1)|}\sum_{j=1}^{|I(t|Rl)|}\\
&HeteSim(O_i(s|R1), I_j(t|Rl)|R2 \circ R3 \circ \cdot\cdot \circ Rl), \qquad (3)
\end{aligned}
$$

where *R* denotes a kind of path relation and *R*1∘ *R*2∘ ··· *Rl* denotes an associated path. *O*(*s*|*R*1) is the out-neighbours of *s* based on the relation *R1* and *I*(*t*|*Rl*) is the in-neighbours of *t* based on the relation *Rl*.

Based on above strategy of HeteSim Score, Fan *et al*. applied a normalized HeteSim Score on a newly constructed heterogeneous network to infer potential MDsAs [39].

Besides, in terms of DgDsAs prediction, Tian *et al*. applied the HeteSim Score of different meta-paths on the newly established network, and designed a novel predictive model named HeteSim_DrugDisease (HSDD) to detect possible DgDsAs [94].

Weighted network (WN)

The WN-based methods assign the *weight* score between nodes in the network first through similarity calculation or corresponding algorithm to construct a weighted network, and then predict potential associations based on the newly established weighted network. In terms of MDsAs prediction, Huang *et al*. designed a calculation model called PBHMDA through implementing a depth-first search algorithm on a weighted heterogeneous network [44]. In addition, Li *et al*. proposed a method named BWNMHMDA, in which a bidirectional weighted network was constructed first, and then a new recommendation algorithm and the KATZ *measure* was integrated to infer potential MDsAs based on the bidirectional weighted network [60].

Besides, in terms of DgDsAs prediction, Lee *et al*. proposed a predictive model by constructing a novel directed gene network and finding the shortest paths from target genes to disease genes in the directed network [135]. In addition, Oh *et al*. designed a detection model through establishing an integrative genetic network [134], in which drug–drug and disease–disease network adjacencies were firstly quantified by using weighted paths between target sets of them in the integrative genetic network, and then the distance between topological drug-module and disease (or disease-module and drug) would be calculated as the input features of RF classifier for possible DgDsAs prediction.

Network consistency projection (NCP)

Fan *et al*. proposed a novel model called HMDA-Pred through integrating multiple data types and adopted the NCP technique to predict latent MDsAs [52], in which the NCP algorithm worked from both perspectives of microbes and diseases. For example, the disease space projection score was calculated as follows:

$$
F(i,j) = \frac{MD(i,:) \cdot DS(:,j)}{|MD(i,:)|}, \qquad (4)
$$

where *MD(i,:)* represents the *i*th row of the adjacency matrix *MD* and *DS(:,j)* denotes the *j*th column of the disease similarity matrix *DS*. Certainly, the microbe space projection score can be calculated in a similar way.

Collaborative filtering (CF)

Recommendation algorithms based on CF can be mainly divided into two categories: one is user-based recommendation, which is dependent on the similarity of users, and the other is item-based recommendation, which is dependent on the similarity of items. Huang *et al*. proposed a method named NGRHMDA by adopting both the user-based and the item-based CF to compute the association possibilities for microbe–disease pairs, in which the microbes and diseases were regarded as ‘items' and ‘ users' separately in a newly constructed microbe–disease similarity network [67].

Advantages and disadvantages

Each model has its strengths and weaknesses; the KATZ measure could reconstruct potential associations simultaneously in a large-scale network, but the calculation of GIP kernel similarity will cause an inevitable bias towards those known associations. The LP and RW algorithms are efficient and straightforward to use, but most prediction

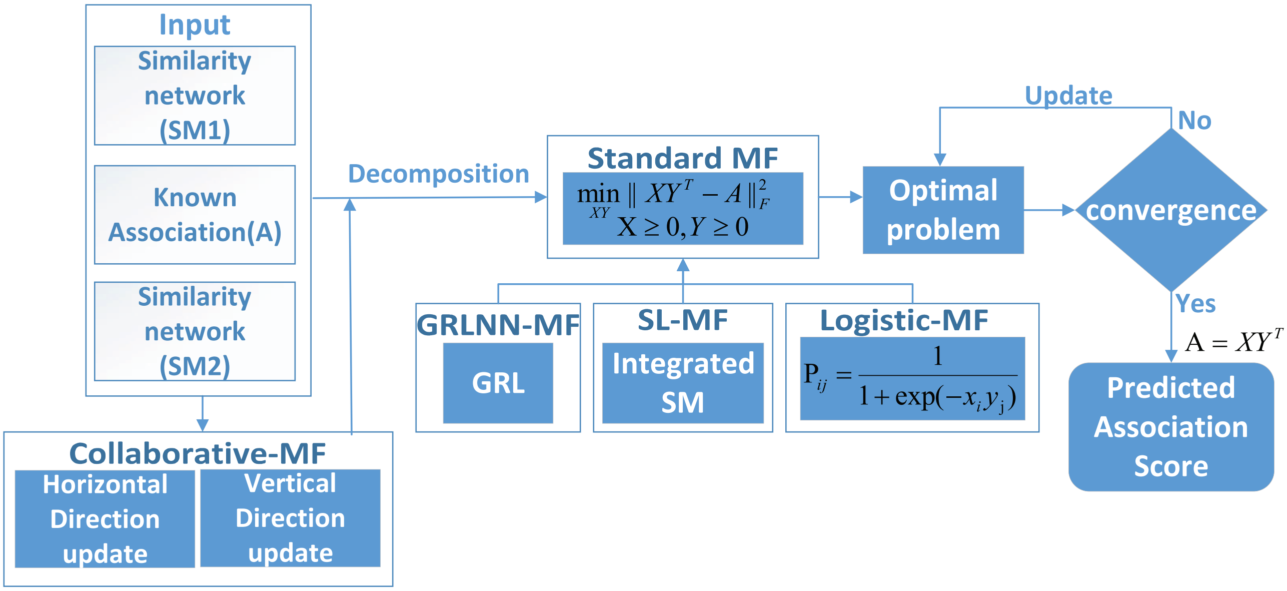

Figure 3. The main framework of matrix factorization-based methods.

methods based on them contain less biological information. The concept of meta path used in NE can explicitly capture essential high-order proximities because if the associations between two nodes are missing, their first-order proximity is zero. However, training the embeddings will increase difficulty when adding more information to the network. The WN-based and HeteSim-based methods have a solid ability to capture potential subtle semantic associations but cannot finish the prediction for microbe (drug, disease) without any known associations.

Matrix factorization (MF)-based methods 基于矩阵因子分解

As shown in Figure 3, the central idea of the MF-based methods is to decompose the input matrix into two low-dimensional matrices, and at the same time guarantee that the product of these two low-dimensional matrices will be approximately equal to the original input matrix. Generally, the mathematical framework of matrix factorization can be described as follows:

$$min_{X,Y}||XY^T - A||_F^2, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm, $A\in R^{m\times n}$ is the known association matrix, $X\in R^{m\times r}$ and $Y\in R^{n\times r}$ are the feature matrices of A and r denotes the subspace dimensionality. For convenience, in this section, we will divide selected representative matrix decomposition-based methods into five subcategories according to the objective function and optimization methods adopted by them for further analysis.

Graph regularized non-negative MF (GRLNN-MF)

In the GRLNN-MF-based methods, different regularization techniques were integrated into the original MF model to prevent overfitting of the model and the geometric structure of the similar network composed of different entities and nearest neighbour information was applied for prediction. Based on the technique of GRLNN-MF, He *et al.* and Liu *et al.* proposed two models named GRNMFHMDA [54] and NMFMDA [38] for possible MDsAs prediction in 2018 separately, in which the Tikhonov regularization [141] and the graph Laplacian regularization were adopted to prevent overfitting, respectively. The definition of the objective function of GRNMFHMDA and NMFMDA can be unified into the following form:

$$\begin{aligned}&min_{X,Y}||W - XY^T||_F^2 + \lambda_1(||X||_F^2 + ||Y||_F^2)+\\&\lambda_2\sum_{i,j=1}^{n}||x_i - x_j||^2 S_{ij}^{m*} + \lambda_3\sum_{p,q=1}^{m}||y_p - y_q||^2 S_{pq}^{d*}\\&s.t. X \geq 0, Y \geq 0,\end{aligned} \tag{6}$$

where W is the known association matrix, X and Y represent microbe and disease features separately; the parameters of λ_1, λ_2 and λ_3 represent the relevant regularization coefficients; x_i and y_p are defined as i_{th} rows of X and j_{th} rows of Y; $S^{m*}$and S^{d*} are introduced to the predictive model to avoid adverse effects. The derivation of the Lagrange function was applied to optimize the objective function and sought the most suitable matrix of X and Y, respectively. Notably, different from NMFMDA, the Weighted K Nearest Known Neighbors (WKNKN) was incorporated to improve the prediction accuracy of GRNMFHMDA.

Besides, in terms of DgDsAs prediction, Xuan *et al.* developed a non-negative matrix factorization-based approach named DisDrugPred by integrating multi-source data of both drugs and diseases [105], in which the l2-regularization on the projection matrices was added to prevent overfitting of model.

Logistic-MF

The logistic MF model is a probabilistic model for MF with implicit feedback, in which the probabilistic distribution was learned by using a collaborative filtering recommender algorithm [149]. In terms of MDsAs prediction, Peng *et al.* developed a computing model named RNMFMDA [69], in which the logistic matrix factorization with neighbourhood regularization was utilized to compute the probabilities of associations for all microbe–disease pairs. The final optimization model is as follows:

$$\begin{aligned} &min_{A,B}\sum_{i=1}^{m}\sum_{j=1}^{n}(1+cy_{ij}-y_{ij})ln[1+exp(a_ib_j^T)]-cy_{ij}a_ib_j^T\\ &+\frac{1}{2}tr[A^T(\lambda_mI+\alpha L_m)A]+\frac{1}{2}tr[B^T(\lambda_dI+\alpha L_d)B].\\ &p_{ij}=\frac{exp(a_ib_j^T)}{1+exp(a_ib_j^T)}, \end{aligned} \tag{7}$$

where A and B represent microbe feature matrix and disease feature matrix, respectively; L_m and L_d were defined as the same to Liu *et al.* [150]. p_{ij} is the association probability between two nodes and tr($\cdot$) denotes the trace of a matrix. Finally, the predictive matrix Y can be calculated as follows:

$$Y = AB^T. \tag{8}$$

Besides, in terms of DgDsAs prediction, Wang *et al.* proposed a method named DDAPRED by adopting the regularized logistic matrix factorization method [109].

Similarity constrained MF (SC-MF)

In the SC-MF-based predictive methods, various biological information and multiple types of similarities were adopted first, and then these different similarities would be introduced into the objective function as constraints. In terms of DgDsAs prediction, Zhang *et al.* proposed a framework named SCMFDD based on the SC-MF [76]. Different from SCMFDD in which only single drug feature-based similarity was utilized Yang *et al.* [92] developed a multi-similarities bilinear MF called MSBMF by incorporating various biological information including the chemical structures, anatomical therapeutic codes, target profiles, drug-drug interaction and side effects to measure multiple drug similarities.

