BioNE: Integration of network embeddings for supervised learning

Poorya Parvizi[1*], Francisco Azuaje[2], Evropi Theodoratou[1,3], and Saturnino Luz[1*]

[1]Usher Institute, The University of Edinburgh, Edinburgh, EH16 4UX, United Kingdom
[2]Genomics England, London, EC1M 6BQ, United Kingdom
[3]Cancer Research UK Edinburgh Centre, Medical Research Council Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, EH4 2XR, United Kingdom
[*]Correspondence: poorya.parvizi@ed.ac.uk, s.luz@ed.ac.uk

Summary: A network embedding approach reduces the analysis complexity of large biological networks by converting them to low-dimensional vector representations (features/embeddings). These lower-dimensional vectors can then be used in machine learning prediction tasks with a wide range of applications in computational biology and bioinformatics. Several network embedding approaches have been proposed with different methods of generating vector representations. These network embedding approaches can be quite diverse in terms of data representation and implementation. Moreover, most were not originally developed for biological networks. Therefore comparing and assessing the performance of these diverse models in practice, in biological contexts, can be challenging. To facilitate such comparisons, we have developed the BioNE framework for integration of different embedding methods in prediction tasks. Using this framework one can easily assess, for instance, whether combined vector representations from multiple embedding methods offer complementary information with regards to the network features and thus better performance on prediction tasks. In this paper, we present the BioNE software suite for embedding integration, which applies network embedding methods following standardised network preparation steps, and integrates the vector representations achieved by these methods using three different techniques. BioNE enables selection of prediction models, oversampling methods, feature selection methods, cross-validation type and cross-validation parameters.

Availability and implementation: BioNE pipeline and detailed explanation of implementation is freely available on GitHub, at https://github.com/pooryaparvizi/BioNE

Keywords: Biological Networks, Graph Embedding, Network Embedding, Node Embedding, Link Prediction, Ensemble Learning, Supervised Learning

Introduction

The advancement in high-throughput technologies has resulted in a substantial increase in available data, providing opportunities to research and gain a deeper understanding of interactions within biological systems. This allows for the analysis and exploration of cells or organisms as systems where molecular parts act together in a dynamic way (1). Network biology analysis, which is based on graph theory, could provide a structure for integrating these high-throughput multi-omics data and investigating interactions within biological systems (2). Although, analysing large amounts of data within the network is valuable, analysis and interpretation results can be challenging using conventional statistical methods (3). Network embedding approaches can provide an effective way to overcome the complexity of large biological network analysis. Embedding approaches map nodes to low-dimensional vectors by preserving the properties of their higher dimensional counterparts. Lower dimensional embeddings are then used in downstream analysis such as supervised learning link prediction tasks.

Embedding methods have been developed and used in a wide range of domains, most notably natural language processing. While research exploring the application of these methods to biological networks is still in its early stages, there is considerable interest. These applications can be loosely divided into three categories; (1) *drug-related applications*, such as drug-target interactions (DTIs) (4–9), drug-disease interactions (10–12), drug side-effects (13, 14), drug-drug interactions (15–17), polypharmacy antagonistic effects (18, 19) and synergistic reactions in drug combination therapy (20); (2) *protein-related applications*, such as protein-protein interactions (PPIs) (21–24) and protein/gene disease interactions (25–31); and (3) *transcriptomics-related applications*, such as lncRNAs-diseases associations (32–35) and miRNA-disease associations (36–43) and many other applications (44–50).

Since network embedding methods were not originally developed for biological networks, their performance in obtaining different biological network features is yet to be established. Different network embedding approaches capture the network's structural properties using different methods; the focus can be on local or global properties (51–53). Biological networks are sparse and incomplete (54). Therefore, it is necessary to develop embedding models that take into account the sparsity and incompleteness of biological networks while also accounting for their local and global structural properties. As no single approach appears to handle these trade-offs satisfactorily, one may ask whether integrating network embeddings from different methods might provide richer feature representations, greater insight into the network, and better prediction performance when used in downstream analysis. To address this question we have developed the BioNE processing pipeline, which provides tools for the preparation of networks, application of different network embedding methods and integration of embeddings in different ways. To the best of our knowledge, this pipeline is the first set of tools to support comprehensive integration of network embeddings.

Implementation

The BioNE pipeline consists of three steps: network preparation, network embedding, and link prediction:

(a) Network preparation involves representing the input data in formats suitable for processing within the network embedding step. In this step, users are required to

convert adjacency matrices to edge list files. The user also has an option to set networks as directed or two or more edge lists can be combined to create heterogeneous networks.

