Supplementary Materials for

"Ensembling graph attention networks for human microbe-drug association prediction"

1. Construction of biological networks

1.1 Microbe-disease associations

We assembled microbe-disease associations from databases HMDAD (*http://www.cuilab.cn/hmdad/*) (*Ma et al. 2017*) and Disbiome (*https://disbiome.ugent.be/home*) (*Janssens et al. 2018*). Here, we derived associations including microbes that are contained in MDAD (*Sun et al. 2018*). As a result, we downloaded 59 microbe-disease associations between 29 microbes and 20 diseases.

1.2 Drug-disease associations

In this work, drug-disease associations were downloaded from CTD database (*http://ctdbase.org/*) (*Davis et al. 2019*). We only selected entries including drugs in MDAD. Finally, we obtained 168 drug-disease associations between 82 drugs and 21 diseases.

1.3 Drug-drug interactions

We collected drug-drug interactions in updated Drugbank (*https://www.drugbank.ca/releases/latest*) (*Wishart et al. 2018*). After intersecting the names of drugs in MDAD, in total, 5586 drug-drug interactions were downloaded involving 181 drugs.

1.4 Microbe-microbe interactions

We attained microbe-microbe interaction data from MIND (*http://www.microbialnet.org/mind_home.html*) . Also, only these interactions including microbes in MDAD were downloaded. After screening, we acquired 138 microbe-microbe interactions including 82 microbes.

1.5 Disease semantic similarity

For calculate the semantic similarity for diseases, we first derived Directed Acyclic Graphs (DAGs) describing diseases from Mesh (Medical Subject Headings) database (*Lipscomb, C.E. 2000*), where there are a large number of descriptors of diseases included. In the DAG of a disease D , nodes represent its ancestor as well as D itself while the edges mean the direct ones from patient nodes to child nodes. Based on the DAG, we can define the contribution value of disease d in DAG(D) to the semantic value of disease D as follows:

$$SV_D(d) = \begin{cases} 1, & if\ d = D, \\ \max\{\Delta * SV_D(d') \mid d' \in children\ of\ d\}, & if\ d \neq D, \end{cases}$$

where Δ is the semantic contribution decay factor (It is often set to 0.5). The semantic value of disease D is defined as follows:

$$SV(D) = \sum_{d \in T(D)} SV_D(d),$$

where $T(D)$ is the set of ancestor diseases of disease D and D itself. Motivated by the assumption that two diseases with larger shared part of their DAGs have greater similar score, we formulate the semantic similarity value between disease and disease as follows:

$$DS(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} (SV_{d_i}(t) + SV_{d_j}(t))}{SV(d_i) + SV(d_j)}.$$

1.6 Drug structure similarity

Following the method SIMCOMP2 (*Hattori et al. 2010*) , we calculated drug structure similarity $SD \in \mathbb{R}^{nd \times nd}$. After screening, 306 out of 1373 drugs have structure similarities with each other. Eventually, we derived structural similarity scores for 2313 drug-drug pairs with the default threshold 0.5.

1.7 Gaussian kernel drug similarity

Motivated by that assumption that drugs with similar treatment functions closely interact with similar microbes, we calculated Gaussian kernel similarity for drugs by leveraging their Gaussian kernel interaction profiles based on known microbe-drug associations. Since the i^{th} row of adjacent matrix A represents the interaction between drug d_i and all drugs, Let us denote the interaction profiles of drug d_i as $IP(d_i)$, respectively. The Gaussian kernel similarity between drug d_i and drug d_j is formulated as follows:

$$GS(d_i, d_j) = \exp(-\lambda \left\| IP(d_i) - IP(d_j) \right\|^2 ,$$

where λ means the normalized kernel bandwidth, and is defined as follows:

$$\lambda = \lambda^{'} / (\frac{1}{nd} \sum_{i=1}^{nd} \left\| IP(d_i) \right\|^2),$$

where $\lambda^{'}$ is the original bandwidth and generally is set to 1.