Kernelized bayesian MF (KB-MF)

The main idea of KB-MF is to use two projection matrices to project biomarkers kernel matrices to a unified low-dimensional subspace first, then estimate their association based on these two low-dimensional spaces U and V and finally, the known association matrix Y is generated from the interaction fraction matrix F [151]. In terms of MDsAs prediction, Chen *et al.* proposed a novel approach based on KBMF [41], in which the Bayesian algorithm was adopted to infer potential MDsAs through constructing a fully conjugate probabilistic model and designing an inferred deterministic variational approximation mechanism.

Collaborative-MF

The purpose of the Collaborative-MF-based method is to iteratively update two decomposed matrices by taking partial derivatives of the objective function. In Collaborative-MF, these two decomposed matrices are randomly initialized, but proper initialization is helpful to accelerate the convergence. In terms of MDsAs prediction, in 2017, Shen *et al.* proposed a computational model based on the Collaborative-MF for Human Microbe–Disease Association prediction (CMFHMDA) [57].

Besides, in terms of DgDsAs prediction, in 2019, a Collective-MF-based multi-task learning method named CMFMTL was designed by Huang *et al.*, in which each task predicts a type of correlation, and these two tasks complement and improve each other by capturing the correlation between them [116].

Advantages and disadvantages

The methods based on matrix factorization can mine deeper potential connections. Meanwhile, matrix factorization has a relatively low spatial complexity because mapping a high-dimensional matrix into two low-dimensional matrices saves storage space. However, the MF-based methods usually contain more parameters. Therefore, selecting the optimal parameters is more challenging, and model training is time-consuming. Furthermore, the models based on matrix factorization is not interpretable. They only update the observed items when optimizing the model and does not consider the unobserved things.

Matrix completion (MC)-based methods

基于矩阵补全的方法

As illustrated in Figure 4, matrix completion aims to restore a matrix with missing values to a complete matrix through decomposing a matrix with missing values into two or more matrices through matrix factorization and then multiplying these decomposed matrices to obtain an approximate matrix of the original one. Some of these methods may also belong to the matrix decomposition category.

In terms of MDgAs prediction, in 2018, Shi *et al.* put forward a predictive model named BMCMDA based on the Binary Matrix Completion [59]. Additionally, in 2019, Yan *et al.* designed a low-rank matrix completion method called MCHMDA through integrating similarities of microbes and diseases with known MDsAs into a heterogeneous network [50], in which a MC-based method was utilized to derive the association scores of unknown microbe–disease pairs by the fast Singular Value Thresholding algorithm [153].

Besides, in terms of DgDsAs prediction, in 2018, Luo *et al.* designed a drug repositioning recommendation system (DRRS) to infer novel drug indications by incorporating related data sources with extra-biological information of drugs and diseases, in which a low-rank matrix

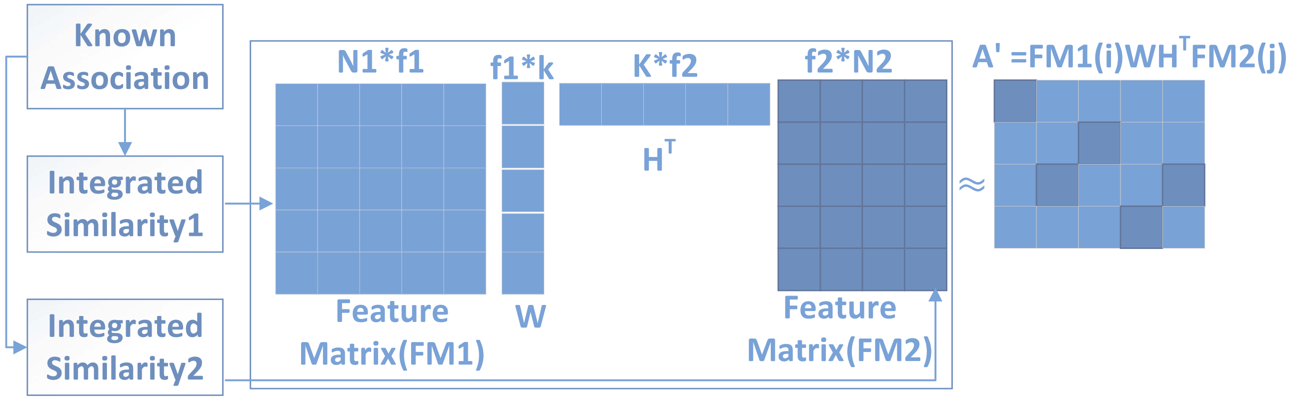

Figure 4. The main framework of matrix completion-based methods.

approximation and a randomized algorithm were utilized to discover new DgDsAs [81]. In 2020, Yang *et al.* proposed a heterogeneous graph inference with matrix completion named HGIMC for drug repositioning [90] through adding more positive and formative drug–disease edges between drug network and disease network to pad a part of the missing entries [91].

Advantages and disadvantages

There are three types of matrix completion models in link prediction: MC model based on kernel norm relaxation, MC model based on matrix decomposition and MC model based on non-convex function relaxation. The advantage of the MC model based on kernel norm relaxation is that it belongs to the convex optimization model, there is a globally optimal solution and the kernel norm nearest neighbour operator has a closed resolution, but the explanation of the model involves complex singular value decomposition, the solution efficiency is limited and the kernel norm cannot closely approximate the actual rank of the target matrix. The MC model based on matrix decomposition avoids the complex matrix singular value decomposition and can be implemented distributed, but it belongs to non-convex optimization and may have the non-global optimal solution. The MC model based on non-convex function relaxation has good completion performance, but it also belongs to non-convex optimization, and there may be a non-global optimal solution.

Regularization (RL)-based methods

RL-based methods aim to establish different regularized least-squares classification, a kind of kernel-based square loss regularization network, to solve different predictive tasks, whose generalization performance is severely influenced by the setting of its kernel and hyper parameters. In terms of MDsAs prediction, Wang *et al.* proposed a novel computing model named LRLSHMDA in 2017 through adopting the Laplacian regularized least squares (LapRLS) classifier [51]. In addition, Xu *et al.* proposed an identification model called MDAKRLS in 2021 by integrating the Kronecker regularized the least square with the Hamming interaction profile similarity to calculate the interaction profile similarities of microbes and diseases [152].

Besides, in terms of DgDsAs prediction, Liang *et al.* designed a calculation model called LRSSL through combining the LapRLS model with the drug chemical information, the drug target domain information and the target annotation information [104]. In addition, an RL-based computing method named RLSDR was proposed to discover new uses of drugs by adopting a semi-supervised learning model [98]. Through combining known DgDsAs with the drug–drug similarity kernel and the disease–disease similarity kernel into a unified and global regression analysis framework, Lu *et al.* developed an RL-based computational tool called DR2DI to infer potential DgDsAs [100].

In terms of MDgAs prediction, in 2021, Zhu proposed a new computational approach called LRLSMDA [125], in which, based on the LapRLS algorithm, two objective functions would be computed by using the minimization of the cost functions and further transformed into a prediction matrix with a linear mean method.

Advantages and disadvantages

The regularization methods generated fewer model parameters, thus saving time and improving robust performance. Meanwhile, the RL-based models have a strong capability of fitting and generalization. Those models enhance the interpretability of the model by introducing regularization terms. However, regularization could bring computational complexity, and it is easy to make the model underfitting by adding penalty terms in regularization. Moreover, models based on regularization are hard to calibrate.

Neural network (NN)-based methods

In this section, selected NN-based methods will be further divided into six different subclasses such as Convolution neural network (CNN), Graph Convolution Network (GCN)-based autoencoder, Graph Attention Network (GAT)-based autoencoder, Collective Variational Autoencoder(CVAE), Sparse Autoencoder(SAE) and Deep Neural Network(DNN) for detailed analysis. The main

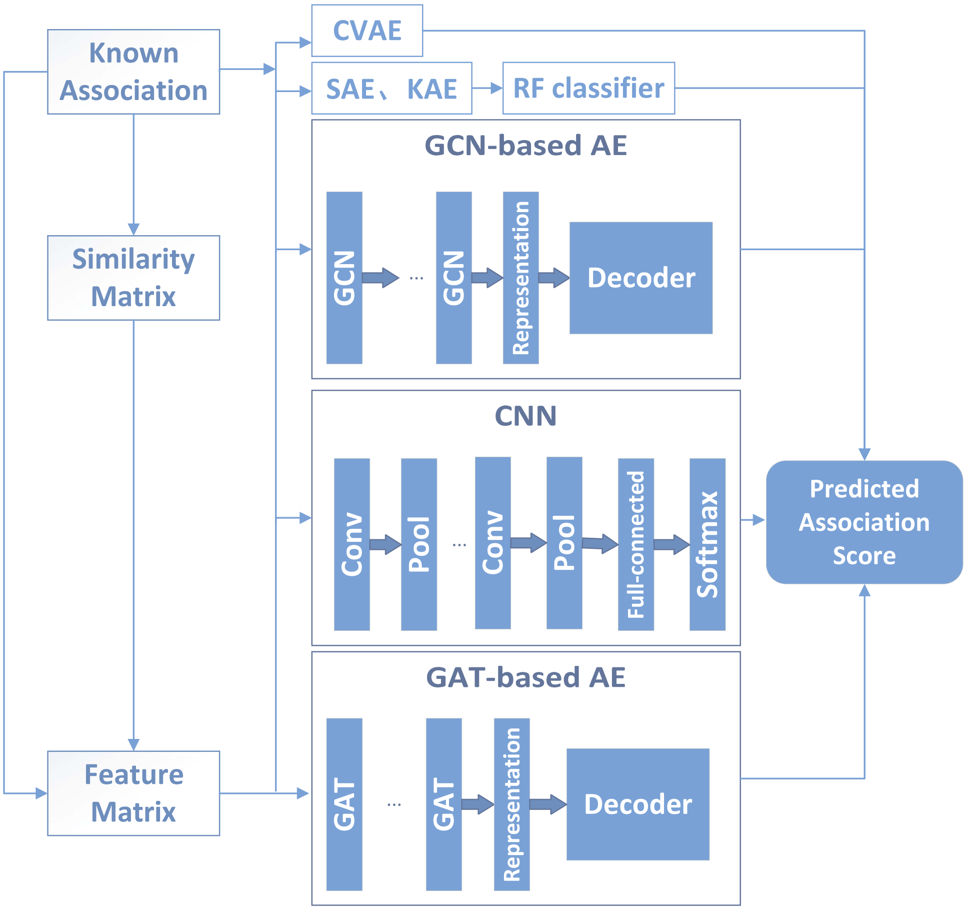

Figure 5. The main framework of neural network-based methods.

framework of neural network-based methods is shown in Figure 5.