(b) BioNE's network embedding step takes the prepared input and applies network embedding methods to learn low-dimensional vector representations for each node on the network. The following embedding methods are available within BioNE; LINE (55), GraRep (56), SDNE (57), HOPE (58), LaplacianEigenmaps (LAP) (59), node2vec (60), DeepWalk (61) and Graph Factorization (62). The user has options to treat the network as directed, weighted or set the vector representations size. The output is a space-delimited file that contains vector representations (features/embeddings) of the nodes.

(c) In the link prediction step, the user needs to provide the annotation file in order to define dependent and independent variables for link prediction tasks. For example, the annotation file should contain information such as, drug A and protein E do not show an association; drug B and protein E show an association. Therefore, the dependent variable of link drug A and protein E is 0 and for drug B and protein E is 1. The concatenation of the drug and protein embeddings are independent variables (see Figure 1). The user is also required to provide the list of embedding files generated by different embedding methods (produced in the embedding step), which the user wishes to integrate. The user can select the cross-validation method and parameters, an oversampling method (for imbalanced data), feature selection techniques and prediction models.

Three techniques, late (eq.1), early (eq.2) and mixed (eq.3) fusions are used to integrate different embeddings.

Late Fusion: For late fusion the resulting embeddings are fed to the classifier (machine learning model) separately. The classifier then calculates the prediction probabilities for each instance. The class with the highest sum of prediction probabilities for each instance is assigned as the prediction, that is:

$$c = \arg\max_{c_i \in C} \left(\sum_{e \in E} p(c_i | e, f) \right) \quad \textbf{(1)}$$

where C is the set of classes $\{0, 1\}$, E is the set of embeddings $\{e_{line}, e_{grarep}, e_{sdne}, e_{hope}, e_{lap}, e_{node2vec}, e_{deepwalk}, e_{gf}\}$ derived from different embedding methods $\{line, grarep, sdne, hope, lap, node2vec, deepwalk, gf\}$, and $p(c_i|e, f)$ is the prediction probability of class c for classifier f (e.g. support-vector machine, SVM) for data e.

Early fusion: Early fusion concatenates the embedding results before inserting them into the prediction model. In the case of the integration of grarep, SDNE and deepwalk embeddings:

$$M_e = e_{grarep} \oplus e_{sdne} \oplus e_{deepwalk}$$

$$c = \arg\max_{c_i \in C} p(c_i | M_e, f) \quad \textbf{(2)}$$

where M_e is a concatenation of embeddings. The classifier f (i.e. SVM) then estimates the conditional class probabilities $p(c_i|M_e, f)$ for the merged data M_e.

Mixed fusion: Mixed fusion merges data as in early fusion and then passes them on to different classifiers. The sum of prediction probabilities from different classifiers for each instance is calculated and the class with the highest sum is considered as the prediction:

$$c = \arg\max_{c_i \in C} \left(\sum_{f \in F} p(c_i | M_e, f) \right) \quad \textbf{(3)}$$

where M_e is a concatenation of embeddings of methods and F is the set of classifiers $\{Random\ Forest, SVM, Naive\ Bayes, XGBoost\}$ and $p(c_i|M_e, f)$ is the prediction probability of class c for classifier f for data M_e.

BioNE outputs the prediction performance of the link prediction in different metrics. In addition, BioNE also provides receiver operating characteristic (ROC) and precision-recall (PR) curves. ROC shows the true positive rate of a model's prediction plotted against its false positive rate as the classification threshold varies. The PR curve displays the trade-off between precision and recall for different classification thresholds. Both plots help evaluate the model by plotting performance trade-offs; ROC are the most common of the two, but PR curves can be useful for highly imbalanced classes, which is a common feature in many biological prediction tasks.

Advantages of BioNE

As mentioned above, the BioNE framework integrates embeddings from different embedding method, enabling the assessment of whether the combined embeddings offer complementary information with regards to the input network features and thus better performance on prediction tasks. In addition, BioNE provides toolsets to overcome the challenges in link prediction and machine learning analysis. In link prediction tasks, machine learning classifiers are used to classify the presence or absence of an interaction between two entities. In biological networks, there is often the problem of class imbalance, as the absence of interactions tend to overwhelmingly dominate the distribution. BioNE provides different oversampling techniques such as SMOTE (63) to overcome this challenge. Another oversampling technique available in BioNE is to equalize the number of positive and negative interactions.

