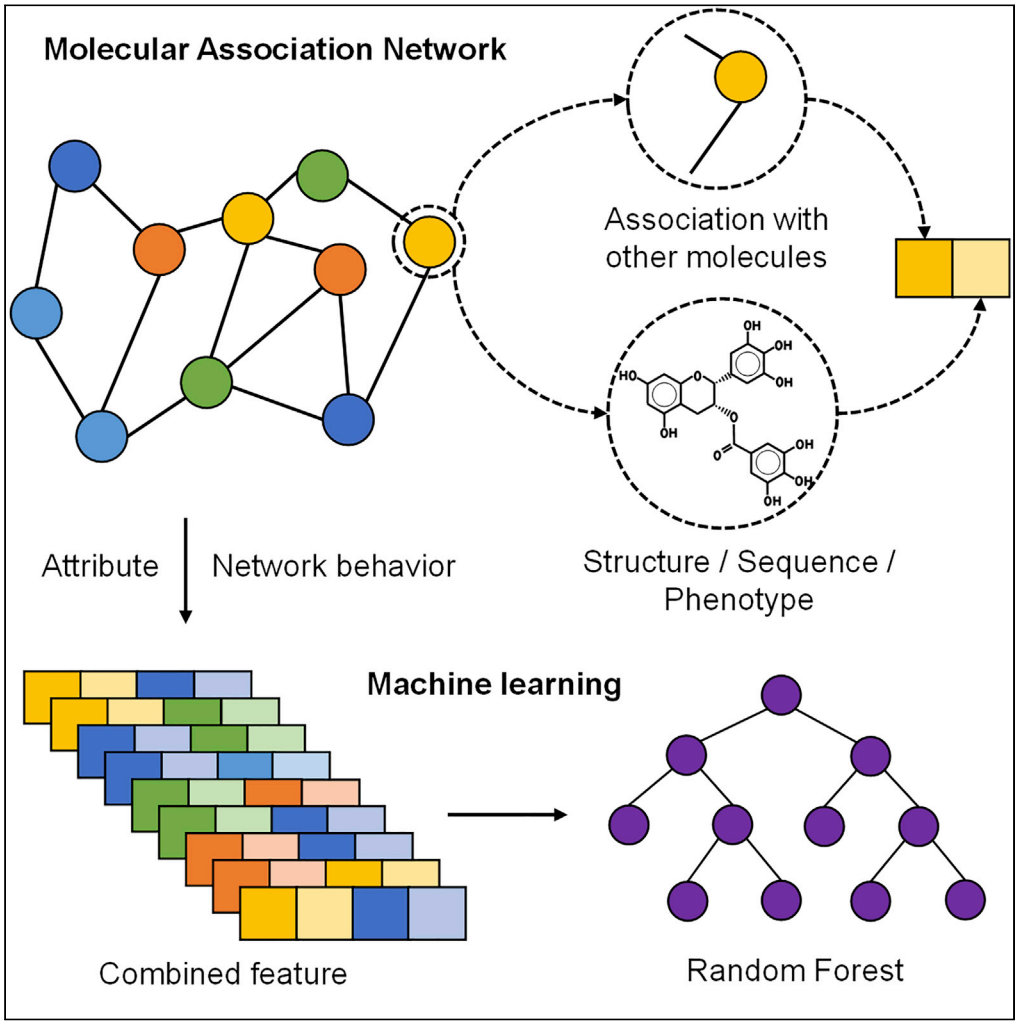
iScience

CellPress
OPEN ACCESS

Article

# Learning Representations to Predict Intermolecular Interactions on Large-Scale Heterogeneous Molecular Association Network




Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, Zhen-Hao Guo, Keith C.C. Chan, Yangming Li

zhuhongyou@ms.xjb.ac.cn

## HIGHLIGHTS

Systematically model a unified association network between multiple molecules

Learning network behavior and attribute feature of biological components

A machine learning method to infer any intermolecular association

Insights into complex systems between biological components from a holistic view

Yi et al., iScience *23*, 101261
July 24, 2020 
https://doi.org/10.1016/j.isci.2020.101261

iScience

CellPress
OPEN ACCESS

Article

# Learning Representations to Predict Intermolecular Interactions on Large-Scale Heterogeneous Molecular Association Network

Hai-Cheng Yi,[1,2] Zhu-Hong You,[1,2,6,*] De-Shuang Huang,[3] Zhen-Hao Guo,[1] Keith C.C. Chan,[4] and Yangming Li[5]

## SUMMARY

**Molecular components that are functionally interdependent in human cells constitute molecular association networks. Disease can be caused by disturbance of multiple molecular interactions. New biomolecular regulatory mechanisms can be revealed by discovering new biomolecular interactions. To this end, a heterogeneous molecular association network is formed by systematically integrating comprehensive associations between miRNAs, lncRNAs, circRNAs, mRNAs, proteins, drugs, microbes, and complex diseases. We propose a machine learning method for predicting intermolecular interactions, named MMI-Pred. More specifically, a network embedding model is developed to fully exploit the network behavior of biomolecules, and attribute features are also calculated. Then, these discriminative features are combined to train a random forest classifier to predict intermolecular interactions. MMI-Pred achieves an outstanding performance of 93.50% accuracy in hybrid associations prediction under 5-fold cross-validation. This work provides systematic landscape and machine learning method to model and infer complex associations between various biological components.**

## INTRODUCTION

A key goal of life science research is to understand the complex association between biomolecules in various functional systems of a cell, which is important for many biomedical researches, for instance, exploring the pathogenesis of cancer, analyzing genetic diseases, and developing drugs and vaccines. Various molecular components and their interactions play important roles in life activities in cells. For example, proteins are the direct bearers of many fundamental life activities (Zhang et al., 2012; You et al., 2010; Marcotte et al., 1999). Most drugs work by binding to a specific protein, altering its biochemical and/or biophysical activities, thereby having multiple effects on multiple functions (Ay et al., 2007). Emerging evidence shows that non-coding RNA (ncRNA), genes that cannot be translated into protein, also play a significant biological role in metabolism, tumorigenesis, and cellular development (Gibb et al., 2011; Yi et al., 2018; Bartel, 2004), including microRNAs, long ncRNAs, and circular RNAs. Microbes, as environment or co-evolved partner, also have critical impacts on human's health and disease (Dethlefsen et al., 2007; Ma et al., 2016; Jostins et al., 2012). These molecules and their synergistic interactions maintain the special cellular activities, operating as part of a highly interconnected molecular association network.