Convolution neural network (CNN)

The basic structure of CNN consists of input layer, convolutional layer, pooling layer, full-connected layer and output layer. The convolutional layers and pooling layers are combined alternately to extract features, which are learned by the full-connected layer. The convolutional operation in the ith layer can be described as follows:

$$\alpha_i = \sigma(\alpha_{i-1} \otimes W_i + b_i), \tag{9}$$

where W_i represents the weight matrix of the convolution kernel; ⊗ represents convolution; b_i is the offset vector;σ is the activation function. The pooling operation in the pooling layer α_i is conducted as follows:

$$\alpha_i = subsampling(\alpha_{i-1}). \tag{10}$$

In terms of DgDsAs prediction, Jiang *et al.* designed a method called SKCNN through combining the Sigmoid Kernel techniques with the CNN [86], in which the convolution and max-pooling operations with a kernel size of 16×16 for the convolutional layer and 2×2 for

需要看论文查看这个做法是怎么实现的，或许我们可以
通过 CNN 学习到疾病之间的相似矩阵。

the subsampling layer were conducted on the inputs of integrated drug or disease similarities to extract features of drug and disease separately, and moreover, the sigmoid function was selected as the activation function, and the random forest was set as the classifier to train the features learned by CNN. In addition, Xuan *et al.* designed a method named CBPred based on a CNN and bidirectional long short-term memory (BiLSTM) [141], in which two CNN layers were constructed to get the final predicted score.

GCN-based autoencoder

Autoencoder is a specific neural network structure whose purpose is to represent the vertices of a graph as low-dimensional vectors using a neural network structure. It mainly consists of the encoder and decoder, where the encoder is used to obtain the node embedding, and the decoder is used to reconstruct the node's neighbourhood statistics.

就是图里面所
有的节点

The graph convolutional network is the graph extension of convolutional neural network for learning low-dimensional representations of nodes and is often used as the encoder in graph autoencoders,in which each layer aggregates neighbour's information to reconstruct the embeddings as the inputs of the next layer. The layerwise propagation rule of GCN can be formulated as follows:

$$H^{(l+1)} = f(H^{(l)}, G) = \sigma(D^{-\frac{1}{2}} G D^{-\frac{1}{2}} H^{(l)} W^{(l)}), \tag{11}$$

where $H^{(l)}$ is the embeddings of nodes at the lth layer, G is the adjacency matrix and D is the degree matrix of G. σ is the activation function.

In terms of MDsAs prediction, Ma *et al*. developed a GCN-based mining model named NinimHMDA by integrating different prior biological knowledge to predict different types of MDsAs [46], in which the encoder was first built by chaining multiple non-linear neural network layers and node embedding update layers to map each node to an embedding, and then the matrix completion technique was used in the decoder to reconstruct the edge weight of the adjacency matrix to predict scores of multiple types of MDsAs by using the learned node embeddings.

这个代码是开源的，也可以学习查看

In terms of DgDsAs prediction, a GCN-based encoder with three graph convolution layers was built to learn the low-dimensional representations of drugs and diseases in LAGCN [117], in which, an attention mechanism was first introduced to combine these embeddings, and then a bilinear decoder [157] was utilized to reconstruct the adjacency matrix to obtain the predicted scores of DgDsAs. Different from LAGCN, in which network topology information from different domains, such as drug and disease, was mixed together without distinction in the heterogeneous network, the GCN-based encoder combined with layer attention mechanism in DRHGCN adopted two feature extraction modules including the intra-domain and the inter-domain modules to extract features [95]; here, the intra-domain module was based on a drug–drug and a disease–disease similarity networks, while the inter-domain module was based on a drug–disease association network only.

这两个感觉也可以学习利用下

In terms of MDgAs prediction, Huang *et al*. proposed a method named Graph2MDA [120] based on variational graph autoencoder (VGAE) [158], in which a two-layer GCN-based encoder was adopted to learn the low-dimensional representations, and a deep neural network-based classifier was designed to infer potential MDgAs by using the learned latent representations. Moreover, Long *et al*. presented a GCN-based framework called GCNMDA through combining a GCN-based encoder with a Conditional Random Field (CRF) layer and a decoder to predict possible MDgAs [122], in which the CRF [159] with an attention mechanism was designed in the hidden layer of GCN to guarantee that similar nodes have similar representations.

这个也可以再学习下，也有开源代码：Graph2MDA

GAT-based autoencoder

The attentional mechanism is prevalent these days, which can amplify the impact of the essential parts of the data. GAT is a spatial-based graph convolution network. Its attentional mechanism is used to determine the weights of nodes' neighbourhoods when feature information is aggregated. The GAT-based encoder has been widely used in possible MDgAs and MDsAs predictions in recent years. For instance, in terms of MDsAs prediction, Long *et al*. proposed a novel framework based on graph attention networks with inductive matrix completion (GATMDA) [42], in which an optimized graph attention network with talking heads was exploited to learn representations for diseases and microbes. MGATMDA is a multicomponent GAT-based framework [71], in which the edges in the microbe–disease bipartite graph were decomposed first by node-level attention mechanism, then recombined by component-level attention mechanism and finally, a full-connected network was utilized to predict potential MDsAs.

In terms of MDgAs prediction, Long *et al*. presented an ensemble GAT framework called EGATMDA [121], in which three different networks, including the microbe–drug bipartite network, the microbe–drug heterogeneous network and the microbe–disease–drug heterogeneous network, and two types of attention mechanisms were included.

Collective variational autoencoder (CVAE)

The CVAE model was proposed by Chen *et al*. that complements the sparse ratings with side information, as feeding side information into the same VAE [160]increases the number of samples for training [161]. In terms of DgDsAs prediction, Zeng *et al*. proposed a method named deepDR by adopting a multimodal deep autoencoder and the CVAE to infer new DgDsAs [112], in which the features of drugs were encoded and decoded via the CVAE to infer candidate diseases for drugs. In addition, Tnj *et al*. introduced a method called SNF-CVAE by adopting the SNF technology and the CVAE to conduct drug–disease interaction prediction [142], in which multiple types of drug similarity networks were integrated in SNF first, and then a non-linear CVAE was trained by using the integrated drug similarity and known DgDsAs to predict novel interactions.

Sparse autoencoder (SAE)

In order to ensure the sparsity of the hidden layer, SAE adds a penalty term based on the autoencoder. The penalty term can be expressed as follows:

$$P_{penalty} = \sum_{t=1}^{S_2} KL(\rho || \widehat{\rho}_t), \tag{12}$$

where S_2 is the number of neurons in the hidden layer, $\widehat{\rho}_t$ represents the average activity of hidden neurons t, $KL(\rho||\widehat{\rho}_t)$ is the relative entropy between two Bernoulli random variables with mean ρ and mean $\widehat{\rho}$ and is defined as follows:

$$KL(\rho || \widehat{\rho}_t) = \rho log \frac{\rho}{\widehat{\rho}_t} + (1-\rho) log \frac{1-\rho}{1-\widehat{\rho}_t}. \tag{13}$$

In terms of DgDsAs prediction, Jiang *et al*. designed a novel calculation model named SAEROF by combining

SAE and rotation forest to predict latent DgDsAs [84], in which a feature extraction module based on SAE and principal component analysis was built first, and then the rotation forest classifier was adopted to deal with the extracted features for final prediction.

Deep neural network (DNN)

The DNN model includes three parts: one input layer, several hidden layers and one output layer, in which each hidden layer extracts more and more generalized features based on the output of the previous layer. In 2020, Liu *et al*. designed a prediction model called HNet-DNN by utilizing a DNN to predict DgDsAs based on the features extracted from the drug–disease heterogeneous network [88], in which four hidden layers were adopted, ReLu was selected as the activation function, the cross entropy was employed as the loss function and the Adam was used to train the model.

Advantages and disadvantages

Neural networks have been widely used in the field of prediction. Compared with traditional neural networks, CNN has a parameter sharing mechanism to avoid overfitting and achieve better performance effectively. But the pooling layer will lose a lot of valuable information and ignore the correlation between the local and the whole. GCN improves the inapplicability 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 neighbours and is sensitive to parameter initialization. SAE can effectively learn important features, suppress secondary characteristics and extract abstract features with lower dimensions and more sparse, but it cannot specify whether a particular node was active or hidden, and sparsity parameters were poorly set. CVAE can generate specific data by selecting tags, but its generality is weak.

Experimental comparisons

In this section, we conducted extensive experiments for assessing selected predictive methods in the fields of MDsAs, DgDsAs predictions. Due to the lack of data and code reproductivity, we did not perform experimental comparison of MDgAs prediction. And as a result, seven state-of-the-art predictors, including KATZHMDA, NBLPIHMDA, BiRWMP, BPNNHMDA, NTSHMDA, HMDA-Pred and LRLSHMDA, were experimentally compared for prediction of latent MDsAs, while five state-of-the-art models, including MBiRW, MSBMF, DR2DI, BNNR, DRHGCN, were implemented for comparison of potential DgDsAs prediction. Moreover, two commonly used evaluation frameworks of leave-one-out cross-validation (LOOCV) and 10-fold cross-validation (10-fold CV) were adopted, and the receiver operating characteristic curve (ROC) with the area under this curve (AUROC) was selected to evaluate the performances of these approaches. Here, ROC depicts the actual positive rate against the false-positive rate at various thresholds. Due to the data distribution being extremely uneven, and the proportion of positive samples being much higher than that of negative samples, we utilized the precision-recall curve (PR) with the area under this curve (AUPR) to further evaluate the performance of different methods. Since precision and recall are often in tension, and improving precision typically reduces recall and vice versa. To fully evaluate the effectiveness of the existing models, we further used the F1-score that is the harmonic mean of the precision and recall assessing these models' performance. Furthermore, while conducting LOOCV, each known association will be rotated as the test set in turn, while the remaining known associations act as the training set.

Thereafter, we can obtain the prediction scores and sort all the predicted values by putting these data into the model. If the prediction score is higher than the given threshold, it will be a successful prediction. Different TPR and FPR can be obtained when setting different thresholds. Subsequently, taking TPR and FPR under different thresholds as Y-axis and X-axis, respectively, ROC can be further drawn, and the area under the ROC line (AUROC) can be taken to evaluate the prediction performance of different methods. Like LOOCV, 10-fold CV divides the data set into 10 parts, and each turn takes nine copies as training data and one copy as test data. Owing to the randomness of dividing samples, we repeated 10-fold CV 10 times to calculate the average value of all AUCs as the final result.The code and data for reproducing experimental results can be found on GitHub (https://github.com/Jappy0/microbe-drug-disease).

MDsAs experimental analysis

To identify predictive performances of different algorithms, in this section, we would conduct the comparative experiments based on DS1, the most used MDsAs related database, for MDsAs prediction first. And then, seven typical MDsAs methods, including KATZHMDA, NBLPIHMDA, BiRWMP, BPNNHMDA, NTSHMDA, HMDA-Pred and LRLSHMDA, would be chosen to conduct comparative experiments under the frameworks of LOOCV and 10-CV separately. As a result, we obtained the ROCs with AUROCs presented in Figure 6a and Figure 6c under the frameworks of LOOCV and 10-fold CV, respectively. Meanwhile, we drew PR curves and calculated the corresponding AUPRs that are illustrated in Figure 6b and Figure 6d under the framework of 10-fold CV, respectively. Furthermore, we set the threshold as 1000 to calculate F1-score to quantitatively analyse the advantages and disadvantages of these competitive methods under the framework of 10-fold CV. Finally, the values of AUC, AUPRs and F1-scores of all compared methods obtained by LOOCV and 10-fold CV were summarized in Table 3.