Following the integration of embeddings, the total number of features increases and can lead to the "curse of dimensionality" which can cause substantial issues for most traditional machine learning algorithms. Insufficiency of training samples and redundancy among features is regarded as a significant issue in the supervised classification of hyperspectral

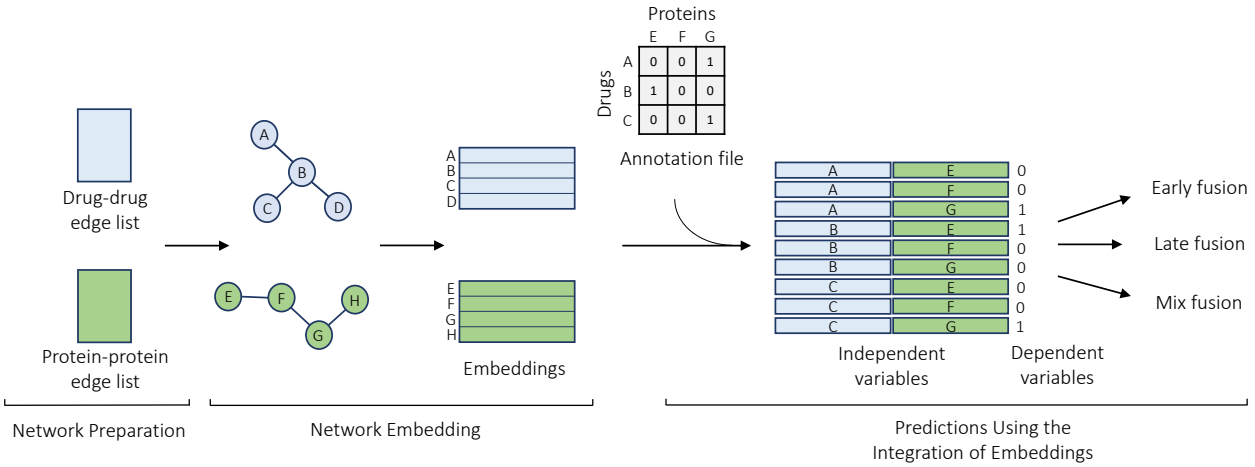

Fig. 1. Use case example of BioNE in the use of embedding methods in a drug-target interaction prediction task without the application of the fusion step. This consists of three parts: preparation of the networks, learning of vector representations using network embedding methods, and predictions using the learnt representations. In this prediction model, embeddings are independent variables and their interactions (i.e. 0 and 1) in the annotation file (i.e. known drug-target interactions) are dependent variables.

data. BioNE provides feature selection methods based on ANOVA and Mutual information (MI) to address this issue. In addition, BioNE provides different machine learning models and helps reduce the risk of over-fitting by providing two different cross-validation methods, namely, k-fold and stratified cross-validation. The performance of the prediction is evaluated using different metrics such as accuracy, precision, recall, specificity, F-scores and area under the ROC and PR curves. As mentioned previously, the PR curve, which mainly focuses on true positive cases, is particularly valuable when measuring model link prediction performance on imbalanced data.

Example Application of BioNE

As an example of the use of BioNE, the late fusion technique has been tested on the drug-target interaction (DTI) prediction task shown in Figure 1. To conduct this prediction, drug-drug interactions (708 drugs) and known DTIs were extracted from the Drug-Bank database (Version 3.0) (64). Protein-protein interactions (1493 proteins) were obtained from the Human Protein Reference Database (HPRD, release 9) (65). Embedding methods (LINE, GraRep, SDNE, HOPE, LAP, DeepWalk and GF) were applied to both drug-drug interactions and protein-protein interactions networks separately using default hyperparameters. Detailed explanation of parameters can be found on the Github[1] page. Vector representations of the drugs are derived from the embedding of drug-drug interaction networks, and vector representations of the proteins are derived from the embeddings of protein-protein interaction networks. The size of vector representations achieved from each network embedding method is 20. The embeddings of the proteins and drugs are then concatenated according to the annotation file that contains known DTIs. These concatenations are considered as predictors, and the absence or presence of associations between drugs and proteins are taken to be the values of the dependent variables, as in (66).