1.8 Integrated drug similarity

Since not all drugs have structure similarity with other drugs, it fails to obtain input feature for such drugs. For complementing and improving the similarity for drugs, we defined the final drug similarity $DS \in \mathbb{R}^{1373 \times 1373}$ as the combination of drug structure similarity SS and Gaussian kernel drug similarity GS . More specifically, if a drug d_i and a drug d_j have structure similarity, then the integrated similarity between d_i and d_j was calculated as the average value of $GS(d_i, d_j)$ and $SS(d_i, d_j)$. Otherwise it was equal to the Gaussian kernel similarity. The new drug similarity was calculated as follows:

$$DS(d_i, d_j) = \begin{cases} \frac{GS(d_i, d_j) + SS(d_i, d_j)}{2}, & if\ SS(d_i, d_j) \neq 0, \\ GS(d_i, d_j), & if\ SS(d_i, d_j) = 0. \end{cases}$$

2. Experiment results

2.1 Comparison performance between our model with seven state-of-the-art methods

To evaluate the effectiveness of our model, we compare our proposed model of EGATMDA with seven state-of-the-art methods on dataset MDAD under the setting CVS1. Table S1 shows the results of 2-fold CV, 5-fold CV and 10-fold CV, respectively. It indicates that our model consistently outperforms all baseline methods in terms of AUC and AUPR while the AUPR value is lower than that of HMDAKATZ in 10-fold CV. Therefore, we can conclude that our model is effective and promising in inferring potential microbe-drug associations.

Table S1. Comparison performance between our method and state-of-the-art methods under the setting CVS1. The best results are marked in bold and the second best is underlined.

Methods	2-fold CV		5-fold CV		10-fold CV	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
HMDAKATZ	0.9166±0.0070	0.9054±0.0050	0.9365±0.0073	0.9305±0.0064	<u>0.9453±0.0101</u>	**0.9405±0.0103**
IMCMDA	0.7236±0.0090	0.7976±0.0052	0.7334±0.0185	0.8038±0.0215	0.7420±0.0143	0.8039±0.0213
NTSHMDA	0.8779±0.0077	0.8983±0.0097	0.8993±0.0137	0.8965±0.0149	0.9092±0.0071	0.9045±0.0080
GCMDR	0.8917±0.0009	0.8940±0.0052	0.8938±0.0137	0.8956±0.0142	0.8942±0.0142	0.8979±0.0144
NetLapRLS	<u>0.9274±0.0052</u>	<u>0.9277±0.0055</u>	<u>0.9372±0.0078</u>	0.9381±0.0085	0.9369±0.0090	0.9379±0.0097
BLM-NII	0.8822±0.0495	0.9051±0.0318	0.9136±0.0484	<u>0.9394±0.0299</u>	0.9071±0.0835	0.9132±0.0537
WNN-GIP	0.7741±0.0413	0.8485±0.0352	0.7799±0.0677	0.8587±0.0456	0.8211±0.0681	0.8869±0.0452
EGATMDA	**0.9508±0.0083**	**0.9315±0.0047**	**0.9586±0.0083**	**0.9460±0.0112**	**0.9601±0.0092**	<u>0.9385±0.0161</u>