Moreover, to better analyse the factors that affect the predictive performances of different models, we summarized the original datasets, additional similarity

Table 3. Performances of seven MDsAs prediction methods evaluated under the frameworks of LOOCV and 10-fold CV. KATZHMDA: KATZ measure for Human Microbe–Disease Association prediction, NBLPIHMDA: Bidirectional Label Propagation for Human Microbe–Disease Association prediction, BiRWMP: Predicting potential microbe–disease associations by bi-random walk on the heterogeneous network, BPNNHMDA: Back-Propagation Neural Network for Human Microbe–Disease Association prediction, NTSHMDA: Predicting human microbe–disease associations based on random walk by integrating network topological similarity, HMDA-Pred: Multi-data integration and network consistency projection for Human Microbe–Disease Associations Prediction, LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe–Disease Association, NN: Neural Network, RL: Regularization.

		LOOCV			10-fold CV		
Methods	Major category	AUROC	AUPR	F1-score	AUROC	AUPR	F1-score
KATZHMDA	Network-based	0.8382	0.6178	0.4041	0.8354	0.6131	0.3972
NBLPIHMDA	Network-based	0.9041	0.6885	0.4400	0.9000	0.6784	0.4345
BiRWMP	Network-based	0.8637	0.6733	0.4483	0.8601	0.6557	0.4510
BPNNHMDA	NN	0.9281	0.7301	0.4593	**0.9188**	**0.7293**	**0.4552**
NTSHMDA	Network-based	0.9070	0.6947	0.4290	0.8882	0.6728	0.4289
HMDA-Pred	Network-based	**0.9589**	**0.8416**	**0.5462**	0.8841	0.4808	0.3324
LRLSHMDA	RL	0.8909	0.6384	0.3972	0.8873	0.6217	0.4093

不是有几种
没有开源代
码吗？那是
怎么训练跑
的实验

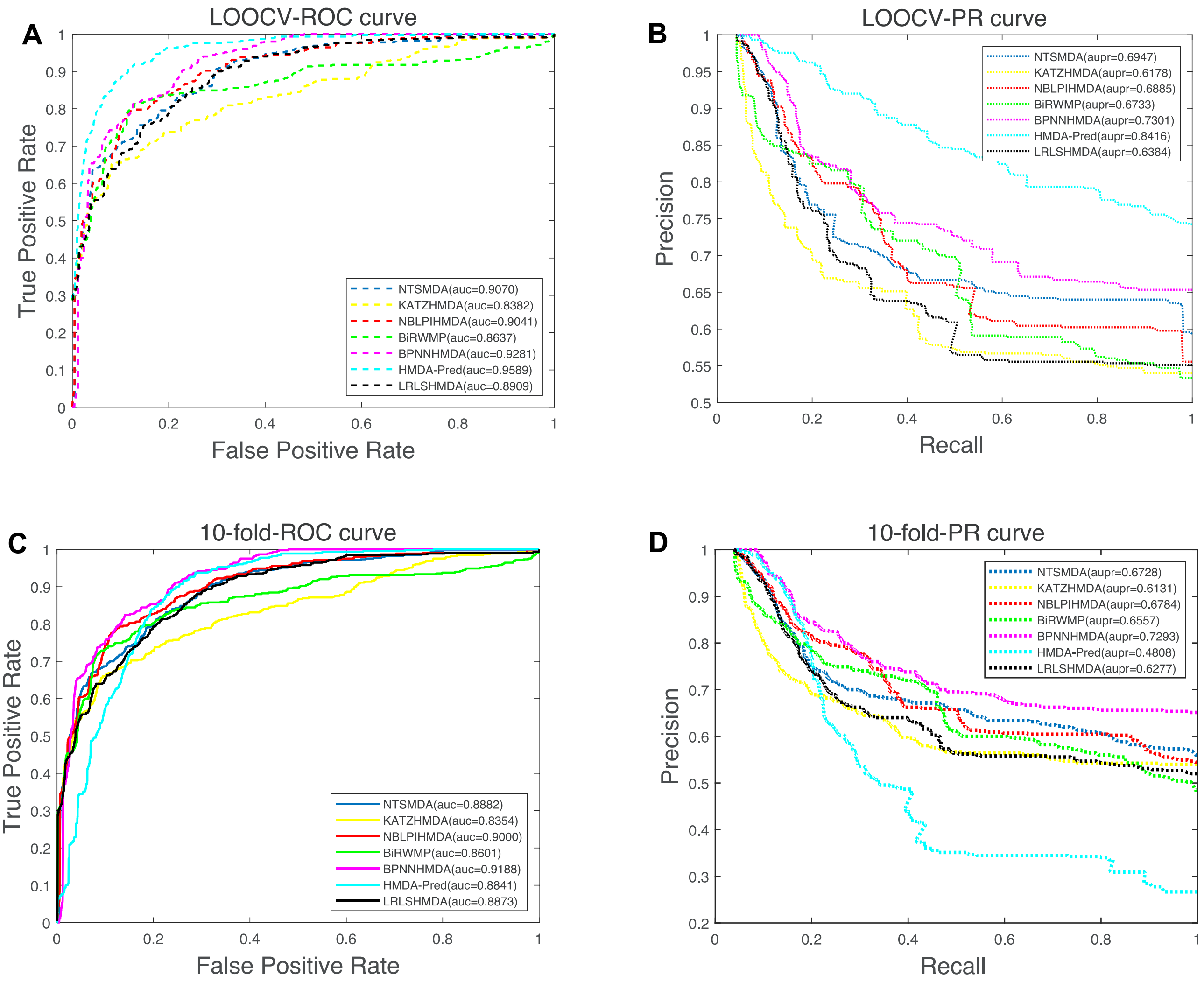

Figure 6. ROCs and PR curves of seven state-of-the-art MDsAs prediction methods. **(A)**: ROCs and AUROCs conducted in LOOCV. **(B)**: PR curves and AUPRs obtained in LOOCV. **(C)**: ROCs and AUROCs implemented in 10-fold CV. **(D)**: PR curves and AUPRs obtained in 10-fold CV.

features and calculation approaches of the seven state-of-the-art methods in Table 4. According to the literatures, the parameters of these approaches were set to default values to derive optimal predictive performance.

From observing the experimental results illustrated in Table 4, it is evident that the network-based methods represented by HMDA-Pred can achieve the best performance in LOOCV, and the excellent performance

Table 4. Original datasets, additional similarity features and calculation approaches utilized in the seven state-of-the-art approaches for MDsAs prediction. KATZHMDA: KATZ measure for Human Microbe–Disease Association prediction, NBLPIHMDA: Bidirectional Label Propagation for Human Microbe–Disease Association prediction, BiRWMP: Predicting potential microbe–disease associations by bi-random walk on the heterogeneous network, BPNNHMDA: Back-Propagation Neural Network for Human Microbe–Disease Association prediction, NTSHMDA: Predicting human microbe–disease associations based on random walk by integrating network topological similarity, HMDA-Pred: Multi-data integration and network consistency projection for Human Microbe–Disease Associations Prediction, LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe–Disease Association, NN: Neural Network, RL: Regularization.

			Similarity	
Method	**Dataset**	**Major category**	**Microbe**	**Disease**
KATZHMDA	DS1	Network-based	GIP	GIP, symptom-based
NBLPIHMDA	DS1	Network-based	GIP	GIP
BiRWMP	DS1	Network-based	GIP	GIP
BPNNHMDA	DS1	NN	GIP	No
NTSHMDA	DS1	Network-based	GIP	GIP
HMDA-Pred	DS1	Network-based	GIP, Cosine similarity	GIP, Cosine similarity
LRLSHMDA	DS1	RL	GIP	GIP

of this model may be derived from the following reasons:

(1) HMDA-Pred fully leveraged multiple similarities of disease and microbe, including the Gaussian interaction profile kernel similarity and the Cosine similarity.

(2) HMDA-Pred proposed a linear network fusion method to fuse multiple similarity networks and derived an informative matrix cleverly.

(3) Network consistency projection conducted on microbe and disease spatial networks is effective.

In addition, it can be seen as well that BPNNHMDA can achieve the best performance in 10-fold CV. Through analysis, the main reasons that why it can outperform other five competing methods may be due to the following three points:

(1) The initial edge weight of the GIP kernel similarity-based network in BPNNHMDA can effectively improve the training efficiency.

(2) In the training process of BPNNHMDA, the input data, edge weights and bias are strictly standardized to ensure training stability.

(3) A new activation function is designed in BPNNHMDA to make up for the shortcomings of the traditional activation function.

According to Table 3 and Table 4, it is easy to see that fusing multiple similarities of microbes and diseases can improve the predictive performance of models significantly. Moreover, it can be seen as well that in all these seven competitive methods, the GIP kernel similarity is adopted to measure the similarity between microorganisms and diseases; through analysis, it can be found that the similarity scores computed by the GIP kernel similarity are closely located in the neighbourhoods of 0 and 1, having a higher distinguishability [33]. In addition, models integrated effectively with prior knowledge can usually achieve better performance than using the GIP kernel similarity only. For example, when incorporated in symptom-based disease similarity, KATZHMDA achieved an AUROC of 0.8644 in LOOCV, which is much higher than the AUROC of 0.8382 achieved by the original model. As a result, multiple source data and various similarity calculation methods need to be an inlet to improve performance and model generalization ability. For example, other data for similarity, including symptom-based disease similarity, microbe functional similarity and disease semantic similarity, have been successfully applied to the MDsAs prediction [24, 25, 29, 32, 36, 37, 40, 145]. Another measure to improve prediction reliability is to define the classification level for each microbe and then excecute prediction at the same level. Furthermore, the introduction of taxonomy is conducive to accurately identifying microbes in microbiological data, contributing to integrating microbiomes, such as microbial genome sequence and patient-derived microbial metagenome, transcription and metabolism, into MDsAs prediction. Finally, most existing models cannot solve the prediction of new diseases and microbes without any known association; this problem can be solved by introducing similarity without relying on the known topology information of the microbial disease association network.

整合疾病症状相似性，可以带来
预测精度的提高

因此需要采用
多源数据和相
似度计算

但是，怎
么更好的
整合这些
相似度，
应该是一
个待分析
的重要任
务

DgDsAs experimental analysis

In this section, we selected five representative methods, such as MBiRW, MSBMF, DR2DI, BNNR and DRHGCN, to compare their performance of latent DgDsAs prediction based on the DS6, the most commonly used DgDsAs related database. As for target proteins of drugs in DS6, we obtained the protein domain and gene ontology information from the InterPro and the UniProt databases, respectively. And then, based on these data, we introduced the original similarity method in DR2DI to calculate the drug target domain-based similarity and drug target-go based similarity for DR2DI, while the disease semantic similarity is replaced with disease phenotype similarity in DR2DI because there is no guarantee that all diseases in the DS6 have corresponding semantic trees in the Mesh to calculate semantic similarity. Moreover, in order to achieve the best predictive performance, during

Table 5. Performances of five DgDsAs related competing prediction methods evaluated under the framework of 10-fold CV. MBiRW: Comprehensive similarity measures and Bi-Random Walk algorithm, MSBMF: Multi-similarities bilinear matrix factorization, DR2DI: A powerful computational tool for predicting novel drug–disease associations, BNNR: Bounded nuclear norm regularization, DRHGCN: Drug Repositioning based on the Heterogeneous information fusion Graph Convolutional Network, MF: Matrix Factorization, RL: Regularization, MC: Matrix Competion, NN: Neural Network.