[1] https://github.com/pooryaparvizi/BioNE

The BioNE pipeline easily integrates different embeddings achieved from network embedding methods and uses the late fusion technique to test the performance of the predictions. For the late fusion step, a 10-fold cross-validation procedure without the application of oversampling and feature selection methods was conducted. In order to eliminate the problem raised due to imbalanced size between classes, the number negative associations is matched to the number of positive associations, as done in other studies (67). This is achieved by randomly (under)sampling the specified number of negative associations from the sample. In this task, the prediction model used was SVM with a radial basis function (RBF) kernel. In addition, we assessed several embeddings individually for comparison to BioNE's late fusion. Prediction performance of this method and other embedding methods is shown in Table 1. This table summarizes the performance of this prediction for each fold, reported values include the mean performance metrics. This table, along with ROC and PR curves are the outputs of the BioNE framework.

	Accuracy	Precision	Recall	F1	ROC-auc	PR-auc
fold1	0.80	0.77	0.78	0.78	0.87	0.88
fold2	0.80	0.79	0.76	0.77	0.87	0.85
fold3	0.72	0.75	0.70	0.73	0.83	0.86
fold4	0.83	0.85	0.81	0.83	0.88	0.90
fold5	0.83	0.90	0.79	0.84	0.89	0.93
fold6	0.82	0.86	0.78	0.82	0.91	0.93
fold7	0.81	0.87	0.75	0.80	0.90	0.91
fold8	0.77	0.75	0.74	0.74	0.85	0.86
fold9	0.77	0.83	0.70	0.76	0.85	0.86
fold10	0.76	0.81	0.71	0.76	0.86	0.89
Mean	0.79	0.82	0.75	0.78	0.87	0.89

Table 1. Cross validation results for late fusion, as implemented through the BioNE pipeline: different metrics to evaluate the performance of the late fusion in the drug-target interaction prediction task. Each line represents the performance in each fold of cross-validation and last line takes the mean of these metrics.

In addition, Table 2 compares the area under ROC and PR curves of other network embedding methods in the drug-target interaction prediction task.
Results show that the BioNE results outperform other network embedding methods when comparing area under ROC and PR curves. With area under the ROC curve of 0.87

	LINE	GraRep	SDNE	HOPE	LAP	DeepWalk	GF	BioNE
ROC	0.75	0.84	0.77	0.66	0.74	0.82	0.83	**0.87**
PR	0.76	0.86	0.77	0.64	0.75	0.82	0.84	**0.89**

Table 2. Mean area under the ROC and PR curves in different embedding methods in DTI prediction task. The column labelled BioNE shows the results of predictions using late fusion for combining different embedding methods.

BioNE outperformed network embedding methods GraRep, GF, DeepWalk, SDNE, LINE, LAP and HOPE by %3, %4, %5, %10, %12, %13 and %21 respectively. This demonstrates that, BioNE is a valuable tool set to easily integrate the embeddings to achieve more comprehensive knowledge of the network and to test their performance in prediction tasks.

Conclusions

As network embedding methods were not originally designed for biological networks, their performance in obtaining different biological network features is not straightforward. We believe, the integration of network embeddings can take into account the sparsity and incompleteness of biological networks and is also capable of hurdling the trade-off between local and global structure properties in network embedding methods.

Therefore, the main purpose of this framework is to enable researchers to easily reuse and combine well-known network embedding methods and test their performance individually and in combination on different link/association predictions. In addition, BioNE provides tool sets to overcome some of the challenges in link prediction and machine learning methods such as oversampling methods to overcome the imbalanced data challenges, feature selection methods to reduce the curse of dimensionality and many other tools. Although, we focused on well-known embedding methods, users can expand this framework by adding other de-novo network embedding methods such as graph convolutional networks (GCN) (68, 69), other feature selection and oversampling methods.

In addition, in the case where users wish to add other features unrelated to the networks to the prediction task, they can integrate these features to vector representations (embedding step's output) and then pass them as inputs to the prediction task. To the best of our knowledge, this pipeline is the first easy to use toolset to support comprehensive integration of network embeddings and test their performance in prediction task, and we intend to develop it further in cooperation with its user community.

ACKNOWLEDGEMENTS
The author P.P. would like to acknowledge Melisa Chuong for their help in proof reading during the preparation of this manuscript.

FUNDING
This work was supported by the University of Edinburgh's Global Research Scholarship and the Chancellor's fellowship awarded to PP via SL, and CRUK Career Development Fellowship [C31250/A22804] awarded to ET.

AVAILABILITY OF DATA AND MATERIALS
The datasets supporting the conclusions of this article, as well as BioNE pipeline and detailed explanation of implementation are freely available on GitHub, at https://github.com/pooryaparvizi/BioNE.