2.2 The impact of different biological data on the performance

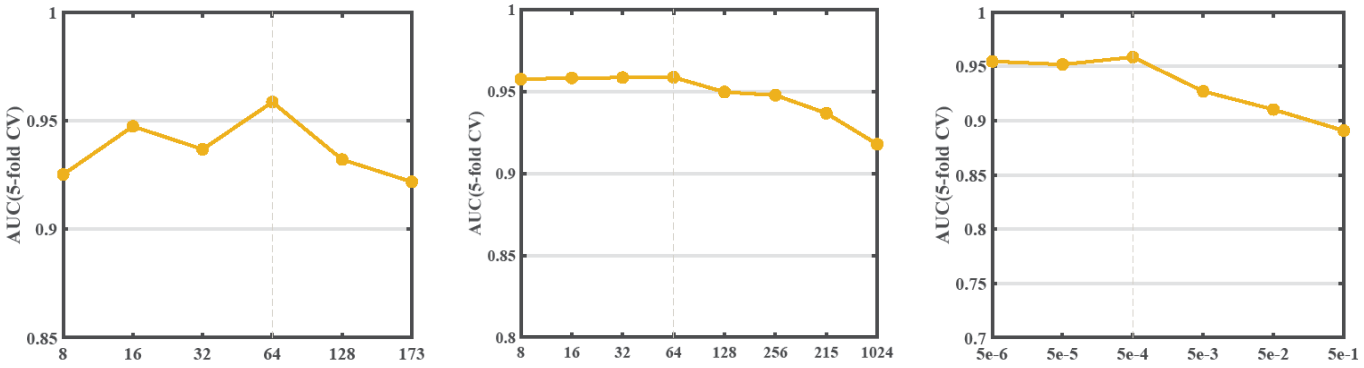
In this work, we fully exploit multiply types of biological networks to predict novel microbe-drug associations, including microbe-drug bipartite network (Net_1), microbe-drug heterogeneous network (Net_2) and microbe-disease-drug heterogeneous network (Net_3). To test the impacts of different networks on the model, we evaluate our model with different combination of networks as inputs using 5-fold CV under the settings CVS1, CVS2 and CVS3, respectively. It should point out that *Global Net* is a global network that is constructed by integrating Net_1, Net_2 with Net_3. The results have been shown in Table S2, from which it could be found that our model achieves the best performance when three networks are simultaneously fed into the model under three different kinds of scenarios, indicating that all of these three biological data are useful for improving the prediction accuracy of our model. Besides, we can conclude that network Net_1 plays the most important role among three networks. And Net_2 contributes more than Net_3. In addition, $Net_1+Net_2+Net_3$ reaches higher AUC and AUPR values than *Global Net*, which demonstrates that our ensemble framework with graph-level attention indeed boosts the prediction performance.

Table S2. Comparison performance between different network combinations under the settings CVS1, CVS2 and CVS3.

Networks	**CVS1**		**CVS2**		**CVS3**	
	AUC	**AUPR**	**AUC**	**AUPR**	**AUC**	**AUPR**
Global Net	0.8943±0.0114	0.8835±0.0153	0.8996±0.0168	0.8704±0.0236	0.8017±0.0597	0.7315±0.0859
Net_1	0.9527±0.0054	0.9189±0.0174	0.9519±0.0073	0.9181±0.0088	0.8032±0.0085	0.7326±0.0137
Net_2	0.9126±0.0140	0.9075±0.0196	0.9105±0.0122	0.9058±0.0186	0.7283±0.0942	0.6911±0.1005
Net_3	0.8677±0.0142	0.8473±0.0157	0.8738±0.0178	0.8630±0.0187	0.5692±0.1726	0.5748±0.1377
Net_1+Net_2	<u>0.9551±0.0054</u>	<u>0.9300±0.0145</u>	<u>0.9541±0.0059</u>	<u>0.9273±0.0161</u>	<u>0.8187±0.0126</u>	<u>0.7573±0.0202</u>
Net_1+Net_3	0.9542±0.0112	0.9170±0.0126	0.9536±0.0064	0.9232±0.0118	0.8064±0.0866	0.7481±0.0849
Net_2+Net_3	0.9139±0.0127	0.8942±0.0197	0.9158±0.0122	0.9064±0.0070	0.7345±0.1088	0.6952±0.0682
$Net_1+Net_2+Net_3$	**0.9586±0.0083**	**0.9460±0.0112**	**0.9612±0.0057**	**0.9398±0.0124**	**0.8232±0.0671**	**0.7655±0.0534**