Method	Major category	AUROC	AUPR	F1-score
MBiRW	Network-based	0.9188	0.4435	0.4848
MSBMF	MF	**0.9507**	**0.6144**	**0.6057**
DR2DI	RL	0.9118	0.4261	0.4627
BNNR	MC	0.9347	0.5993	0.5858
DRHGCN	NN	0.9480	0.5694	0.5677

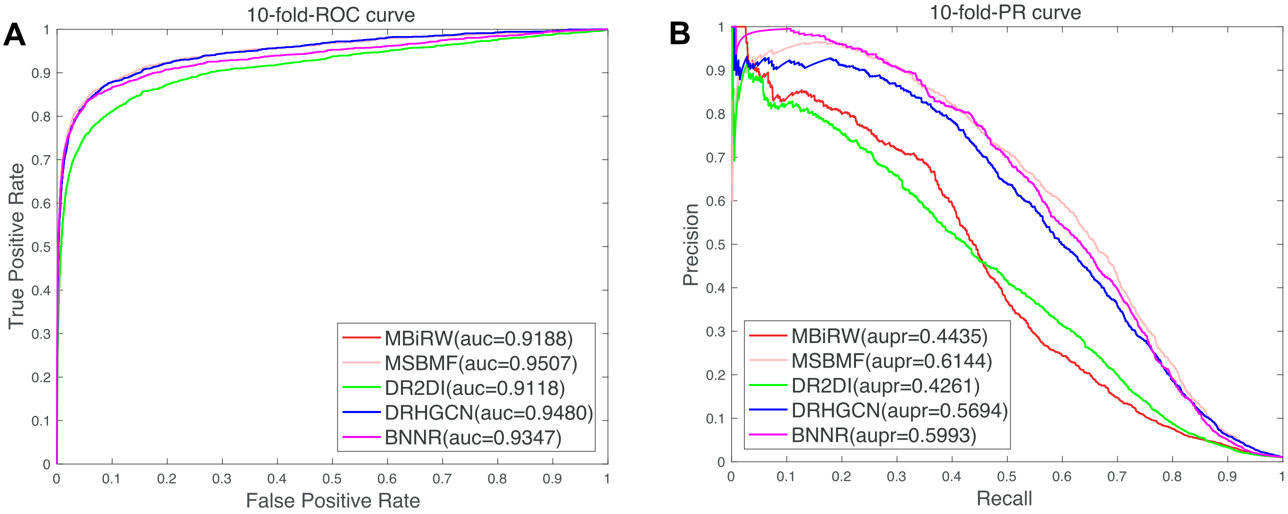

Figure 7. ROCs and PR curves of all comparative DgDsAs related methods in 10-fold CV. **(A)**: ROCs and AUROCs achieved by five competitive methods in 10-fold CV. **(B)**: PR curves and AUPRs achieved by five competitive methods in 10-fold CV.

experiments, we settled with the original parameters for all these competitive methods and evaluated them under the framework of the 10-fold CV adopted in DRHGCN, which is a little different from the 10-fold CV mentioned above. In the new type of 10-fold CV, we randomly selected 10% of known associations and 10% of unknown associations as the testing set, and the remaining 90% of known associations and unknown associations as the training set. The ROCs and PR curves were subsequently drawn and illustrated in Figure 7a and Figure 7b, respectively. Meanwhile, the AUROCs, AUPRs and F1-scores of all compared methods were shown in Table 5.

Similarly, to better analyse the factors affecting the performance of prediction models by combining the experimental results with the prediction model method, we summarized the original datasets and similarity calculation approaches of these carefully selected methods in Table 6.

According to Table 5, it is easy to find that the network-based method MBiRW and RL-based method DR2DI are slightly inferior to other competing methods. From descriptions in previous sections, it is easy to see that in MBiRW, the BiRW strategy was applied on the heterogeneous network in MBiRW, while in DR2DI, a novel Kronecker product kernel was adopted to capture the topologies of different similarity networks. Hence, through analysis, the possible reason that why MBiRW and DR2DI cannot achieve as good performance as other comparative method may be that sparse known associations will lead to weak associations between nodes in the network. Moreover, it is easy to see as well that compared with other selected methods, the MF-based method MSBMF and the NN-based method DRHGCN can achieve better predictive performances. After analysis, the possible reason may be that MSBMF used multiple kinds of biological information, and at the same time the MF-based method has strong generalization ability, which can solve the sparse problem to a certain extent. Besides, in DRHGCN, different feature embeddings were fused from different domains to avoid the loss of much network-specific information caused by undifferentiated and mixed network topology information. Simultaneously, the attention mechanism is adopted in DRHGCN to enhance the feature representation capability.

However, there is no doubt that each type of selected approach has its own advantages and disadvantages. For instance, the NN-based methods are more widely used in prediction of latent DgDsAs, while the network-based approaches are more widely adopted in detection of potential MDsAs. Besides, the MF-based method, such as NMFMDA, performs better in DgDsAs prediction but failed to obtain a satisfactory result in MDsAs prediction

Table 6. Original datasets and similarity calculation approaches adopted in selected DgDsAs related competitive methods. MBiRW: Comprehensive similarity measures and Bi-Random Walk algorithm, MSBMF: Multi-similarities bilinear matrix factorization, DR2DI: A powerful computational tool for predicting novel drug–disease associations, BNNR: Bounded nuclear norm regularization, DRHGCN: Drug Repositioning based on the Heterogeneous information fusion Graph Convolutional Network, MF: Matrix Factorization, RL: Regularization, MC: Matrix Competion, NN: Neural Network.

			Similarity	
Method	**Dataset**	**Major category**	**Drug**	**Disease**
MBiRW	DS6	Network-based	chemical structure	phenotype
MSBMF	DS6	MF	chemical structure, ATC code, side effects, drug interactions, target profiles	phenotype, ontology
DR2DI	DS7	RL	GIP, chemical structure, target domain, target go	GIP, semantic
BNNR	DS6	MC	chemical structure	phenotype
DRHGCN	DS6	NN	chemical structure	phenotype

[38]. Moreover, HMDA-Pred [52] is a network-based method that achieved better performance in experiments illustrated in the previous section. The Machine Learning-based methods are suitable for processing large amounts of data, and noisy fields such as SSI-DDI [70] perform better in the drug–drug interactions prediction field.

Disscusion

This review provides an all-sided review about the prediction of pairwise relationships between human microbes, drugs and diseases from biological data to computational models. We aim to address the connections between microbes, medicines and illness and their benefits from three perspectives.

First of all, extensive studies discussed in the section of Introduction have demonstrated that microbes, drugs and diseases are inherently linked to human health. Few efforts have investigated the microbe–drug–disease associations' network. For example, Wu *et al*. constructed a framework of various connections, including drugs, receptors, microorganisms and diseases, forming a comprehensive knowledge base and network. It can provide potential mechanisms based on quorum sensing to report causal links between drugs and microorganisms at the phenotypic level [190]. Systematic studies of the various interactions between drugs, microbes and diseases can advance understanding of personalized medicine, promote early diagnosis and develop potential therapies for a wide range of diseases.

Secondly, from Figure 1, we can easily see common microbes, drugs and diseases between the newly built DS18, DS19 and DS20. These overlapped items can be additional features to contribute to any specific prediction problem of MDsAs, DgDsAs and MDgAs. For example, Long *et al*. utilized microbe–disease associations and drug–disease associations to predict possible microbe–drug associations [121].

Thirdly, although most of the existing computational approaches and related investigations discussed in this review only focus on a specific prediction problem, these models are closely related. We firstly divided all these state-of-the-art methods of three different prediction topics into five major categories, with multiple subclasses for each primary type, according to methodology. From Table 2, similar techniques can be adapted to different prediction problems provide us intending to understand the model from the perspective of the problem and vice versa.

Challenges and prospects

Based on the existing studies on identifying possible relationships between microbes, drugs and diseases, through in-depth analysis, in this section, some challenges and opportunities were discussed in detail, based on which some valuable suggestions were provided for further improving predictive performances as well.

Integrating multi-type data for a single task

In this review, we briefly summarized 17 widely used state-of-the-art datasets and summarized 39, 48 and 6 computational methods related to prediction issues of latent MDsAs, DgDsAs and MDgAs, respectively. To achieve better predictive performance, first of all, the simplest idea is to integrate all these typical databases illustrated in Table1 together as a whole to predict any single problem since they are closely related to each other. Besides, in these 95 selected methods, many other kinds of datasets were introduced as well, for example, the widely used chemical structure-based and phenotype-based data in DgDsAs predictions [92, 95, 100, 132], the widely used symptom-based disease similarity and disease semantic similarity in MDsAs predictions [39, 50, 62, 162] and so on. Certainly, it is a challenge that will improve the prediction model's performance to integrate various kinds of bioinformatics data to target one predictive task reasonably. Moreover, some other relevant data including microbe–microbe interactions (MMIs), drug–drug interactions (DgDgIs) [163–165] and disease–disease associations (DsDsAs) [166–170] can be an addition to contribute the prediction problems of MDsAs, DgDsAs and MDgAs.

DgDgIs data

In the past few years, researchers have constructed a series of datasets related to DgDgIs. For instance, Wishart DS developed a database called DrugBank in 2018, a comprehensive and free-to-access online database that includes detailed drug data with complete drug target information. The latest release of DrugBank (version 5.1.8) contains 14 556 drug entries, including 2699 approved small molecule drugs, 1475 approved biologics, 131 nutraceuticals and over 6653 experimental (discovery-phase) drugs. Additionally, 5259 non-redundant protein sequences are linked to these drug entries, and each entry contains more than 200 data fields, with half of the information being devoted to drug/chemical data and the other half to drug target or protein data [163]. Besides, KEGG DRUG contains 501 689 interactions involving 10 979 drugs and is a comprehensive database of drug-related information, covering all drugs listed in Japan and some prescription drugs in the United States and Europe. The basic information recorded in KEGG DRUG, such as drug chemical structure, target gene and metabolism-related enzymes, can be utilized as drug features [164, 165]. Furthermore, the database called TWOSIDES was created for DgDgIs side effects collection, in which 868 221 associations between 59 220 pairs of drugs and 1301 adverse events were contained [184].

Introducing new mechanisms

Most existing 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 mechanism [171–173] also work for this problem. For example, the attention mechanism can learn the importance of different neighbouring nodes and the importance of different node (information) types to a current node. Many GNN models, such as the Spatial Convolution concept [164], 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. For example, suppose an entity (e.g. drug) is not associated with another entity in an unknown relational data (e.g. drug–disease associations) set. In that case, we cannot predict its relationships with other entities (e.g. diseases), nor can we predict the relationships between entities (e.g. microbes) outside the known relational data set. However, unsupervised approaches and enriching data may solve this kind of problem.