AVAILABILITY AND REQUIREMENTS OF BioNE
Project name: BioNE
Project home page: https://github.com/pooryaparvizi/BioNE
Archived version: https://doi.org/10.5281/zenodo.5500712
Operating system(s): Platform independent
Programming language: Python
Other requirements: Bash
License: GNU General Public License v3.0
Any restrictions to use by non-academics: No

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
No ethical approval or consent to participate required. All of the data used in this study are publicly available.

COMPETING FINANCIAL INTERESTS
The authors declare that they have no competing interests.

CONSENT FOR PUBLICATION
Not applicable.

AUTHOR CONTRIBUTIONS
P.P. designed, developed, implemented and wrote the manuscript; S.L, F.A. and E.T. supervised P.P. in this study and reviewed the development and implementation of the pipeline and edited the manuscript.

Bibliography

1. Maria V. Schneider and Sandra Orchard. Omics Technologies, Data and Bioinformatics Principles. *Methods in molecular biology (Clifton, N.J.)*, 719:3–30, 2011. ISSN 19406029. doi: 10.1007/978-1-61779-027-0_1.
2. Jingwen Yan, Shannon L. Risacher, Li Shen, and Andrew J. Saykin. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Briefings in Bioinformatics*, 19(6):1370–1381, 2017. ISSN 14774054. doi: 10.1093/bib/bbx066.
3. Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina*, 56(9):1–10, sep 2020. ISSN 16489144. doi: 10.3390/MEDICINA56090455.
4. Bo Ya Ji, Zhu Hong You, Han Jing Jiang, Zhen Hao Guo, and Kai Zheng. Prediction of drug-target interactions from multi-molecular network based on LINE network representation method. *Journal of Translational Medicine*, 18(1):1–11, 2020. ISSN 14795876. doi: 10.1186/s12967-020-02490-x.
5. Zhan Heng Chen, Zhu Hong You, Zhen Hao Guo, Hai Cheng Yi, Gong Xu Luo, and Yan Bin Wang. Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model. *Frontiers in Bioengineering and Biotechnology*, 8(June):1–9, 2020. ISSN 22964185. doi: 10.3389/fbioe.2020.00338.
6. Qi An and Liang Yu. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Briefings in bioinformatics*, 22(6):1–10, 2021. ISSN 14774054. doi: 10.1093/bib/bbab275.
7. Bo Wei Zhao, Zhu Hong You, Lun Hu, Zhen Hao Guo, Lei Wang, Zhan Heng Chen, and Leon Wong. A novel method to predict drug-target interactions based on large-scale graph representation learning. *Cancers*, 13(9):1–12, 2021. ISSN 20726694. doi: 10.3390/cancers13092111.
8. Yang Yue and Shan He. DTI-HeNE: a novel method for drug-target interaction prediction based on heterogeneous network embedding, 2021. ISSN 14712105.
9. Tianyi Zhao, Yang Hu, Linda R. Valsdottir, Tianyi Zang, and Jiajie Peng. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Briefings in Bioinformatics*, 22(2):2141–2150, 2021. ISSN 14774054. doi: 10.1093/bib/bbaa044.
10. Kai Yang, Xingzhong Zhao, David Waxman, and Xing Ming Zhao. Predicting drug-disease associations with heterogeneous network embedding. *Chaos*, 29(12), 2019. ISSN 10541500. doi: 10.1063/1.5121900.
11. Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. DeepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019. ISSN 14602059. doi: 10.1093/bioinformatics/btz418.
12. Renyi Zhou, Zhangli Lu, Huimin Luo, Ju Xiang, Min Zeng, and Min Li. NEDD: A network embedding based method for predicting drug-disease associations. *BMC Bioinformatics*, 21(Suppl 13):1–12, 2020. ISSN 14712105. doi: 10.1186/s12859-020-03682-4.
13. Baofang Hu, Hong Wang, Lutong Wang, and Weihua Yuan. Adverse drug reaction predictions using stacking deep heterogeneous information network embedding approach. *Molecules*, 23(12), 2018. ISSN 14203049. doi: 10.3390/molecules23123193.
14. Baofang Hu, Hong Wang, and Zhenmei Yu. Drug side-effect prediction via random walk on the signed heterogeneous drug network. *Molecules*, 24(20):1–15, 2019. ISSN 14203049. doi: 10.3390/molecules24203668.
15. Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, and Mohammad Sadoghi. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *Journal of Web Semantics*, 44:104–117, 2017. ISSN 15708268. doi: 10.1016/j.websem.2017.06.002.
16. Remzi Celebi, Huseyin Uyar, Erkan Yasar, Ozgur Gumus, Oguz Dikenelli, and Michel Dumontier. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinformatics*, 20(1):1–14, 2019. ISSN 14712105.
17. Yang Zhang, Yang Qiu, Yuxin Cui, Shichao Liu, and Wen Zhang. Predicting drug-drug interactions using multi-modal deep auto-encoders based network embedding and positive-

unlabeled learning. *Methods*, 179(June 2020):37–46, 2020. ISSN 10959130. doi: 10.1016/j.ymeth.2020.05.007.