Owing to the rapid development of related molecular biology, computational biology, and omics research, many valuable researches on individual intermolecular associations in human were developed and a variety of valuable experimental data have been released, e.g., mRNA-protein interactions (McCarthy and Kollmus, 1995; Peritz et al., 2006), long non-coding RNA (lncRNA)-protein interactions (Yi et al., 2018), protein-protein interactions (You et al., 2017b), micro RNA (miRNA)-protein interactions (Dweep and Gretz, 2015), and miRNA-lncRNA interactions (Huang et al., 2017, 2018). Considering exogenous chemical compound or complex disease, there are drug-disease interactions (Wang et al., 2018; Dumbreck et al., 2015; Zhang et al., 2018), miRNA-disease associations (You et al., 2017a; Wang et al., 2019), drug-protein interactions (Hu et al., 2016; Li et al., 2017), protein-disease associations (Lee et al., 2012; Wang et al., 2018), and lncRNA-disease associations (Chen, 2015). Emerging research on circular RNA (circRNA) shows there are also circRNA-miRNA associations (Zhang et al., 2017), circRNA-protein interactions (Chen et al., 2017), and circRNA-disease associations (Zhao et al., 2018). The microbes and drugs also have been involved into many biological systems (Ma et al., 2016; Sun et al., 2018).

[1]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

[4]Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR 999077, China

[5]College of Engineering Technology, Rochester Institute of Technology, Rochester, NY 14623, USA

[6]Lead Contact

*Correspondence: zhuhongyou@ms.xjb.ac.cn

https://doi.org/10.1016/j.isci.2020.101261

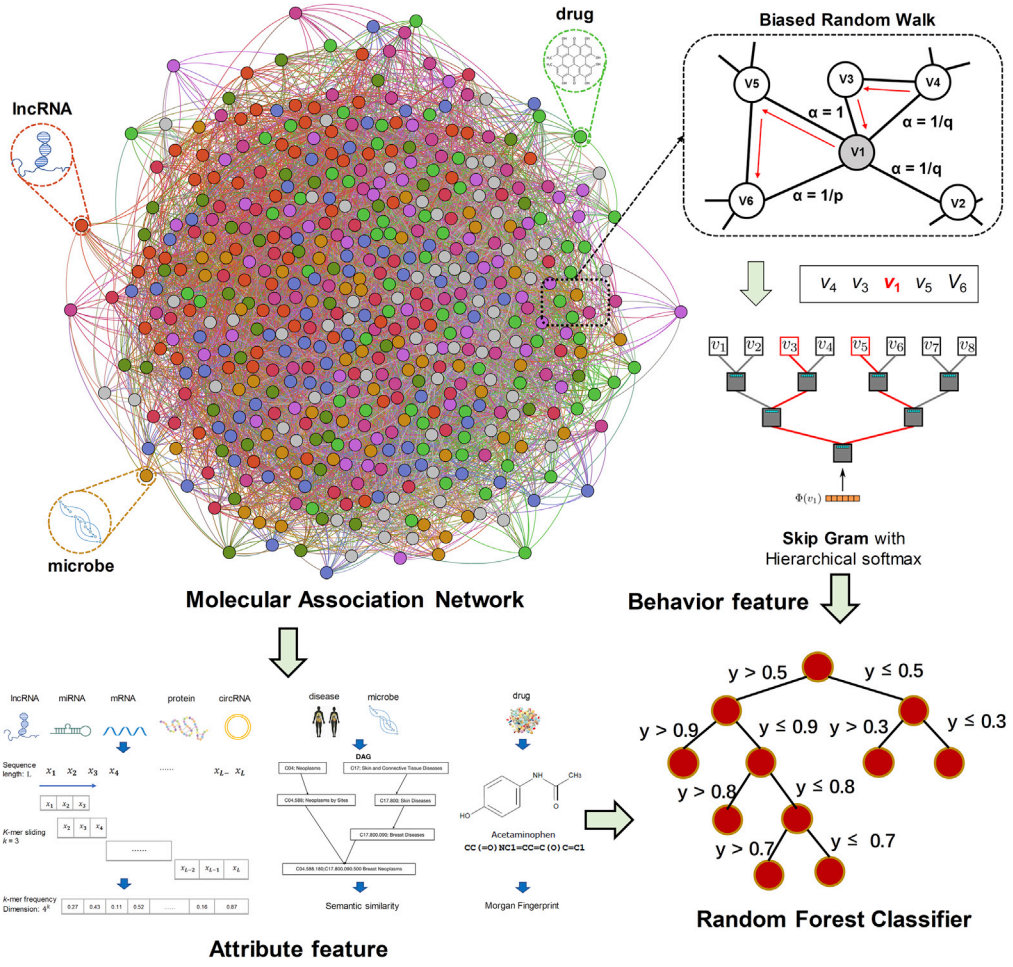



**Figure 1. The Workflow of the MMI-Pred**

The molecular association network is formed by connecting multitype intermolecular associations among mRNAs, proteins, miRNAs, lncRNAs, circRNAs, drugs, microbes, and diseases. Both the handcrafted attribute features and behavior features learned by network embedding method of biomolecules are jointly fed into a random forest classifier for training to predict potential intermolecular interactions.

These researches focus on individual associations between two molecules, and there are several studies that have considered the association between multiple biomolecules; e.g., Davis et al. manually compiled interactions among chemical, gene, and disease from publications to construct a chemical-gene-disease network (Davis et al., 2008). Liu et al. connected the associations between miRNA, target gene, lncRNA, and disease to build a network for calculating the similarity of miRNA and disease to predict miRNA-disease associations (Liu et al., 2017). The concerns of these studies are on two or very limited intermolecular relationships. However, intermolecular interactions are widespread and interconnected.

Inspired by this systematic perspective, to address some limitations of existing studies, we propose a molecular association network (MAN)-based framework to predict molecule-molecule associations by learning behavior and attribute feature of biomolecules, named MMI-Pred. The workflow of MMI-Pred is shown in Figure 1. First, a comprehensive MAN network is generated by connecting extensive associations between miRNAs, lncRNAs, circular RNAs, mRNAs, proteins, drugs, microbes, and diseases. It contains 14,315 molecular nodes and 18 kinds of, 114,150 association entries. Then, a random walk and skip-gram algorithm-based network embedding model node2vec is adopted to learn the behavioral features of biomolecular nodes. And the attribute feature is also calculated from sequence, structure, and phenotype information of different biomolecules. Moreover, both the attribute and behavior features are combined to train a Random Forest classifier (Breiman, 2001) to predict intermolecular associations. To evaluate the performance of MMI-Pred, the predictive ability of the entire MAN is