Benchmark evaluation

LOOCV and *K*-fold CV have been widely used in all the above-reviewed literature, have been benchmark evaluation frameworks for link predictions. Moreover, visually appealing ROC plots and AUROC provide an overview of a predictor's performance and are commonly used to assess the prediction results for the above prediction problems. The developed computational approaches for the prediction problems of MDsAs and DgDsAs and MDgAs always use the strongly imbalanced datasets in which the number of negatives outweighs the number of positives significantly. However, the ROC plots could be misleading when applied in imbalanced prediction scenarios. Alternative measures such as F-score, positive predictive value and PR plots are used less frequently [174]. Moreover, the researchers reported that PR curves could give a more informative picture of an algorithm's performance when dealing with highly skewed datasets, and algorithms that optimize the area under the ROC are not guaranteed to optimize the area under the PR curve [175]. Therefore, exploring reasonable benchmark evaluation is essential and urgent in the interaction between three terms of microbes, drugs and diseases predictions.

Handling negative samples

To our knowledge, no actual negative samples have been collected and utilized in these predictive tasks presented in this survey. The loss of negative samples significantly affects the prediction performance of the proposed model. Therefore, on the one hand, it is crucial to collect negative samples from the biomedical databases and literature. On the other hand, developing computational methods to generate high-quality negative samples is an alternative to solve this problem. Until now, few works have been conducted for improving the prediction performance by selecting high-quality negative samples. In addition, study demonstrated that selected negative samples can achieve substantial performance improvement in the domain of Protein–RNA Interactions identification [176].

Moreover, researchers have proposed technical solutions to deal with the negative samples for the proposed prediction tasks [177, 178]. For instance, Li *et al.* proposed a negative-aware training approach by introducing negative samples and training them with the original training set [179]. Plus, to balance the imbalanced bioinformatics data, Zhang *et al.* proposed a pseudo-negative sampling method based on the max-relevance and min-redundancy Pearson correlation coefficient in supervised learning [180].

Multi-type associations identification

Our analysis from biological data to computational methods shows that pairwise relationships between microbes, drugs and diseases are closely related. However, few efforts on biomedical data and computational methods have been made to simultaneously identify multi-type associations of MDsAs, DgDsAs and MDgAs which can give us new insights into how they are related.

In addition, multi-task learning (MTL), one of the computational frameworks to handle multi-type associations prediction, has been widely used in bioinformatics. For instance, Huang *et al.* proposed a collective matrix factorization-based multi-task learning method (CMFMTL) to predict two types of DgDsAs since the database CTD provides the drug–disease associations as therapeutic or marker/mechanism [116]. Wang *et al.* developed two MTL approaches, NetML and NetSML, to predict common differentially expressed genes shared across different cancer types and differentially expressed genes specific to each cancer type [181]. Zhou *et al.* introduced an MTL formulation to identify the disease progression measured by the cognitive scores and predict potential markers of the disease progression [182]. Therefore, the MTL framework has provided us with ideas and opportunities to explore pairwise associations between microbes, drugs and diseases.

Conclusions

Researchers have developed many computational methods in bioinformatics for the past few years. This work presented a comprehensive overview from biological data to computational models on predicting pairwise relationships between human microbes, drugs and diseases. Firstly, 17 state-of-the-art datasets relevant to identifying possible relationships between microbes, drugs and diseases were extracted and reviewed, and then three new datasets of MDsAs, MDgAs and DgDsAs were constructed based on these collected datasets. Secondly, this work analysed most state-of-the-art predictive methods related to these three predictive tasks in detail according to their core strategies. Meanwhile, 7 and 5 representative methods that belong to two predictive topics of MDsAs and DgDsAs separately were selected for comparison to analyse the effects of original datasets and core strategies on the predictive performances. Finally, this work provided suggestions for further improving predictive performances and outlined some challenges and opportunities.

Key Points

- Prediction of pairwise associations between microbes, drugs and diseases provides essential insights into the underlying understanding of disease mechanisms from the perspective of human microbes and drugs, which are greatly helpful for investigating pathogenesis, promoting early diagnosis and improving precision medicine. A total of 17 state-of-the-art datasets and 93 computational methods for predicting potential microbe–disease associations (MDsAs), drug–disease associations (DgDsAs) and microbe–drug associations (MDgAs) are reviewed, respectively.
- Computational approaches for MDsAs, DgDsAs and MDgAs predictions based on diverse strategies and core ideas, to our knowledge, are classified and discussed. A total of 12 state-of-the-art methods are performed, evaluated and analysed in terms of raw data sets, additional similarity features and computational approaches, and possible performance improvements are discussed based on the analysis result.
- Based on the development of identifying possible relationships between microbes, drugs and diseases, current computational challenges, opportunities and prospects are presented and valuable suggestions for improving predictive performances are discussed.

Author contributions statement

Y.Q.T. and X.Y.Y. collected data and conducted the experiments, L.W., P.Y.P. ,Y.Q.T. and X.Y.Y. wrote this paper. L.W. and L.A.K. provided suggestions and reviewed the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding

The National Natural Science Foundation of China (No.61873221) and the Natural Science Foundation of Hunan Province (No.2019JJ70010).

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**(7420):207–14.
2. Ron S, Shai F, Ron M. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* 2016;**14**(8):e1002533.
3. Ventura M, O'Flaherty S, Claesson MJ, *et al.* Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat Rev Microbiol* 2009;**7**:61–71.
4. Kau AL, Ahern PP, Griffin NW, *et al.* Human nutrition, the gut microbiome and the immune system. *Nature* 2011;**474**:327–36.
5. Sommer F, Bäckhed F. The gut microbiota - masters of host development and physiology. *Nat Rev Microbiol* 2013;**11**:227–38.
6. Ley RE, Turnbaugh PJ, Klein S, *et al.* Human gut microbes associated with obesity. *Nature* 2006;**444**:1022–3.
7. Durack J, Lynch SV. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med* 2019;**216**(1):20–40.
8. Xiang Y-T, Li W, Zhang Q, *et al.* Timely research papers about COVID-19 in China. *Lancet* 2020;**395**(10225):684–5.
9. McCoubrey LE, Gaisford S, Orlu M, *et al.* Predicting drug-microbiome interactions with machine learning. *Biotechnol Adv* 2022;**54**:107797.
10. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, *et al.* Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 2019;**570**(7762):462–7.
11. Cummings J, Lee G, Ritter A, *et al.* Alzheimer's disease drug development pipeline: 2018. *Alzheimers Dement (N Y)* 2018;**4**:195–214.
12. Adams CP, Brantner VV. Estimating The Cost Of New Drug Development: Is It Really $802 Million? *Health Aff* 2006;**25**:420–8.
13. Ma W, Zhang L, Zeng P, *et al.* An analysis of human microbe-disease associations. *Brief Bioinform* 2017;**18**(1):85–97.

14. Janssens Y, Nielandt J, Bronselaer A, *et al.* Disbiome database: linking the microbiome to disease. *BMC Microbiol* 2018;**18**(1):50.
15. Yao G, Zhang W, Yang M, *et al.* MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes. *Genomics Proteomics Bioinformatics* 2020;**18**(6):760–72.
16. Wu C, Xiao X, Yang C, *et al.* Mining microbe-disease interactions from literature via a transfer learning model. *BMC Bioinformatics* 2021;**22**(1):432.
17. Skoufos G, Kardaras FS, Alexiou A, *et al.* Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res* 2021;**49**:D1328–33.
18. Davis AP, Grondin CJ, Johnson RJ, *et al.* The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res* 2019;**47**:D948–54.
19. Davis AP, Grondin CJ, Johnson RJ, *et al.* The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2017;**45**:D972–8.
20. Davis AP, Murphy CG, Johnson R, *et al.* The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res* 2013;**41**:D1104–14.
21. Davis AP, King BL, Mockus S, *et al.* The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res* 2011;**39**:D1067–72.
22. Mattingly CJ, Rosenstein MC, Colby GT, *et al.* The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J Exp Zool A Comp Exp Biol* 2006;**305**(9):689–92.
23. Mattingly CJ, Colby GT, Forrest JN, *et al.* The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect* 2003;**111**(6):793–5.
24. Davis AP, Grondin CJ, Johnson RJ, *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res* 2021;**49**:D1138–43.
25. Wishart DS, Craig K, Guo AC, *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**:D901–6.
26. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res* 2002;**30**:412–5.
27. Hamosh A, Scott AF, Amberger J, *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002;**30**(1):52–5.
28. de Leon J. Highlights of drug package inserts and the website DailyMed: the need for further improvement in package inserts to help busy prescribers. *J Clin Psychopharmacol* 2011;**31**(3):263–5.
29. Kim S, Thiessen PA, Bolton EE, *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res* 2016;**44**:D1202–13.
30. Sun Y-Z, Zhang D-H, Cai S-B, *et al.* MDAD: A Special Resource for Microbe-Drug Associations. *Front Cell Infect Microbiol* 2018;**8**:424.
31. Rajput A, Thakur A, Sharma S, *et al.* aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Res* 2018;**46**:D894–900.
32. Andersen PI, Ianevski A, Lysvand H, *et al.* Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseasesm* 2020;**93**:268–76.
33. Wen Z, Yan C, Duan G, *et al.* A survey on predicting microbe-disease associations: biological data and computational methods. *Brief Bioinform* 2021;**22**(3):bbaa157.
34. Long Y, Luo J. WMGHMDA: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. *BMC Bioinformatics* 2019;**20**:541.
35. Zhang W, Yang W, Lu X, *et al.* The Bi-Direction Similarity Integration Method for Predicting Microbe-Disease Associations. *IEEE Access* 2018;**6**:38052–61.
36. Niu Y-W, Qu C-Q, Wang G-H, *et al.* RWHMDA: Random Walk on Hypergraph for Microbe-Disease Association Prediction. *Front Microbiol* 2019;**10**:1578.
37. Wu C, Gao R, Zhang D, *et al.* PRWHMDA: Human Microbe-Disease Association Prediction by Random Walk on the Heterogeneous Network with PSO. *Int J Biol Sci* 2018;**14**(8):849–57.
38. 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.
39. Fan C, Lei X, Guo L, *et al.* Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. *Neurocomputing* 2019;**323**:76–85.
40. 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.
41. Chen S, Liu D, Zheng J, *et al.* Predicting Microbe-Disease Association by Kernelized Bayesian Matrix Factorization. *Intelligent Computing Theories and Application* 2018;**10955**:389–94.
42. Long Y, Luo J, Zhang Y, *et al.* Predicting human microbe-disease associations via graph attention networks with inductive matrix completion. *Brief Bioinform* 2021;**22**:bbaa146.
43. Shen X, Chen Y, Jiang X, *et al.* Predicting disease-microbe association by random walking on the heterogeneous network. In: Tianhai Tian, Qinghua Jian, Yunlong Liu, Kevin Burrage, Jiangning Song, Yadong Wang, Xiaohua Hu, Shinichi Morishita, Qian Zhu and Guohua Wang (eds), *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China: IEEE, 2016, 771–4.
44. Huang Z-A, Chen X, Zhu Z, *et al.* PBHMDA: Path-Based Human Microbe–Disease Association Prediction. *Front Microbiol* 2017;**8**:233.
45. 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 Bioinform* 2018;**17**(4):1341–51.
46. Ma Y, Jiang H. NinimHMDA: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human Microbe-Disease association. *Bioinformatics* 2021;**36**:5665–71.
47. Yin M-M, Liu J-X, Gao Y-L, *et al.* NCPLP: A Novel Approach for Predicting Microbe-Associated Diseases With Network Consistency Projection and Label Propagation. *IEEE Transactions on Cybernetics* 2020. Early Access, doi: 10.1109/TCYB.2020.3026652.
48. 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.
49. Xu D, Xu H, Zhang Y, *et al.* MDAKRLS: Predicting human microbe-disease association based on Kronecker regularized least squares and similarities. *J Transl Med* 2021;**19**(1):1–12.
50. Yan C, Duan G, Wu F-X, *et al.* MCHMDA: Predicting Microbe-Disease Associations Based on Similarities and Low-Rank Matrix Completion. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:611–20.