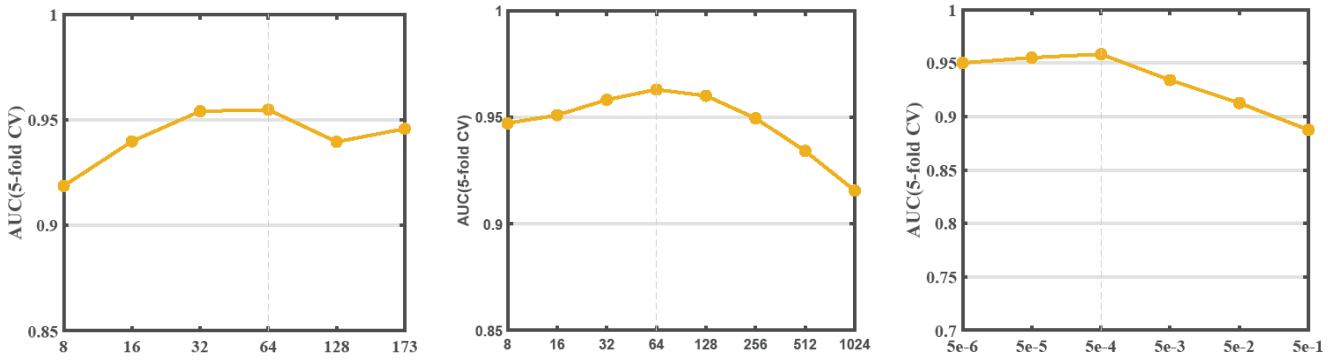
2.3 Parameter analysis

There are several important parameters in our model, such as the size of microbe genome feature k, the dimension of latent factor in GCN layer l and the weight factor γ. We also conduct experiments to measure their influences on the proposed model under the settings CVS1, CVS2 and CVS3. The results in Fig. S1, Fig. S2 and Fig. S3 indicate that our model achieves better performance when $k=64$, $l=64$ and $\gamma=0.0005$.



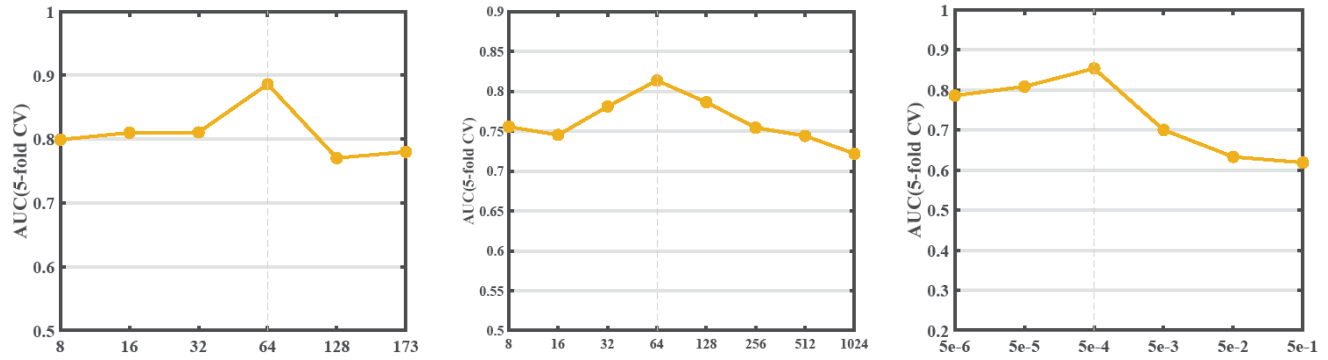
(a) k (b) l (c) γ

Fig. S1. Parameter sensitivity w.r.t. Size of microbe genome feature k, Dimension of latent factor in GCN layer l and Weight factor γ, under the setting CVS1.



(a) k (b) l (c) γ

Fig. S2. Parameter sensitivity w.r.t. Size of microbe genome feature k, Dimension of latent factor in GCN layer l and Weight factor γ, under the setting CVS2.



(a) k (b) l (c) γ

Fig. S3. Parameter sensitivity w.r.t. Size of microbe genome feature k, Dimension of latent factor in GCN layer l and Weight factor γ, under the setting CVS3.

2.4 Case study

For further validating the effectiveness of our model, we select two popular drugs (i.e., Ciprofloxacin and Moxifloxacin) and two common microbes (i.e., Pseudomonas aeruginosa and Escherichia coli) for case studies. Table S3 and Table S4 show the top predicted candidate microbes for Ciprofloxacin and Moxifloxacin, respectively. Table S5 and Table S6 show the top predicted associated drugs for Pseudomonas aeruginosa and Escherichia coli, respectively.

Table S3. Prediction results of the top 50 Ciprofloxacin-associated microbes. The first column records top 1-25 associated microbes. The third column records top 26-50 associated microbes.