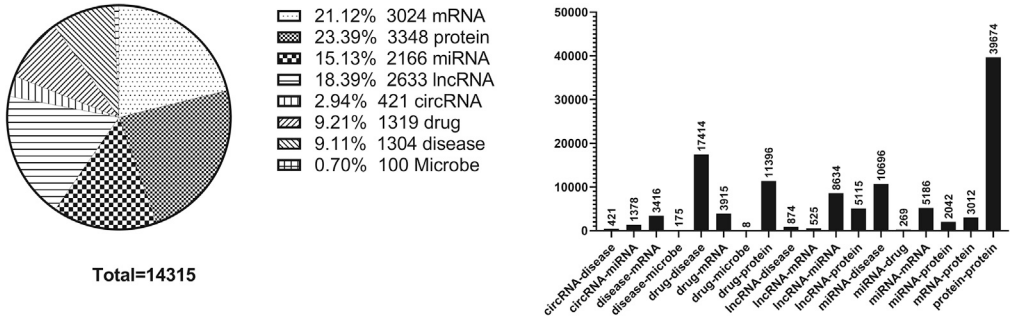



**Figure 2. The Number and Type Distribution of Biomolecule Nodes and Intermolecular Associations in the Molecular Association Network**

first evaluated under 5-fold cross-validation. Furthermore, MMI-Pred was applied to predict miRNAs most relevant to Breast Neoplasms and Colon Neoplasms as a case study. Experimental results demonstrate that this work brings new insights and a promising prediction method for discovering and understanding intermolecular associations.

## RESULTS

### Molecular Association Network

The extensive associations between mRNAs, proteins, miRNAs, drugs, lncRNAs, circRNAs, microbes, and diseases are interconnected and form a complex molecular association network. Considering that same biomolecule may have different naming in different databases, we used the same naming convention to unify the naming of the eight molecules (nodes), e.g., STRING ID for protein (Szklarczyk et al., 2018), miRBase ID for miRNA (Kozomara et al., 2018), NONCODE ID for lncRNA (Fang et al., 2017), circBase ID for circRNA (Glažar et al., 2014), DrugBank ID for drug (Wishart et al., 2017), and NIH MeSH ID for microbe and disease. Then, the duplicate and completely isolated associations are removed. Finally, there are 14,315 molecule nodes and 114,150 association links in the MAN. The distribution of molecules nodes and association types is shown in Figure 2. The MAN obtained 39,674 protein-protein interactions from STRING v11 (Szklarczyk et al., 2018); 421 circRNA-disease associations (links) from Circ2Disease (Yao et al., 2018), CircRNA disease (Zhao et al., 2018), LncRNADisease 2.0 (Bao et al., 2018), and CircR2Disease (Fan et al., 2018); 1,378 circRNA-miRNA associations from SomamiR 2.0 (Bhattacharya and Cui, 2015); 3,416 mRNA-disease associations from DisGeNET (Piñero et al., 2017); 175 microbe-disease associations from HMDAD (Ma et al., 2016); 17,414 drug-disease interactions from CTD (Davis et al., 2018); 3,915 drug-mRNA associations from PharmGKB (Hewett et al., 2002); 8 drug-microbe associations from PharmacoMicrobiomics (R Rizkallah et al., 2012); 11,396 drug-protein interactions from DrugBank (Wishart et al., 2017); 874 lncRNA-disease associations from LncRNADisease (Chen et al., 2012) and lncRNASNP2 (Miao et al., 2017); 525 lncRNA-mRNA interactions from LncRNA2Target (Cheng et al., 2018); 8,634 lncRNA-miRNA interactions from lncRNASNP2 (Miao et al., 2017); 5,115 lncRNA-protein interactions from NPInter v2.0 (Yuan et al., 2013); 10,696 miRNA-disease associations from HMDD (Li et al., 2013); 3,012 mRNA-protein associations from NCBI data; 269 miRNA-drug associations from SM2miR (Liu et al., 2012); 5,186 miRNA-mRNA associations from MiRTarBase (Chou et al., 2017); and 2,042 miRNA-protein interactions from NPInter v2.0 (Yuan et al., 2013) and TransmiR v2.0 (Tong et al., 2018).

### Predictive Performance Evaluation of MMI-Pred

The overall performance of MMI-Pred for predicting potential associations between arbitrary molecules in the MAN network was first evaluated under 5-fold cross-validation. In each fold validation, only the associations in the train set can be used to exploit the latent high-level representation of biomolecules nodes by network embedding model, which can avoid label leakage. As many studies have confirmed, there is a bias in measuring the performance of machine learning models using only precision or recall rates. When evaluating the classification performance of a model, the precision-recall curve and area under the precision-recall curve (AUPR) values that balance these two metrics are adopted. The overall performance of MMI-Pred is shown in Figure 3 and Table 1.

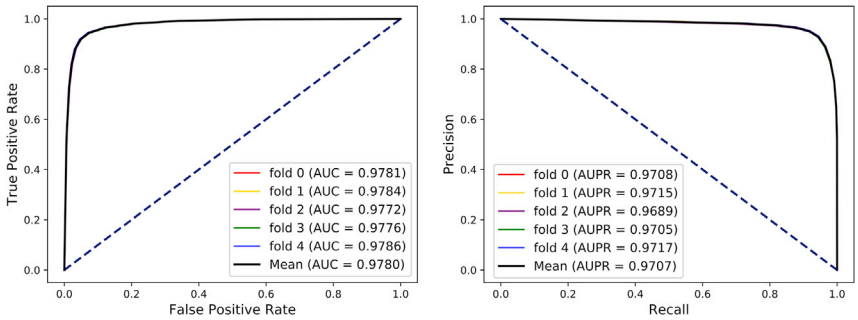



**Figure 3. The Performance of MMI-Pred on Entire MAN Dataset under 5-Fold Cross-Validation**
On the left is the ROC curve and AUC value, and on the right is the precision-recall curve and AUPR value.

As Figure 3 shows, in each fold cross-validation, the performance of MMI-Pred is very closed, which means the robust of our model. In whole MAN network, the model obtained a remarkable performance with high accuracy of 93.50% and high area under the curve (AUC) value of 0.9780. And the sensitivity, specificity, and precision of the model are 91.75%, 95.24%, 95.07%, respectively. The MMI-Pred receives a high AUPR value of 0.9707. In the case of class imbalance in classification tasks, accuracy is meaningless, for example, suppose there are 90 negative samples and 10 positive samples in a dataset, even if the model directly classified all samples into negative samples, the accuracy even is 90%, but this is obviously meaningless. And when the thresholds are different, the outputs are different. So, receiver operating characteristic (ROC) curve that can avoid these problems was used to measure our model's performance. The standard deviation (SD) of each performance value is 0.09%, 0.20%, 0.24%, 0.23%, 0.19%, and 0.06%, respectively, which can show the stable and robust of MMI-Pred in predicting any molecule-molecule associations in the MAN.