51. Wang F, Huang Z-A, Chen X, *et al.* LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe-Disease Association prediction. *Sci Rep* 2017;**7**:7601.
52. Fan Y, Chen M, Zhu Q, *et al.* Inferring Disease-Associated Microbes Based on Multi-Data Integration and Network Consistency Projection. *Front Bioeng Biotechnol* 2020;**8**:831.
53. 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.
54. 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.
55. Peng L-H, Yin J, Zhou L, *et al.* Human Microbe-Disease Association Prediction Based on Adaptive Boosting. *Front Microbiol* 2018;**9**:2440.
56. Liu Y, Wang S-L, Zhang J-F, *et al.* DMFMDA: Prediction of Microbe-Disease Associations Based on Deep Matrix Factorization Using Bayesian Personalized Ranking. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:1763–72.
57. Shen Z, Jiang Z, Bao W. CMFHMDA: Collaborative Matrix Factorization for Human Microbe-Disease Association Prediction. In: De-Shuang Huang, Kang-Hyun Jo and Juan Carlos Figueroa-Garcia (eds), *International Conference on Intelligent Computing*, Liverpool, United Kingdom: Springer, 2017, 261–9.
58. Yan C, Duan G, Wu F, *et al.* BRWMDA: Predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**17**(5):1595–604.
59. Shi J-Y, Huang H, Zhang Y-N, *et al.* BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinformatics* 2018;**19**:281.
60. Li H, Wang Y, Jiang J, *et al.* A Novel Human Microbe-Disease Association Prediction Method Based on the Bidirectional Weighted Network. *Front Microbiol* 2019;**10**:676.
61. 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**:e0184394.
62. Chen X, Huang Y-A, You Z-H, *et al.* A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 2017;**33**:733–9.
63. Shen X, Zhu H, Jiang X, *et al.* A Novel Approach Based on Bi-Random Walk to Predict Microbe-Disease Associations. In: De-Shuang Huang, M. Michael Gromiha, Kyungsook Han and Abir Hussain (eds), *International Conference on Intelligent Computing*, Wuhan, China: Springer, 2018; pp:746–52.
64. Wang L, Wang Y, Li H, *et al.* A Bidirectional Label Propagation Based Computational Model for Potential Microbe-Disease Association Prediction. *Front Microbiol* 2019;**10**:684.
65. Wang Y, Lei X, Lu C, *et al.* Predicting Microbe-disease Association Based on Multiple Similarities and LINE Algorithm. *IEEE/ACM Trans Comput Biol Bioinform* 2021. Early Access, doi: 10.1109/TCBB.2021.3082183.
66. Fu C, Zhong R, Jiang X, *et al.* An Integrated Knowledge Graph for Microbe-Disease Associations. *In: Health Information Science(HIS)*, Amsterdam and Leiden, Netherlands: Springer, 2020;77–90.
67. Huang Y-A, You Z-H, Chen X, *et al.* Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J Transl Med* 2017;**15**:209.
68. Wang D, Cui Y, Cao Y, *et al.* Human Microbe-Disease Association Prediction by a Novel Double-Ended Random Walk with Restart. *Biomed Res Int* 2020;**2020**:3978702.
69. Peng L, Shen L, Liao L, *et al.* 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.
70. Li H, Wang Y, z. z. zhen, et al. Identifying Microbe-Disease Association Based on a Novel Back-Propagation Neural Network Model. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**(6): 2502–13.
71. Dayun L, Junyi L, Yi L, *et al.* MGATMDA: Predicting microbe-disease associations via multi-component graph attention network. *IEEE/ACM Trans Comput Biol Bioinform* 2021. Early Access, doi: 10.1109/TCBB.2021.3116318.
72. Gottlieb A, Stein GY, Ruppin E, *et al.* PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496.
73. Moghadam H, Rahgozar M, Gharaghani S. Scoring multiple features to predict drug disease associations using information fusion and aggregation. *SAR QSAR Environ Res* 2016;**27**: 609–28.
74. Jiang HJ, Huang YA, You ZH. Predicting Drug-Disease Associations via Using Gaussian Interaction Profile and Kernel-Based Autoencoder. *Biomed Res Int* 2019;**2019**:1–11.
75. Wang J, Wang W, Yan C, *et al.* Predicting Drug-Disease Association Based on Ensemble Strategy. *Front Genet* 2021;**12**:666575.
76. Zhang W, Yue X, Lin W, *et al.* Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 2018;**19**:233.
77. Zhang W, Yue X, Huang F, *et al.* Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* 2018;**145**:51–9.
78. Wang W, Yang S, Zhang X, *et al.* Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;**30**(20):2923–30.
79. Yang K, Zhao X, Waxman D, *et al.* Predicting drug-disease associations with heterogeneous network embedding. *Chaos* 2019;**29**:123109.
80. Luo H, Wang J, Li M, *et al.* Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;**32**:2664–71.
81. Luo H, Li M, Wang S, *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12.
82. Wu G, Liu J, Yue X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinformatics* 2019;**20**:134.
83. Wu G, Liu J, Wang C. Predicting drug-disease interactions by semi-supervised graph cut algorithm and three-layer data integration. *BMC Med Genomics* 2017;**10**:79.
84. Jiang HJ, Huang YA, You ZH. SAEROF: an ensemble approach for large-scale drug-disease association prediction by incorporating rotation forest and sparse autoencoder deep neural network. *Sci Rep* 2020;**10**:4972.
85. Zhou R, Lu Z, Luo H, *et al.* NEDD: a network embedding based method for predicting drug-disease associations. *BMC Bioinformatics* 2020;**21**:387.
86. Jiang HJ, You ZH, Huang YA. Predicting drugdisease associations via sigmoid kernel-based convolutional neural networks. *J Transl Med* 2019;**17**(1):382.
87. Yang M, Luo H, Li Y, *et al.* Overlap matrix completion for predicting drug-associated indications. *PLoS Comput Biol* 2019;**15**:e1007541.
88. Liu H, Zhang W, Song Y, *et al.* HNet-DNN: inferring new drug-disease associations with deep neural network based on heterogeneous network features. *J Chem Inf Model* 2020;**60**(4): 2367–76.

89. Liu J, Zuo Z, Wu G. Link Prediction Only With Interaction Data and its Application on Drug Repositioning. *IEEE Trans Nanobioscience* 2020;**19**:547–55.
90. Yang M, Huang L, Xu Y, *et al.* Heterogeneous graph inference with matrix completion for computational drug repositioning. *Bioinformatics* 2021;**36**:5456–64.
91. Yang M, Luo H, Li Y, *et al.* Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 2019;**35**: i455–63.
92. Yang M, Wu G, Zhao Q, *et al.* Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief Bioinform* 2021;**22**:1–14.
93. Yan C-K, Wang W-X, Zhang G, *et al.* BiRWDDA: A Novel Drug Repositioning Method Based on Multisimilarity Fusion. *J Comput Biol* 2019;**26**(11):1230–42.
94. Tian Z, Teng Z, Cheng S, *et al.* Computational drug repositioning using meta-path-based semantic network analysis. *BMC Syst Biol* 2018;**12**:134.
95. Cai L, Lu C, Xu J, *et al.* Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform* 2021;**22**(6):bbab319.
96. Dai W, Liu X, Gao Y, *et al.* Matrix Factorization-Based Prediction of Novel Drug Indications by Integrating Genomic Space. *Comput Math Methods Med* 2015;**2015**:275045.
97. Liu H, Song Y, Guan J, *et al.* Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* 2016;**17**:539.
98. Le D-H, Nguyen-Ngoc D. Repositioning by Integrating Known Disease-Gene and Drug-Target Associations in a Semi-supervised Learning Model. *Acta Biotheor* 2018;**66**(4): 315–31.
99. Wang Y, Chen S, Deng N, *et al.* Drug Repositioning by Kernel-Based Integration of Molecular Structure, Molecular Activity, and Phenotype Data. *PLoS ONE* 2013;**8**(11):e78518.
100. 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.
101. Olivier B. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**(Database issue):D267–70.
102. Martínez V, Navarro C, Cano C, *et al.* DrugNet: Network-based drug-disease prioritization by integrating heterogeneous data. *Artif Intell Med* 2015;**63**(1):41–9.
103. Schriml LM, Arze C, Nadendla S, *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2012;**40**(D1):D940–6.
104. Liang X, Zhang P, Yan L, *et al.* LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. *Bioinformatics* 2017;**33**(8):1187–96.
105. Xuan P, Song Y, Zhang T, *et al.* Prediction of Potential Drug-Disease Associations through Deep Integration of Diversity and Projections of Various Drug Features. *Int J Mol Sci* 2019;**20**(17):4102.
106. Xuan P, Cui H, Shen T, *et al.* HeteroDualNet: A Dual Convolutional Neural Network With Heterogeneous Layers for Drug-Disease Association Prediction via Chou's Five-Step Rule. *Front Pharmacol* 2019;**10**:1301.
107. Xuan P, Zhao L, Zhang T, *et al.* Inferring Drug-Related Diseases Based on Convolutional Neural Network and Gated Recurrent Unit. *Molecules* 2019;**24**(15):2712.
108. Yangkun PX, *et al.* Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* 2019;**35**(20):4108–19.
109. Wang X, Yan R. DDAPRED: a computational method for predicting drug repositioning using regularized logistic matrix factorization. *J Mol Model* 2020;**26**(3):60.
110. Song Y, Cui H, Zhang T, *et al.* Prediction of drug-related diseases through integrating pairwise attributes and neighbor topological structures. *IEEE/ACM Trans Comput Biol Bioinform* 2021; Early Access, doi: 10.1109/TCBB.2021.3089692.
111. Xuan P, Gao L, Sheng N, *et al.* Graph Convolutional Autoencoder and Fully-Connected Autoencoder with Attention Mechanism Based Method for Predicting Drug-Disease Associations. *IEEE J Biomed Health Inform* 2021;**25**(5):1793–804.
112. Zeng X, Zhu S, Liu X, *et al.* deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**(24):5191–8.
113. Brown AS, Patel CJ. A standard database for drug repositioning. *Scientific Data* 2017;**4**(1):170029.
114. Jin S, Niu Z, Jiang C, *et al.* HeTDR: Drug repositioning based on heterogeneous networks and text mining. *Patterns* 2021;**2**(8):100307.
115. Davis AP, *et al.* The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 2016;**45**(D1):D972–8.
116. Huang F, Qiu Y, Li Q, *et al.* Predicting Drug-Disease Associations via Multi-Task Learning Based on Collective Matrix Factorization. *Front Bioeng Biotechnol* 2020;**8**:218.
117. Yu Z, Huang F, Zhao X, *et al.* Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021;**22**(4):bbaa243.
118. Zhao BW, You ZH, Wong L, *et al.* MGRL: Predicting Drug-Disease Associations Based on Multi-Graph Representation Learning. *Front Genet* 2021;**12**:657182.
119. Li Z, Huang Q, Chen X, *et al.* Identification of Drug-Disease Associations Using Information of Molecular Structures and Clinical Symptoms via Deep Convolutional Neural Network. *Front Chem* 2020;**7**:924.
120. Deng L, Huang Y, Liu X, *et al.* Graph2MDA: a multi-modal variational graph embedding model for predicting microbe-drug associations. *Bioinformatics* 2021;**38**(4):1118–25.
121. Long Y, Wu M, Liu Y, *et al.* Ensembling graph attention networks for human microbe-drug association prediction. *Bioinformatics* 2020;**36**(2):i779–86.
122. Long Y, Wu M, Kwoh CK, *et al.* Predicting human microbe-drug associations via graph convolutional network with conditional random field. *Bioinformatics* 2020;**36**(19):4918–27.
123. Zhu L, Duan G, Yan C, *et al.* Prediction of Microbe-Drug Associations Based on KATZ Measure. In: Illhoi Yoo, Jinbo Bi and Xiaohua Hu (eds), *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA: IEEE, 2019, 183–7.
124. Long Y, Luo J. Association Mining to Identify Microbe Drug Interactions Based on Heterogeneous Network Embedding Representation. *IEEE J Biomed Health Inform* 2021;**25**(1):266–75.
125. Zhu L, Wang J, Li G, *et al.* Predicting Microbe-Drug Association based on Similarity and Semi-Supervised Learning. *American Journal of Biochemistry and Biotechnology* 2021;**17**(1):50–8.
126. Hamosh A, Scott AF, Amberger JS, *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**(Database issue):D514–7.
127. J. Li, Z. Lu. A new method for computational drug repositioning using drug pairwise similarity. In Lyle Ungar and Cathy Wu (eds), *IEEE International Conference on Bioinformatics and Biomedicine(BIBM)*, Pennsylvania, USA: IEEE, 2012; pp:1–4.
128. Fan C, Lei X, Guo L, et al. Predicting the associations between microbes and diseases by integrating multiple data sources