Microbes	Evidences	Microbes	Evidences
Human immunodeficiency virus 1	PMID:9566552	Streptococcus sanguis	PMID: 11347679
Candida albicans	PMID:31471074	Micrococcus luteus	PMID:26954218
Staphylococcus epidermis	PMID:10632381	Clostridium perfringens	PMID:29978055
Staphylococcus epidermidis	PMID:28481197	Streptococcus gordonii	PMID:22887906
Enterococcus faecalis	PMID:27790716	Aggregatibacter actinomycetemcomitans	PMID: 23952779
Streptococcus mutans	PMID:30468214	Klebsiella pneumoniae	PMID:27257956
Vibrio harveyi	PMID:27247095	Mycobacterium avium	PMID:30012773
Salmonella enterica	PMID:26933017	Enterococcus faecium	PMID:28193670
Eikenella corrodens	PMID:16875802	Vibrio anguillarum	PMID:19146525
Burkholderia pseudomallei	PMID:24502667	Bacteroides fragilis	PMID: 12111577
Plasmodium falciparum	PMID:17214980	Yersinia enterocolitica	PMID: 29649750
Streptococcus pneumoniae	PMID:26100702	Burkholderia cepacia	PMID: 24976594
Enteric bacteria	PMID:27436461	Streptococcus mitis	PMID: 29968346
Actinomyces oris	Unconfirmed	Enterococcus gallinarum	PMID:29030312
Serratia marcescens	PMID:23751969	Candida tropicalis	Unconfirmed
Streptococcus epidermidis	Unconfirmed	Aspergillus fumigatus	PMID: 21911564
Listeria monocytogenes	PMID:28355096	Aeromonas hydrophila	PMID:23084650
Vibrio vulnificus	PMID:28971862	Kocuria rhizophila	Unconfirmed
Burkholderia cenocepacia	PMID:27799222	Salmonella Typhi	PMID: 30428828
Porphyromonas gingivalis	PMID: 15231772	Propionibacterium acnes	PMID:27801379
Proteus mirabilis	PMID:26953206	Streptococcus parasanguinis	PMID: 21193474
Streptococcus sanguinis	PMID:21507381	Hafnia alvei	PMID:28537065
Campylobacter jejuni	PMID:27900889	Shewanella oneidensis	PMID: 30673286
Burkholderia multivorans	PMID: 19633000	Bacillus cereus	PMID:26358183
Vibrio cholerae	PMID:28270803	Lysinibacillus sphaericus	Unconfirmed

Table S4. Prediction results of the top 50 Moxifloxacin-associated microbes. The first column records top 1-25 associated microbes. The third column records top 26-50 associated microbes.

Microbes	Evidences	Microbes	Evidences
Pseudomonas aeruginosa	PMID:31691651	Vibrio vulnificus	PMID: 12384368
Staphylococcus aureus	PMID:31689174	Campylobacter jejuni	PMID:16027651
Escherichia coli	PMID:31542319	Enteric bacteria and other eubacteria	Unconfirmed
Staphylococcus epidermis	PMID: 11249827	Mycobacterium tuberculosis	PMID:31713607
Staphylococcus epidermidis	PMID:31516359	Vibrio cholerae	PMID:16341343
Human immunodeficiency virus 1	Unconfirmed	Streptococcus epidermidis	Unconfirmed
Streptococcus mutans	PMID:29160117	Enterococcus faecium	PMID:23524466
Enterococcus faecalis	PMID:31763048	Streptococcus gordonii	PMID:29160117
Vibrio harveyi	Unconfirmed	Proteus vulgaris	PMID: 19692210
Salmonella enterica	PMID:22151215	Aggregatibacter actinomycetemcomitans	PMID: 31516229
Plasmodium falciparum	PMID:15125930	Yersinia enterocolitica	PMID: 15992072
Bacillus subtilis	PMID:30036828	Burkholderia cepacia	PMID: 11605808
Eikenella corrodens	PMID:14614671	Streptococcus mitis	PMID: 10629010
Streptococcus pneumoniae	PMID:31542319	Aeromonas hydrophila	PMID:12821471
Burkholderia pseudomallei	PMID:15731198	Aspergillus fumigatus	PMID: 19109335
Actinomyces oris	PMID: 26538502	Bacteroides fragilis	PMID: 30831235
Streptococcus sanguinis	PMID:10629010	Providencia stuartii	Unconfirmed
Burkholderia cenocepacia	Unconfirmed	Kocuria rhizophila	Unconfirmed
Listeria monocytogenes	PMID:28739228	Enterococcus gallinarum	PMID: 21444994
Serratia marcescens	PMID:17592324	Candida tropicalis	PMID:20455400
Burkholderia multivorans	Unconfirmed	Vibrio anguillarum	Unconfirmed