## Evaluate the Impact of Network Behavior and Attribute Feature

Molecules in the association network are similar to people in social networks, and they have both attributes and network behavior features. Both the network behavior and attribute features are adopted as representations of biomolecules. For mRNA, miRNA, lncRNA, circRNA, and protein, their attributes are nucleic acid or amino acid sequence. The k-mer is used to transfer sequences into numerical vector. For disease and microbe, their direct attribute is hard to gain, their phenotypes are employed to calculate their semantic similarity as attribute feature. The fingerprints of drug compounds that stand for the chemical structure are used as their attribute. All nodes in the MAN network can be calculated for their network embedding based on their behavior with other nodes in the network. We tested them under the same experimental conditions to verify the performance of these features and their impact on the predicted results.

As Figure 4 and Table 2 show, the MMI-Pred model can achieve high accuracy more than 90% whether using attribute features or behavior features, which indicate that the distinguishing power of features is acceptable. In general, the performance of behavior feature is a bit better than the attribute features, whereas the best performance is obtained when using both two features. In addition, when the nodes or network behavior attributes of some new molecules are missing, the combination of these two features can enhance the robustness of the model and ensure that the prediction can be performed normally.

## Compared with Widely Used Machine Learning Classifiers

To verify the impact of different machine learning models on performance, in this section, we compared the performance of the Logistic Regression (LR), AdaBoost, Naive Bayes (NB), XGBoost, and Random Forest as classifier of our framework using the attribute and behavior feature under the same experimental conditions. The Random Forest classifier and other contrast classifiers are implemented by Scikit-learn (Pedregosa et al., 2013) and use only default parameters.

As shown in Figure 5 and Table 3, the proposed method MMI-Pred that uses Random Forest classifier achieves the best performance. LR is a commonly used binary classification algorithm that directly models

| Fold | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC (%) |
|---|---|---|---|---|---|---|
| 0 | 93.42 | 91.64 | 95.2 | 95.02 | 86.9 | 97.81 |
| 1 | 93.51 | 91.60 | 95.43 | 95.25 | 87.09 | 97.84 |
| 2 | 93.48 | 92.08 | 94.87 | 94.72 | 86.99 | 97.72 |
| 3 | 93.43 | 91.62 | 95.24 | 95.06 | 86.91 | 97.76 |
| 4 | 93.64 | 91.82 | 95.47 | 95.3 | 87.35 | 97.86 |
| Average | 93.50 ± 0.09 | 91.75 ± 0.20 | 95.24 ± 0.24 | 95.07 ± 0.23 | 87.05 ± 0.19 | 97.80 ± 0.06 |

**Table 1. The 5-Fold Cross-Validation Performance of MMI-Pred on MAN Dataset**

the possibility of classification without the assumption of data distribution in advance. AdaBoost is the most famous representative of the Boosting algorithm. It requires the base classifier to learn specific data distributions, which can be achieved by re-weighting. The NB classifier is a series of simple probability classifiers based on Bayesian theorem based on independence between hypothetical features. XGBoost is an improvement of the Gradient Boosting Decision Tree (GBDT) implementation. Random Forest is an efficient, fast, and easy-to-use decision tree-based algorithm, which was proved to be the most effective model in this task by rigorous experimental results.

## Case Study: Predicting *Human* Disease-Associated miRNAs

To demonstrate the predictive ability of the proposed model on specific types of interactions, the MMI-Pred was executed to predict the miRNAs that are most relevant to two diseases, including Breast neoplasms and Colon neoplasms, as case studies. In the MAN, all miRNA-disease associations are from the HMDD database. When conducting case studies for individual disease, we trained the MMI-Pred predictor with a MAN network that removed those miRNA-Breast neoplasms (or Colon neoplasms) association pairs that overlapped with the dbDEMC 2.0 database (Yang et al., 2016). Then, the trained model performs prediction on testing Breast neoplasms or Colon neoplasms-miRNAs pairs. This processing can also be considered as cross-dataset validation. In the context of screening for disease-associated miRNAs, the candidate rankings are more valuable than the report of the overall false-positive, false-negative, and other indicators of the framework. Therefore, when the MMI-Pred is executed on the test samples, we rank the possible associated miRNAs based on the probability values output by the MMI-Pred. And then, the top 30 high-scored miRNAs for each disease are validated through the dbDEMC database.

Breast cancer is the most terrible killer of women's health. In 2018, about 2.1 million new cases of Breast tumor in women were diagnosed globally. And breast cancer accounts for about a quarter of the globally diagnosed cases of female cancer (Bray et al., 2018). Among the world's latest cancer incidence rates, female breast cancer also ranks second, accounting for 11.6% of the total cancer population. Studies have shown that miRNAs have the most significant expression difference between normal and cancer tissues, which can be used as tumor markers (Iorio et al., 2005). As shown in Table 4, the top 30 highest ranked breast cancer-associated miRNAs are predicted by MMI-Pred, and 25 of them were confirmed.

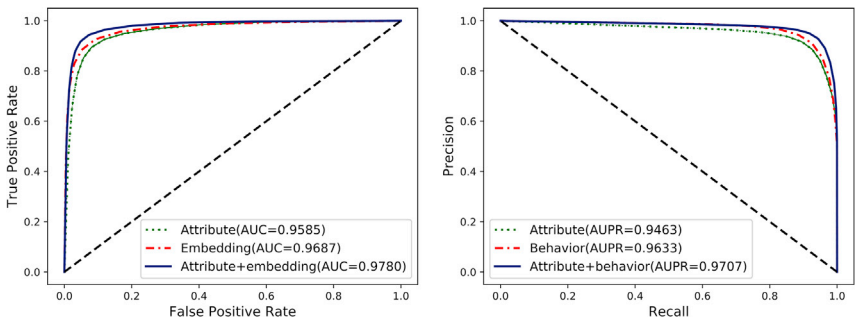



**Figure 4. The Comparison of Network Behavior and Attribute Features Using Random Forest Classifier**
On the left is the ROC curve and AUC value, and on the right is the PR curve and AUPR value.

| Feature | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC (%) |
|---|---|---|---|---|---|---|
| Attribute | 90.69 ± 0.14 | 89.49 ± 0.19 | 91.89 ± 0.15 | 91.69 ± 0.15 | 81.40 ± 0.27 | 95.85 ± 0.12 |
| Behavior | 91.64 ± 0.18 | 88.44 ± 0.15 | 94.83 ± 0.21 | 94.48 ± 0.23 | 83.45 ± 0.36 | 96.87 ± 0.17 |
| Combined | 93.50 ± 0.09 | 91.75 ± 0.20 | 95.24 ± 0.24 | 95.07 ± 0.23 | 87.05 ± 0.19 | 97.80 ± 0.06 |

**Table 2. Comparison of Attribute and Behavior Features Using Random Forest Classifier under 5-Fold Cross-Validation**

Colon cancer ranks fourth in overall cancer incidence, accounting for 6.1%, but ranks second in mortality, accounting for 9.2% (Bray et al., 2018). And recent research confirms that miRNAs play a role in carcinogenesis through DNA methylation and histone modifications and *human* colorectal tumorigenesis (Bandres et al., 2009). The predicted top 30 miRNAs with the highest score that associated with Colon Neoplasms are shown in Table 5; among them, 26 of miRNA-disease associations were confirmed.