18. S. S. Deepika, T. V. Geetha, and Deepika S.S. A meta-learning framework using representation learning to predict drug-drug interaction. *Journal of Biomedical Informatics*, 84 (January):136–147, 2018. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2018.06.015.
19. Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty294.
20. Benedek Rozemberczki, Anna Gogleva, Sebastian Nilsson, Gavin Edwards, Andriy Nikolov, and Eliseo Papa. *MOOMIN: Deep Molecular Omics Network for Anti-Cancer Drug Combination Therapy*, volume 1. Association for Computing Machinery, 2021.
21. Jiongmin Zhang, Man Zhu, and Ying Qian. protein2vec: Predicting Protein-Protein Interactions Based on LSTM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5963(c):1–1, 2020. ISSN 1545-5963. doi: 10.1109/tcbb.2020.3003941.
22. Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics*, 21(1): 1–16, 2020. ISSN 14712105. doi: 10.1186/s12859-020-03646-8.
23. Xiao Rui Su, Zhu Hong You, Lun Hu, Yu An Huang, Yi Wang, and Hai Cheng Yi. An Efficient Computational Model for Large-Scale Prediction of Protein–Protein Interactions Based on Accurate and Scalable Graph Embedding. *Frontiers in Genetics*, 12(February), 2021. ISSN 16648021. doi: 10.3389/fgene.2021.635451.
24. Elahe Nasiri, Kamal Berahmand, Mehrdad Rostami, and Mohammad Dabiri. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Computers in Biology and Medicine*, 137(August):104772, 2021. ISSN 18790534. doi: 10.1016/j.compbiomed.2021.104772.
25. Sezin Kircali Ata, Le Ou-Yang, Yuan Fang, Chee Keong Kwoh, Min Wu, Xiao Li Li, Ata S.K., Ou-Yang L., Fang Y., Kwoh C.-K., and Wu M. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC systems biology*, 12(Supplement 9):138, 2018. ISSN 1752-0509. doi: http://dx.doi.org/10.1186/s12918-018-0662-y.
26. Mona Alshahrani, Robert Hoehndorf, and Alshahrani M. Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics (Oxford, England)*, 34(17):i901–i907, 2018. ISSN 1367-4811. doi: http://dx.doi.org/10.1093/bioinformatics/bty559.
27. Lei Chen, Yu Hang Zhang, Guohua Huang, Xiaoyong Pan, Tao Huang, and Yu Dong Cai. Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms. *Gene Therapy*, 26(12):465–478, 2019. ISSN 14765462. doi: 10.1038/s41434-019-0099-y.
28. Xiaochan Wang, Yuchong Gong, Jing Yi, and Wen Zhang. Predicting gene-disease associations from the heterogeneous network using graph embedding. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, pages 504–511, 2019. doi: 10.1109/BIBM47256.2019.8983134.
29. Jiajie Peng, Jiaojiao Guan, and Xuequn Shang. Predicting Parkinson's disease genes based on node2vec and autoencoder. *Frontiers in Genetics*, 10(APR):1–6, 2019. ISSN 16648021. doi: 10.3389/fgene.2019.00226.
30. Haijie Liu, Jiaojiao Guan, He Li, Zhijie Bao, Qingmei Wang, Xun Luo, and Hansheng Xue. Predicting the Disease Genes of Multiple Sclerosis Based on Network Representation Learning. *Frontiers in Genetics*, 11(April):1–7, 2020. ISSN 16648021. doi: 10.3389/fgene.2020.00328.
31. Haijie Liu, Liping Hou, Shanhu Xu, He Li, Xiuju Chen, Juan Gao, Ziwen Wang, Bo Han, Xiaoli Liu, and Shu Wan. Discovering Cerebral Ischemic Stroke Associated Genes Based on Network Representation Learning. *Frontiers in Genetics*, 12(September):1–9, 2021. ISSN 16648021. doi: 10.3389/fgene.2021.728333.
32. Jianwei Li, Jianing Li, Mengfan Kong, Duanyang Wang, Kun Fu, and Jiangcheng Shi. SVDNVLDA: predicting lncRNA-disease associations by Singular Value Decomposition and node2vec. *BMC Bioinformatics*, 22(1):1–18, 2021. ISSN 14712105. doi: 10.1186/s12859-021-04457-1.
33. Qing Wen Wu, Jun Feng Xia, Jian Cheng Ni, and Chun Hou Zheng. GAERF: Predicting lncRNA-disease associations by graph auto-encoder and random forest. *Briefings in Bioinformatics*, 22(5):1–12, 2021. ISSN 14774054. doi: 10.1093/bib/bbaa391.