Streptococcus sanguis	PMID:10629010	Hafnia alvei	Unconfirmed
Porphyromonas gingivalis	PMID:30048853	Streptococcus parasanguinis	Unconfirmed
Proteus mirabilis	PMID:27351708	Salmonella Typhi	PMID: 27097699
Micrococcus luteus	PMID:24231380	Shewanella oneidensis	Unconfirmed

Table S5. Prediction results of the top 50 Pseudomonas aeruginosa-associated drugs. The first column records top 1-25 associated drugs. The third column records top 26-50 associated drugs.

Microbes	Evidences	Microbes	Evidences
3-tetradecanoylthiotetronic	PMID: 27999062	D/L-Aspartate	PMID: 27988856
Tea tree oil	PMID: 23581401	(10R,11R)-Hydnocarpin	Unconfirm
Clary sage oil	PMID: 23157022	(10R,11R)-Hydnocarpin D	Unconfirm
cat-AGE-RK1	Unconfirm	(10S,11S)-Hydnocarpin D	Unconfirm
cat-AGE-RK2	Unconfirm	(2Z)-N-hexyl-2-hydroxy-3-[4-(trifluoromethyl)phenyl]prop-2-enamide	Unconfirm
Tiliroside	PMID: 17137105	(5Z)-2-(1,1-dichloro-2-phenylethyl)-5-{[4-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
2,3-Dehydrosilybin	Unconfirm	(5Z)-2-(1,1-dichloro-3,3-dimethylbutyl)-5-(phenylmethylidene)-1,3-oxazolidin-4-one	Unconfirm
Citropin 1.1	PMID: 24779193	(5Z)-2-(1,1-dichloro-3,3-dimethylbutyl)-5-{[4-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
Ch_GG_Nt-Dhvar5	PMID: 25818458	(5Z)-2-(1,1-dichloro-7-methyloctyl)-5-{[4-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
Olive oil	PMID: 21153811	(5Z)-2-(1,1-dichlorobut-3-yn-1-yl)-5-{[4-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
Resorcinol	PMID: 28989246	(5Z)-2-(1,1-dichloropropyl)-5-{[4-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
3-tetra- decanoyltetronic	PMID: 27999062	(5Z)-2-(2,2-dimethylpropyl)-5-{[3-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	Unconfirm
Ethylene glycol tetraacetic acid	PMID: 19445464	2,3-dihydroxybenzoic acid	PMID: 24449781
(5Z)-2-(1,1-dichloro-3,3-dimethylbutyl)-5-[(3-methoxyphenyl)methylidene]-1,3-oxazolidin-4-one	Unconfirm	3-(2-Pyridinyl)alanine	Unconfirm
Daptomycin	PMID: 31364338	5- hydroxyethyl-3-tetradecanoyltetramic acid	PMID: 27999062
PTP-7	PMID: 27105736	alpha-tocopherol	PMID: 25458794
Phloridzin	Unconfirm	CAMA	PMID: 30849936
Silver dihydrogen citrate	Unconfirm	Cecropin (1-7)-melittin A (2-9) amide	Unconfirm
Tangerine oil	Unconfirm	Ch Dhvar5 ads	Unconfirm
cis-Stillbene	Unconfirm	Clavulanate	PMID: 31969294
Colostrum Hexasaccharide	Unconfirm	DASamP1	PMID: 28672834
Dichloromethane	PMID: 29375509	D-Aspartate	PMID: 25656450
(5Z)-2-tert-butyl-5-{[3-(trifluoromethyl)phenyl]methylidene}-1,3-oxazolidin-4-one	DD13	PMID: 28161851	
3-Hydroxytyrosine	Unconfirm	Dicinnamyl	Unconfirm
Biphenyl	PMID: 26519802	Diphenyl methane	PMID: 16345223

Table S6. Prediction results of the top 50 Escherichia coli-associated drugs. The first column records top 1-25 associated drugs. The third column records top 26-50 associated drugs.