## DISCUSSION

In this research, we proposed a computational framework based on network representation learning to predict any associations between molecules. First, the molecular association network is constructed by integrating 18 types of associations, 14,315 nodes, 114,150 molecular associations between mRNA, lncRNA, protein, miRNA, circRNA, drug, disease, and microbe. The performance of the framework is evaluated on the entire network under 5-fold cross-validation. To demonstrate the predictive ability, we use MMI-Pred to predict miRNAs most relevant to Breast cancer and Colon cancer as case studies. Experimental results proved that the MMI-Pred can predict any potential associations between molecules. Moreover, network embedding representations obtained based on MAN network and network representation learning algorithms can serve as efficient low-rank representations of disease, microbes, and other biological components whose features are difficult to be extracted by computational algorithms. In addition, randomly sampled unknown samples without known association are used as negative samples in this work; high-quality negative samples or sampling techniques are worth studying. It is anticipated that this work can help to advance related intermolecular associations research in a long term.

### Limitations of the Study

In this study, we provide a systematic and holistic perspective on intermolecular interactions and provide a machine learning method to model molecular properties and intermolecular behaviors in order to promote understanding and discover new intermolecular interactions. This work still has some limitations that deserve attention and further study. First, the interactions screened from public databases for building molecular association networks are not complete, although to our best knowledge, these databases are already of high quality and relatively comprehensive. For nodes that do not exist in the network, the network embedding feature will be not applicable. More complete data will be more conducive to comprehensive modeling of the relationship between biomolecules. Second, the MAN network is a heterogeneous information network that contains many types of molecules and many different association relationships. When characterizing the network behavior of biomolecule nodes, the network embedding algorithm

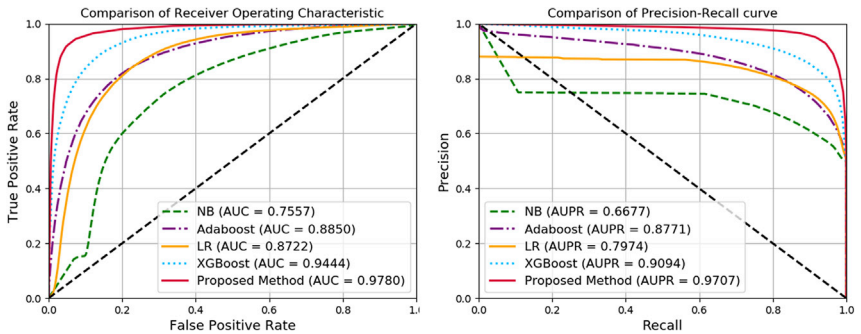



**Figure 5. The Performance Comparison between MMI-Pred and Four Different Comparison Models Include Naive Bayes, Adaboost, Logistic Regression, and XGBoost Classifiers**

| Method | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC |
|---|---|---|---|---|---|---|
| NB | 59.64 | 31.35 | 87.94 | 72.2 | 23.38 | 75.57 |
| LR | 80.61 | 82.44 | 78.79 | 79.54 | 61.27 | 87.21 |
| AdaBoost | 80.91 | 82.68 | 79.14 | 79.86 | 61.86 | 88.5 |
| XGBoost | 85.67 | 78.66 | 92.68 | 91.48 | 72.05 | 94.44 |
| Proposed method | 93.50 | 91.75 | 95.24 | 95.07 | 87.05 | 97.80 |

**Table 3. The Performance Comparison of Different Machine Learning Classifiers**

does not use the heterogeneous information. The further study of network representation learning algorithms for heterogeneous information networks will be very helpful.

## METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

### Resource Availability

#### *Lead Contact*

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Zhu-Hong You (zhuhongyou@ms.xjb.ac.cn).

#### *Materials Availability*

This study did not generate new materials.

#### *Data and Code Availability*

The datasets/code generated during this study are available at https://github.com/haichengyi/MAN.

| miRNA | dbDEMC | miRNA | dbDEMC |
|---|---|---|---|
| hsa-mir-186-5p | Confirmed | hsa-mir-539-5p | Confirmed |
| hsa-mir-216a-5p | Unconfirmed | hsa-mir-330-5p | Confirmed |
| hsa-mir-154-5p | Confirmed | hsa-mir-543 | Confirmed |
| hsa-mir-181d-5p | Confirmed | hsa-mir-4262 | Unconfirmed |
| hsa-mir-449b | Confirmed | hsa-mir-384 | Confirmed |
| hsa-mir-211-5p | Confirmed | hsa-mir-4458 | Confirmed |
| hsa-mir-504-5p | Unconfirmed | hsa-mir-28-5p | Confirmed |
| hsa-mir-1271-5p | Confirmed | hsa-mir-136-5p | Confirmed |
| hsa-mir-300 | Confirmed | hsa-mir-99b-5p | Confirmed |
| hsa-mir-337-5p | Confirmed | hsa-mir-518-5p | Unconfirmed |
| hsa-mir-637 | Confirmed | hsa-mir-217 | Confirmed |
| hsa-mir-517a-3p | Confirmed | hsa-mir-664 | Confirmed |
| hsa-mir-671-5p | Confirmed | hsa-mir-508-5p | Confirmed |
| hsa-mir-525-5p | Unconfirmed | hsa-mir-431-5p | Confirmed |
| hsa-mir-532-5p | Confirmed | hsa-mir-483-5p | Confirmed |

**Table 4. The Top 30 miRNAs Relevant to Breast Cancer Predicted by MMI-Pred**

| miRNA | dbDEMC | miRNA | dbDEMC |
|---|---|---|---|
| hsa-mir-186-5p | Confirmed | hsa-mir-16-5p | Confirmed |
| hsa-mir-485-5p | Confirmed | hsa-mir-497-5p | Confirmed |
| hsa-mir-206 | Confirmed | hsa-mir-33b-5p | Confirmed |
| hsa-mir-19b-3p | Confirmed | hsa-mir-7-5p | Unconfirmed |
| hsa-mir-361-5p | Confirmed | hsa-mir-185-5p | Confirmed |
| hsa-mir-154-5p | Confirmed | hsa-mir-26b-5p | Confirmed |
| hsa-mir-9-5p | Unconfirmed | hsa-mir-34c-5p | Confirmed |
| hsa-mir-122-5p | Confirmed | hsa-mir-449b-5p | Confirmed |
| hsa-mir-590-5p | Confirmed | hsa-mir-139-5p | Confirmed |
| hsa-mir-340-5p | Confirmed | hsa-mir-134-5p | Unconfirmed |
| hsa-mir-211-5p | Confirmed | hsa-mir-153-3p | Unconfirmed |
| hsa-mir-149-5p | Confirmed | hsa-mir-449a-5p | Confirmed |
| hsa-mir-183-5p | Confirmed | hsa-mir-129-5p | Confirmed |
| hsa-mir-503-5p | Confirmed | hsa-mir-136-5p | Confirmed |
| hsa-mir-324-5p | Confirmed | hsa-mir-10a-5p | Confirmed |

**Table 5. The Top 30 miRNAs Relevant to Colon Cancer Predicted by MMI-Pred**

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101261.