and path-based HeteSim scores. *Neurocomputing* 2019; **323**: 76–85.

129. Chen H, Zhang H, Zhang Z, *et al.* Network-Based Inference Methods for Drug Repositioning. *Comput Math Methods Med* 2015;**2015**:130620.

130. Huang YF, Yeh HY, Soo VW. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med Genomics* 2013;**6**(3):S4.

131. Lotfi Shahreza M, Ghadiri N, Mousavi SR, *et al.* Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning. *J Biomed Inform* 2017;**68**:167–83.

132. Chen H, Zhang Z, Peng W. miRDDCR: a miRNA-based method to comprehensively infer drug-disease causal relationships. *Sci Rep* 2017;**7**(1):15921.

133. Wang Y, Guo M, Ren Y, *et al.* Drug repositioning based on individual bi-random walks on a heterogeneous network. *BMC Bioinformatics* 2019;**20**(15):547.

134. Oh M, Ahn J, Yoon Y. A Network-Based Classification Model for Deriving Novel Drug-Disease Associations and Assessing Their Molecular Actions. *Plos One* 2014;**9**(10):e111668.

135. Lee T, Yoon Y. Drug repositioning using drug-disease vectors based on an integrated network. *BMC Bioinformatics* 2018;**19**(1):446.

136. Zhang W, Lu X, Yang W, *et al.* HNGRNMF: Heterogeneous Network-based Graph Regularized Nonnegative Matrix Factorization for predicting events of microbe-disease associations. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, 803–7.

137. Ma Y, Liu G, Ma Y, *et al.* Integrative Analysis for Identifying Co-Modules of Microbe-Disease Data by Matrix Tri-Factorization With Phylogenetic Information. *Front Genet* 2020;**11**:83.

138. Yang J, Li Z, Fan X, *et al.* Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization. *J Chem Inf Model* 2014;**54**(9):2562–9.

139. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. *AMIA Annual Symposium proceedings* 2014;**2014**:1258–67.

140. Yang M, Huang L, Xu Y, et al. Heterogeneous graph inference with matrix completion for computational drug repositioning. *Bioinformatics* 2020; **36**(22-23): 5456–64.

141. Xuan P, Ye Y, Zhang T, *et al.* Convolutional Neural Network and Bidirectional Long Short-Term Memory-Based Method for Predicting Drug-Disease Associations. *Cell* 2019;**8**(7):705.

142. Tnj A, Jgr A, Raab C. SNF-CVAE: Computational method to predict drug-disease interactions using similarity network fusion and collective variational autoencoder. *Knowledge-Based Systems* 2020;**212**:106585.

143. Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953;**18**(1):39–43.

144. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. In: Association for Computing Machinery, San Francisco California USA, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen and Rajeev Rastogi (eds), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 855–64.

145. Fu T, Lee W-C, Lei Z. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, 1797–806.

146. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, 135–44.

147. Gu C, Liao B, Li X, *et al.* Network Consistency Projection for Human miRNA-Disease Associations Inference. *Sci Rep* 2016;**6**(1):36054.

148. Shi C, Kong X, Huang Y, *et al.* HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. *IEEE Transactions on Knowledge and Data Engineering* 2014;**26**(10):2479–92.

149. Johnson CC. Logistic Matrix Factorization for Implicit Feedback Data. *Advances in Neural Information Processing Systems27* 2014.

150. Liu Y, Wu M, Miao C, *et al.* Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLoS Comput Biol* 2016;**12**(2):e1004760.

151. Gönen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;**28**(18):2304–10.

152. Xu D, Xu H, Zhang Y, et al. MDAKRLS: Predicting human microbe-disease association based on Kronecker regularized least squares and similarities. *J Transl Med* 2021; **19**(1): 66.

153. Cai J-F, Candés EJ, Shen Z. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization* 2010;**20**(4):1956–82.

154. Joachims T. Making large-scale SVM learning practical. *MIT Press* 1998;169–84.

155. Oyama S, Manning CD. Using Feature Conjunctions across Examples for Learning Pairwise Classifiers. *Transactions of the Japanese Society for Artificial Intelligence* 2005;**20**(2): 105–116.

156. Shawe-Taylor J. Kernel Methods for Pattern. *Analysis*. Cambridge: Cambridge University Press, 2014;**140**.

157. Huang YA, Hu P, Chan K, *et al.* Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics* 2019;**36**(3):851–8.

158. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: Yoshua Bengio and Yann LeCun (eds), *the 5th International Conference on Learning Representations (ICLR), OpenReview*, Toulon, France, 2017.

159. Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, 2017, 282–9.

160. Im DJ, Ahn S, Memisevic R, *et al.* Auto-Encoding Variational Bayes. *CoRR* 2014; abs/1312.6114.

161. Chen Y, Rijke MD. A Collective Variational Autoencoder for Top-N Recommendation with Side Information. In: *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems (DLRS 2018)*, 2018, 3–9.

162. Lowe H, Barnett G. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 1994;**271**(14):1103–8.

163. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.

164. Kanehisa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(Database issue):D354–7.

165. Kanehisa M, Araki M, Goto S, *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2007;**36**(Database issue):D480–4.

166. Oti M, Brunner H. The modular nature of genetic diseases. *Clin Genet* 2007;**71**(1):1–11.

167. Goh K-I, Cusick ME, Valle D, *et al.* The human disease network. *Proc Natl Acad Sci* 2007;**104**(21):8685–90.

168. Oti M, Huynen MA, Brunner HG. Phenome connections. *Trends Genet* 2008;**24**(3):103–6.
169. van Driel MA, Bruggeman J, Vriend G, *et al.* A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**(5):535–42.
170. Zhou X, Menche J, Barabási A-L, *et al.* Human symptoms-disease network. *Nature. Communications* 2014;**5**(1):4212.
171. Jing Y, Yang Y, Wang X, *et al. Amalgamating Knowledge From Heterogeneous Graph Neural Networks* 2021;15709–18.
172. Wang X, Ji H, Shi C, *et al.* Heterogeneous Graph Attention Network. In: Association for Computing Machinery, San Francisco CA USA, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates and Leila Zia (eds), *The World Wide Web Conference*, 2019;2022–32.
173. Li X, Shang Y, Cao Y, *et al.* Type-Aware Anchor Link Prediction across Heterogeneous Networks Based on Graph Attention Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 2020;**34**:147–55.
174. 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.
175. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning*, 2006, 233–40.
176. Cheng Z, Huang K, Wang Y, *et al.* Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst Biol* 2017;**11**(2):9.
177. Peng L, Shen L, Liao L, et al. RNMFMDA: A Microbe-Disease Association Identification Method Based on Reliable Negative Sample Selection and Logistic Matrix Factorization With Neighborhood Regularization. *Front Microbiol* 2020; **11**:592430–0.
178. Shi J-Y, Huang H, Zhang Y-N, et al. BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinformatics* 2018; **19**(9): 281.
179. Li X, Jia X, Jing X-Y. Negative-Aware Training: Be Aware of Negative Samples. *Santiago de Compostela* 2020.
180. Zhang Y, Qiao S, Lu R, *et al.* How to balance the bioinformatics data: pseudo-negative sampling. *BMC Bioinformatics* 2019;**20**(25):695.
181. Wang Z, He Z, Shah M, *et al.* Network-based multi-task learning models for biomarker selection and cancer outcome prediction. *Bioinformatics* 2020;**36**(6):1814–22.
182. Zhou J, Yuan L, Liu J, *et al.* A multi-task learning formulation for predicting disease progression. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, 814–22.
183. Zhao Y, Wang C-C, Chen X. Microbes and complex diseases: From experimental results to computational models. *Brief Bioinform* 2020;**22**(3):bbaa158.
184. Tatonetti NP, Ye PP, Daneshjou R, *et al.* Data-Driven Prediction of Drug Effects and Interactions. *Sci Transl Med* 2012;**4**(125):125ra31.
185. Vieira-Silva S, Falony G, Belda E, *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* 2020;**581**(7808):310–5.
186. Javdan B, Lopez JG, Chankhamjon P, *et al.* Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. *Cell* 2020;**181**(7):1661–1679.e22.
187. Hassan R, Allali I, Agamah FE, *et al.* Drug response in association with pharmacogenomics and pharmacomicrobiomics:towards a better personalized medicine. *Brief Bioinform* 2021;**22**(4):bbaa292.
188. Maier L, Pruteanu M, Kuhn M, *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* 2018;**555**(7698):623–8.
189. Panebianco C, Andriulli A, Pazienza V. Pharmacomicrobiomics: exploiting the drug-microbiota interactions in anticancer therapies. *Microbiome* 2018;**6**(1):92.
190. Wu S, Yang S, Wang M, *et al.* Quorum Sensing-Based Interaction Network Construction for Drugs, Microbes, and Diseases. *Research Square* 2021. https://doi.org/10.21203/rs.3.rs-845581/v1.