Microbes	Evidences	Microbes	Evidences
OSIP108	PMID: 27491841	(2-(4-chlorophenyl)- 4-{[(6-methyl-2-pyridinyl)amino]methylene}-1,3-oxazol-5(4H)-one)	Uncomfirmed
L-carnosine	PMID: 20952637	[(ethanesulfinyl)sulfanyl]ethane	Uncomfirmed
(1Z)-ethylidene-lambda4-sulfanyliumolate	Uncomfirmed	1-(4-chlorophenyl)-3-m-tolyurea	Uncomfirmed
3-[(prop-2-ene-1-sulfinyl)sulfanyl]prop-1-ene	Uncomfirmed	1-(4-methoxyphenyl)-3-m-tolyurea	Uncomfirmed
S-(2-Aminoethyl)-L-cysteine hydrochloride	PMID: 18154269	1,3-dim-tolyurea	Uncomfirmed
S-ethyl-L-cysteine	PMID: 13768478	1-mesityl-3(4-nitrophenyl)urea	PMID: 27375583
S-methyl- L-cysteine	PMID: 13768478	2-amino-3-(prop-2-ene-1-sulfinyl)propanoic acid	Uncomfirmed
Farnesol	PMID: 27440491	2-amino-3-(propane-1-sulfinyl)propanoic acid	Uncomfirmed
S6L3-33	PMID: 31974490	2-amino-imidazole/ triazole conjugate	PMID: 24763714
C16-33	PMID: 22607313	2C-4	Uncomfirmed
M8-33	PMID: 17025157	AAP2	PMID: 19679242
M8G2	PMID: 9150204	Aloe-emodin	PMID: 15615409
KSL	PMID: 18503516	Anthraflavic acid	PMID: 30154376
Listerine	PMID: 29513232	Carminic acid	Uncomfirmed
Econazole	PMID: 27686306	Erythritol	Uncomfirmed
Chrysophsin-1	PMID: 26388176	Hsn5 12-mer	PMID: 16595638
MUC7 20-mer	PMID: 12543672	Hypericin	PMID: 29926201
Danthron	PMID: 2200948	IMB-2	Uncomfirmed
Melittin B	PMID: 30245684	K4-S4	PMID: 15019214
1-(4-chlorophenyl)-3-p-tolyurea	Uncomfirmed	Kaurenoic acid	PMID: 29340903
Caprolacton	PMID: 15576175	L-K6	PMID: 24496141
MUC7 12-mer-D	PMID: 17996119	MUC7 12-mer-L	PMID: 16595638
Sennidin A	Uncomfirmed	MUC7 12-mer-L4	Uncomfirmed
Sm6(L2)B33	Uncomfirmed	N -[(3-chloro-isoquinolin-4-yl)methylene] benzohydrazide	Uncomfirmed
(1E)-1-{[(1E)-prop-1-ene-1-sulfinyl]sulfanyl}prop-1-ene	Uncomfirmed	P15-CSP	Uncomfirmed

Reference

[1]. Ma, W. et al. (2017) An analysis of human microbe-disease associations. Brief. Bioinf., 18.

[2]. Janssens, Y. et al. (2018) Disbiome database: linking the microbiome to disease. BMC Microbiology, 18.

[3]. Sun, Y.-Z. et al. (2018) Mdad: a special resource for microbe-drug associations. Frontiers in cellular and infection microbiology,8.

[4]. Davis, A.P. et al. (2019) The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 47.

[5]. Wishart, D.S. et al. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res., 46.

[6]. Lipscomb, C.E. (2000) Medical subject headings (MeSH). Bulletin of the Medical Library Association, 88.

[7]. Hattori, M. et al. (2010). Simcomp/subcomp: chemical structure search servers for network analyses. Nucleic acids research,38(suppl_2), W652–W656.