## ACKNOWLEDGMENTS

This work was supported by the National Outstanding Youth Science Foundation of NSFC, under Grant no. 61722212, in part by the National Natural Science Foundation of China under Grants nos. 61873212, 61861146002, and 61732012.

## AUTHOR CONTRIBUTIONS

H.-C.Y. and Z.-H.Y. designed and conceived the algorithm, carried out analyses, prepared the datasets, carried out experiments, and wrote the manuscript; D.-S.H., Z.-H.G., K.C.C.C., and Y.L. performed and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 3, 2019
Revised: April 29, 2020
Accepted: June 8, 2020
Published: July 24, 2020

## REFERENCES

Ay, M., Goh, K.-I., Cusick, M.E., Barabasi, A.-L., and Vidal, M. (2007). Drug–target network. Nat. Biotechnol. *25*, 1119–1127.

Bandres, E., Agirre, X., Bitarte, N., Ramirez, N., Zarate, R., Romangomez, J., Prosper, F., and Garciafoncillas, J. (2009). Epigenetic regulation of microRNA expression in colorectal cancer. Int. J. Cancer *125*, 2737–2743.

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2018). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. *47*, D1034–D1037.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bhattacharya, A., and Cui, Y. (2015). SomamiR 2.0: a database of cancer somatic mutations altering microRNA–ceRNA interactions. Nucleic Acids Res. *44*, D1005–D1010.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of

incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Chen, X. (2015). Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci. Rep. *5*, 13186.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2012). LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. *41*, D983–D986.

Chen, Y.G., Satpathy, A.T., and Chang, H.Y. (2017). Gene regulation in the immune system by long noncoding RNAs. Nat. Immunol. *18*, 962.

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2018). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47*, D140–D144.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., and Lee, W.-H. (2017). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. *46*, D296–D302.

Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., and Mattingly, C.J. (2008). Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. Nucleic Acids Res. *37*, D786–D792.

Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2018). The comparative toxicogenomics database: update 2019. Nucleic Acids Res. *47*, D948–D954.

Dethlefsen, L., McFall-Ngai, M., and Relman, D.A. (2007). An ecological and evolutionary perspective on human–microbe mutualism and disease. Nature *449*, 811.

Dumbreck, S., Flynn, A., Nairn, M., Wilson, M., Treweek, S., Mercer, S.W., Alderson, P., Thompson, A., Payne, K., and Guthrie, B. (2015). Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines. BMJ *350*, h949.

Dweep, H., and Gretz, N. (2015). miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. Nat. Methods *12*, 697.

Fan, C., Lei, X., Fang, Z., Jiang, Q., and Wu, F.-X. (2018). CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database (Oxford) *2018*, bay044.

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., and Sun, X. (2017). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. *46*, D308–D314.

Gibb, E.A., Brown, C.J., and Lam, W.L. (2011). The functional role of long non-coding RNA in human carcinomas. Mol. Cancer *10*, 38.

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA *20*, 1666–1670.

Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B., and Klein, T.E. (2002). PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res. *30*, 163–165.

Chan, K.C., Hu, P.-W., and You, Z.-H. (2016). Large-scale prediction of drug-target interactions from deep representations. In 2016 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 1236–1243.

Huang, Y.-A., Chan, K.C., and You, Z.-H. (2017). Constructing prediction models from expression profiles for large scale lncRNA–miRNA interaction profiling. Bioinformatics *34*, 812–819.

Huang, Z.-A., Huang, Y.-A., You, Z.-H., Zhu, Z., and Sun, Y. (2018). Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph. BMC Med. Genomics *11*, 113.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., and Campiglio, M. (2005). MicroRNA gene expression deregulation in human breast cancer. Cancer Res. *65*, 7065–7070.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A., et al. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2018). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, D155–D162.

Lee, H.S., Bae, T., Lee, J.-H., Kim, D.G., Oh, Y.S., Jang, Y., Kim, J.-T., Lee, J.-J., Innocenti, A., and Supuran, C.T. (2012). Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. BMC Syst. Biol. *6*, 80.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2013). HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. *42*, D1070–D1074.

Li, Z., Han, P., You, Z.H., Li, X., Zhang, Y., Yu, H., Nie, R., and Chen, X. (2017). In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. Sci. Rep. *7*, 11174.

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., Yu, X., Li, X., and Jiang, W. (2012). SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. Bioinformatics *29*, 409–411.

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans. Comput. Biol. Bioinform. *14*, 905–915.

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., Yang, J., Kong, W., Zhou, X., and Cui, Q. (2016). An analysis of human microbe–disease associations. Brief. Bioinform. *18*, 85–97.

Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751–753.

McCarthy, J.E., and Kollmus, H. (1995). Cytoplasmic mRNA-protein interactions in eukaryotic gene expression. Trends Biochem. Sci. *20*, 191–197.

Miao, Y.-R., Liu, W., Zhang, Q., and Guo, A.-Y. (2017). lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. Nucleic Acids Res. *46*, D276–D280.

Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Vanderplas, J. (2013). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Peritz, T., Zeng, F., Kannanayakal, T.J., Kilk, K., Eiríksdóttir, E., Langel, U., and Eberwine, J. (2006). Immunoprecipitation of mRNA-protein complexes. Nat. Protoc. *1*, 577.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45* (D1), D833–D839.

R Rizkallah, M., Gamal-Eldin, S., Saad, R., and K Aziz, R. (2012). The pharmacomicrobiomics portal: a database for drug-microbiome interactions. Curr. Pharmacogenomics Personal. Med. *10*, 195–203.