34. Qing Wen Wu, Rui Fen Cao, Junfeng Xia, Jian Cheng Ni, Chun Hou Zheng, and Yansen Su. Extra Trees Method for Predicting LncRNA-Disease Association Based on Multi-layer Graph Embedding Aggregation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5963(c):1–9, 2021. ISSN 15579964. doi: 10.1109/TCBB.2021.3113122.
35. Lei Deng, Wenkai Li, and Jingpu Zhang. LDAH2V: Exploring Meta-Paths across Multiple Networks for lncRNA-Disease Association Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(4):1572–1581, 2021. ISSN 15579964. doi: 10.1109/TCBB.2019.2946257.
36. Ya-Wei Niu, Guang-Hui Wang, Gui-Ying Yan, and Xing Chen. Integrating random walk and binary regression to identify novel miRNA-disease association. *BMC bioinformatics*, 20(1): 59, 2019. ISSN 1471-2105. doi: https://dx.doi.org/10.1186/s12859-019-2640-9.
37. Yuchong Gong, Yanqing Niu, Wen Zhang, and Xiaohong Li. A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinformatics*, 20(1):1–13, 2019. ISSN 14712105. doi: 10.1186/s12859-019-3063-3.
38. Bo Ya Ji, Zhu Hong You, Zhan Heng Chen, Leon Wong, and Hai Cheng Yi. NEMPD: A network embedding-based method for predicting miRNA-disease associations by preserving behavior and attribute information. *BMC Bioinformatics*, 21(1):1–17, 2020. ISSN 14712105. doi: 10.1186/s12859-020-03716-x.
39. Bo Ya Ji, Zhu Hong You, Li Cheng, Ji Ren Zhou, Daniyal Alghazzawi, and Li Ping Li. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Scientific Reports*, 10(1):1–12, 2020. ISSN 20452322. doi: 10.1038/s41598-020-63735-9.
40. Lei Zhang, Bailong Liu, Zhengwei Li, Xiaoyan Zhu, Zhizhen Liang, and Jiyong An. Predicting MiRNA-disease associations by multiple meta-paths fusion graph embedding model. *BMC Bioinformatics*, 21(1):1–19, 2020. ISSN 14712105. doi: 10.1186/s12859-020-03765-2.
41. Dong Ling Yu, Zu Guo Yu, Guo Sheng Han, Jinyan Li, and Vo Anh. Heterogeneous types of miRNA-disease associations stratified by multi-layer network embedding and prediction. *Biomedicines*, 9(9):1–14, 2021. ISSN 22279059. doi: 10.3390/biomedicines9091152.
42. Cunmei Ji, Yutian Wang, Jiancheng Ni, Chunhou Zheng, and Yansen Su. Predicting miRNA-Disease Associations Based on Heterogeneous Graph Attention Networks. *Frontiers in Genetics*, 12(August):1–12, 2021. ISSN 16648021. doi: 10.3389/fgene.2021.727744.
43. Hao Yuan Li, Hai Yan Chen, Lei Wang, Shen Jian Song, Zhu Hong You, Xin Yan, and Jin Qian Yu. A structural deep network embedding model for predicting associations between miRNA and disease based on molecular association network. *Scientific Reports*, 11 (1):1–13, 2021. ISSN 20452322. doi: 10.1038/s41598-021-91991-w.
44. Mona Alshahrani, Mohammad Asif Khan, Omar Maddouri, Akira R Kinjo, Nuria Queralt-Rosinach, and Robert Hoehndorf. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics (Oxford, England)*, 33(17):2723–2730, 2017. ISSN 1367-4811. doi: https://dx.doi.org/10.1093/bioinformatics/btx275.
45. Gamal Crichton, Yufan Guo, Sampo Pyysalo, and Anna Korhonen. Neural networks for link prediction in realistic biomedical graphs: A multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, 19(1):1–11, may 2018. ISSN 14712105. doi: 10.1186/S12859-018-2163-9/TABLES/7.
46. Zhen Hao Guo, Zhu Hong You, De Shuang Huang, Hai Cheng Yi, Zhan Heng Chen, and Yan Bin Wang. A learning based framework for diverse biomolecule relationship prediction in molecular association network. *Communications Biology*, 3(1):1–9, 2020. ISSN 23993642. doi: 10.1038/s42003-020-0858-8.
47. Leon Wong, Zhu Hong You, Zhen Hao Guo, Hai Cheng Yi, Zhan Heng Chen, and Mei Yuan Cao. MIPDH: A Novel Computational Model for Predicting microRNA-mRNA Interactions by DeepWalk on a Heterogeneous Network. *ACS Omega*, 5(28):17022–17032, 2020. ISSN 24701343. doi: 10.1021/acsomega.9b04195.