Sun, Y.-Z., Zhang, D.-H., Cai, S.-B., Ming, Z., Li, J.-Q., and Chen, X. (2018). MDAD: a special resource for microbe-drug associations. Front. Cell. Infect. Microbiol. *8*, 424.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., and Bork, P. (2018). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tong, Z., Cui, Q., Wang, J., and Zhou, Y. (2018). TransmiR v2. 0: an updated transcription factor-microRNA regulation database. Nucleic Acids Res. *47*, D253–D258.

Wang, R., Li, S., Wong, M.H., and Leung, K.S. (2018). Drug-protein-disease association prediction and drug repositioning based on tensor decomposition. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), pp. 305–312.

Wang, L., You, Z.-H., Chen, X., Li, Y.-M., Dong, Y.-N., Li, L.-P., and Zheng, K. (2019). LMTRDA: using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput. Biol. *15*, e1006865.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., johnson, D., LI, C., and Sayeeda, Z. (2017). DrugBank 5.0: a major update to the DrugBank

database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., and Teschendorff, A.E. (2016). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. *45*, D812–D818.

Yao, D., Zhang, L., Zheng, M., Sun, X., Lu, Y., and Liu, P. (2018). Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci. Rep. *8*, 11018.

Yi, H.-C., You, Z.-H., Huang, D.-S., Li, X., Jiang, T.-H., and Li, L.-P. (2018). A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. Mol. Ther. Nucleic Acids *11*, 337–344.

You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics *26*, 2744–2751.

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., and Chen, X. (2017a). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

You, Z.-H., Zhou, M., Luo, X., and Li, S. (2017b). Highly efficient framework for predicting interactions between proteins. IEEE Trans. Cybern. *47*, 731–743.

Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2013). NPInter v2. 0: an updated database of ncRNA interactions. Nucleic Acids Res. *42*, D104–D108.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., and Hunter, T. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature *490*, 556.

Zhang, P., Meng, X., Chen, H., Liu, Y., Xue, J., Zhou, Y., and Chen, M. (2017). PlantCircNet: a database for plant circRNA–miRNA–mRNA regulatory networks. Database (Oxford) *2017*, bax089.

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018). Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods *145*, 51–59.

Zhao, Z., Wang, K., Wu, F., Wang, W., Zhang, K., Hu, H., Liu, Y., and Jiang, T. (2018). circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell Death Dis. *9*, 475.

**iScience, Volume *23***

## Supplemental Information

# Learning Representations to Predict Intermolecular Interactions on Large-Scale Heterogeneous Molecular Association Network

**Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, Zhen-Hao Guo, Keith C.C. Chan, and Yangming Li**

## Transparent Methods

### Construction of balanced data set

The molecular association network contains 14,315 nodes and 114,150 known associations. These known associations are positive samples. The association between these biomolecule nodes forms a symmetric adjacency matrix of 14,315 * 14,315, the possible intermolecular association is $(14315 \times 14315 - 14315) \div 2 = 102{,}452{,}455$ (subtract the samples on the diagonal of the adjacency matrix). However, the known positive samples only accounts for about 0.1%, hence there is a serious imbalance between positive and negative samples. In order to train and build a machine learning model, molecule pairs without known associations are randomly sampled and used as negative samples. The number of negative samples is the same as the number of positive samples. Finally, a balanced data set with 114,150 positive samples and 114,150 negative samples is constructed.

### Learning network behavior of molecular nodes

As an important data structure, network contains rich information about the nodes and the associations between nodes. The connection of molecular nodes to other nodes in the MAN network can be regarded as its synergistic relationship with other biomolecules in biological functions. These extensive connections can be regarded as the network behavior of molecules. To obtain the behavior of molecular nodes in the MAN, we introduce a random walk and skip-gram based network representation learning method, node2vec (Grover and Leskovec, 2016) to learn the latent network behavior embedding of molecules.

Node2vec is a graph embedding method that considers Breadth-First-Search (BFS) neighborhoods and Depth-First Search (DFS) neighborhoods, which be seen as an extension of DeepWalk (Perozzi et al., 2014) that combines DFS and BFS random walks. The method simulates a random walk of each node with a step size of *l*, wherein the *i*-th node $c(i)$ in the walk can be described as follows:

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{v,x}}{Z}, & if\ v,x\ in\ E \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Where $Z$ represents normalization constant, and $\pi_{v,x}$ indicates the unstandardized transition probability of nodes $v$ and $x$:

$$\pi_{v,x} = \alpha_{pq}(t,x) * \omega_{v,x} \quad (2)$$

where $\omega_{v,x}$ is the weight of the edges $v$ and $x$, and the unweighted graph used in this experiment is set to 1. Node2vec introduces two hyperparameters *q* and *p* to control node chain sampling strategy, suppose the current random walk passes the edge (*t*, *v*) to the vertex *v*, $\alpha_{pq}(t,x)$ is used to adjust the random walk process, interpolating between BFS and DFS. It can be defined as follows:

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & if\ d_{tx} = 0 \\ 1 & if\ d_{tx} = 1 \\ \frac{1}{q} & if\ d_{tx} = 2 \end{cases} \quad (3)$$

Where $d_{tx}$ is the shortest distance between $t$ and $x$. $p$ and $q$ are the Return parameter and the In-out parameter, respectively.

Suppose $f(u)$ be a mapping function that maps the vertex *u* to the embedding vector. For

each vertex *u* in the graph, define $N_s(U)$ as the set of neighboring vertices of the vertex *u* sampled by the sampling strategy *S* .The goal of node2vec optimization is to maximize the probability that its neighbor vertices (how important it is to define neighbor vertices) will appear under the condition of each vertex.

$$max_f \sum_{u \in v} log P_r(N_S(U)|f(u)) \quad (4)$$

In order to make the above optimization problem solvable, the article proposes two hypotheses: conditional independence hypothesis and feature space symmetry hypothesis.

Suppose that under a given source vertex, the probability of its neighbor vertices appearing is independent of the rest of the neighbors in the set of neighbors.

$$P_r\big(N_S(u)\big|f(u)\big) = \prod_{n_i \in N_S(u)} P_r\big(n_i\big|f(u)\big) \quad (5)$$

Feature space symmetry hypothesis, that is, one vertex shares the same set of embedding vectors as the source vertex and as the neighbor vertex. Under this assumption, the above conditional probability formula can be expressed as:

$$P_r\big(n_i\big|f(u)\big) = \frac{\exp f(n_i)\cdot f(u)}{\sum_{u \in v} \exp f(v)\cdot f(u)} \quad (6)$$

According to the above two assumptions, the final objective function is expressed as:

$$max_f \sum_{u \in v}[-\log Z_u + \sum_{n_i \in N_S(u)} \exp\left(f(n_i)\cdot f(u)\right)] \quad (7)$$

$$Z_u = \sum_{n_i \in N_S(u)} \exp\left(f(n_i)\cdot f(u)\right) \quad (8)$$

Since the normalization factor $Z_u$ is computationally expensive, it is optimized using negative sampling techniques.

**Attribute feature exaction of multiple biological components**

As the behavior feature of molecule nodes are learned from the associations with other molecules. These nodes also have its attribute features, such as sequence information for mRNA, miRNA, lncRNA, circRNA, and protein, chemical structure for drug compounds, phenotype for disease and microbe. To fully exploit these attribute feature for molecules, the widely used feature extraction method in each related research is adopted. For biological sequence, we applied the *k*-mer frequency to represent the attribute information. That is, for mRNA, miRNA, lncRNA and circRNA, the 3-mer is uses to encode the sequence information; for protein, the 20 amino acids are firstly reduced into 4 group inspired by (Shen et al., 2007), then, each sequence of protein also can be represented by 3-mer. On chemical structure, we convert Simplified Molecular Input Line Entry Specification (SMILES) (Weininger, 1988) into Morgan fingerprints. For disease and microbe, the Directed Acyclic Graph (DAG) that constructed by Medical Subject Headings (MeSH) descriptor were employed to obtain the semantic similarity as attribute.

The MeSH descriptor is used to represent the semantic information of disease and microbe and construct a DAG. In the DAG, the two ends of each edge side are the parent and child nodes, respectively (Wang et al., 2010). If the disease *d(j)* is the parent of the disease *d(i)*, their relationship can be described as:

$$DAG_{d(i)} = (d(i), N_{d(i)}, E_{d(i)}) \quad (9)$$

where $N_{d(i)}$ is the points set for all diseases in the $DAG_{d(i)}$. $E_{d(i)}$ is an set that contains all edges between nodes in the $DAG_{d(i)}$.

The relationship between a superior disease $d(i)$ and a subordinate disease $s$ can be defined as:

$$D_{d(i)}(s) = \begin{cases} 1, & if\ s = p(i) \\ \max\{\varepsilon \cdot D_{p(i)}(\acute{s}) | \acute{s} \in children\ of\ s\}, & if\ s \neq p(i) \end{cases} \quad (10)$$

Where $\varepsilon$ is the semantic contribution factor associated with disease $s$ and $\acute{s}$. The contribution of *s* will be reduced when *s* and $d(i)$ are different. In addition, the contribution of disease $d(i)$ is set to 1. Therefore, the semantic value $DV(d(i))$ of $d(i)$ can be obtained by the formula as follows:

$$DV(\mathrm{d(i)}) = \sum_{\mathrm{s} \in \mathrm{N_{d(i)}}} \mathrm{D_{d(i)}(s)} \quad (11)$$

The calculation method of DAG similarity value $Simi(d(i), d(j))$ of the disease $d(i)$ and disease $d(j)$ is similar to the Jaccard similarity coefficient, which is calculated as:

$$Simi(\mathrm{d(i), d(j)}) = \frac{\sum_{\mathrm{s} \in \mathrm{N_{d(i)}} \cap \mathrm{N_{d(j)}}} (\mathrm{D_{d(i)}(s) + D_{d(j)}(s)})}{\mathrm{DV(d(i)) + DV(d(j))}} \quad (12)$$

**Unify attribute feature dimension of different components**

Considering that the features of different types of biomolecules extracted from the abovementioned attribute representation methods are not uniform in dimension, and each is in a different feature space. Deep Stacked Autoencoder (SA) model is employed to learn latent representations with uniform dimensions of different types of nodes. Autoencoder is a data compression algorithm, in which data compression and decompression functions are data-dependent, lossy, and automatically learned from samples (Vincent et al., 2010). In most cases where autoencoders are mentioned, the compression and decompression functions are implemented by neural networks. At present, there are two main applications of autoencoders. The first is data denoising, and the second is dimension reduction for visualization. With proper dimensionality and sparse constraints, SA can learn more effective data projections than data dimensionality reduction technologies such as Principal Component Analysis (PCA) (Wold et al., 1987). The output *O*(*x*) of SA for an input *x* can be defined as:

$$O_{(w,b)}(x) = f(W^T x) = f(\sum_{i=1}^{n} w_i x_i + b) \quad (13)$$

Where the *f* means a nonlinear activation function:

$$f(\mathrm{t}) = max(0, Wt + b) \quad (14)$$

The learning goal of SA is to minimize the loss between $\hat{x}$ and *x*. We can define the loss function as:

$$\mathcal{L} = \sum_{i=1}^{n} \|\hat{x}_i - x_i\|_2^2 \quad (15)$$

**Performance evaluation strategy**

The widely used machine learning model evaluation metrics are applied to evaluate the performance of MMI-Pred. And the five-fold cross-validation strategy is also followed to obtain robust performance evaluation results. All the data are randomly divided into five parts, four of which are taken as the training data each time, and the remaining part is treated as the test data, and the cycle was repeated five times. There is no overlap between the training and the test data, and the mean result of five times is taken as the final performance. Suppose *TP*, *TN* represents the count of positive and negative samples predicted correctly, and *FP*, *FN* indicates the count of predicted incorrectly, these measures including Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spec.), Precision (Prec.), and Matthews Correlation Coefficient (MCC) are defined as:

$$\mathrm{Acc.} = \frac{TN + TP}{TN + TP + FN + FP} \quad (16)$$

$$\text{Sen.} = \frac{TP}{TP+FN} \tag{17}$$

$$\text{Spec.} = \frac{TN}{TN+FP} \tag{18}$$

$$\text{Prec.} = \frac{TP}{TP+FP} \tag{19}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{20}$$

To avoid the bias, the Receiver Operating Characteristic (ROC) curve, the Area Under the ROC curve (AUC) are adopted. In order to avoid the impact of special situations on the results such as imbalanced data sets and accurately and comprehensively reflect the quality of predictions, the Area Under the Precision-Recall curve (AUPR) are also evaluated and reported.

**Supplemental References**


GROVER, A. & LESKOVEC, J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016. ACM, 855-864.

PEROZZI, B., AL-RFOU, R. & SKIENA, S. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014. ACM, 701-710.

SHEN, J., ZHANG, J., LUO, X., ZHU, W., YU, K., CHEN, K., LI, Y. & JIANG, H. 2007. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences,* 104**,** 4337-4341.

VINCENT, P., LAROCHELLE, H., LAJOIE, I., BENGIO, Y. & MANZAGOL, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research,* 11**,** 3371-3408.

WANG, D., WANG, J., LU, M., SONG, F. & CUI, Q. 2010. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics,* 26**,** 1644-1650.

WEININGER, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences,* 28**,** 31-36.

WOLD, S., ESBENSEN, K. & GELADI, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems,* 2**,** 37-52.