48. Zhen Hao Guo, Zhu Hong You, and Hai Cheng Yi. Integrative Construction and Analysis of Molecular Association Network in Human Cells by Fusing Node Attribute and Behavior Information. *Molecular Therapy - Nucleic Acids*, 19(March):498–506, 2020. ISSN 21622531. doi: 10.1016/j.omtn.2019.10.046.
49. Xiaoqi Wang, Yaning Yang, Kenli Li, Wentao Li, Fei Li, and Shaoliang Peng. BioERP: biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. *Bioinformatics*, 37(24):4793–4800, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab565.
50. Haitao Fu, Feng Huang, Xuan Liu, Yang Qiu, and Wen Zhang. MVGCN: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks. *Bioinformatics*, 38(2):426–434, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab651.
51. Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network Representation Learning: A Survey. *IEEE Transactions on Big Data*, 6(1):3–28, jun 2018. doi: 10.1109/TBDATA.2018.2850013.
52. Haochen Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. A Tutorial on Network Embeddings. *arxiv*, pages 1–23, 2018.
53. Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A Survey on Network Embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, may 2019. ISSN 15582191. doi: 10.1109/TKDE.2018.2849727.
54. Margaret G. Guo, Daniel N. Sosa, and Russ B. Altman. Challenges and opportunities in network-based solutions for biological questions. *Briefings in Bioinformatics*, 23(1):1–4, jan 2022. ISSN 14774054. doi: 10.1093/BIB/BBAB437.
55. Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale information network embedding. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, may 2015. doi: 10.1145/2736277.2741093.
56. Shaosheng Cao, Wei Lu, and Qiongkai Xu. GraRep: Learning graph representations with global structural information. *International Conference on Information and Knowledge Management, Proceedings*, 19-23-Oct-:891–900, oct 2015. doi: 10.1145/2806416.2806512.
57. Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1225–1234, aug 2016. doi: 10.1145/2939672.2939753.
58. Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1105–1114, aug 2016. doi: 10.1145/2939672.2939751.
59. Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.
60. Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:855–864, aug 2016. doi: 10.1145/2939672.2939754.
61. Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, aug 2014. doi: 10.1145/2623330.2623732.
62. Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, and Alexander J. Smola. Distributed large-scale natural graph factorization. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pages 37–47, may 2013. doi: 10.1145/2488388.2488393.
63. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, jun 2002. ISSN 1076-9757. doi: 10.1613/JAIR.953.
64. Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Database issue):D1035, jan 2011. ISSN 03051048. doi: 10.1093/NAR/GKQ1126.
65. T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee,

Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. I. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database—2009 update. *Nucleic Acids Research*, 37(Database issue):D767, 2009. ISSN 03051048. doi: 10.1093/NAR/GKN892.

66. Poorya Parvizi, Francisco Azuaje, Evropi Theodoratou, and Saturnino Luz. A Network-Based Embedding Method for Drug-Target Interaction Prediction. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020-July:5304–5307, jul 2020. doi: 10.1109/EMBC44109.2020.9176165.
67. Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8(1), dec 2017. ISSN 20411723. doi: 10.1038/s41467-017-00680-8.
68. Giulia Muzio, Leslie O'Bray, and Karsten Borgwardt. Biological network analysis with deep learning. *Briefings in Bioinformatics*, 22(2):1515–1530, 2021. ISSN 14774054. doi: 10.1093/bib/bbaa257.
69. Xiao Meng Zhang, Li Liang, Lin Liu, and Ming Jing Tang. Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics*, 12(July):1–22, 2021. ISSN 16648021. doi: 10.3389/fgene.2021.690049.