A deep learning approach to predict inter-omics interactions in multi-layer networks

Niloofar Borhani
Isfahan University of Technology
Jafar Ghaisari
Isfahan University of Technology
Maryam Abedi
Isfahan University of Medical Sciences
Marzieh Kamali
Isfahan University of Technology
Yousof Gheisari (ygheisari@med.mui.ac.ir)
Isfahan University of Medical Sciences

Research Article

Keywords: deep learning, integrative networks, datasets

Posted Date: June 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-595180/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on January 26th, 2022.
See the published version at https://doi.org/10.1186/s12859-022-04569-2.

A deep learning approach to predict inter-omics interactions in multi-layer networks

Niloofar Borhani[1], Jafar Ghaisari[1,*], Maryam Abedi[2,3], Marzieh Kamali[1], and Yousof Gheisari[2,3,*]

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
[2]Department of Genetics and Molecular Biology, Isfahan University of Medical Sciences, Isfahan, Iran
[3]Regenerative Medicine Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
[*]ghaisari@iut.ac.ir; ygheisari@med.mui.ac.ir

ABSTRACT

Despite enormous achievements in production of high throughput datasets, constructing comprehensive maps of interactions remains a major challenge. The lack of sufficient experimental evidence on interactions is more significant for heterogeneous molecular types. Hence, developing strategies to predict inter-omics connections is essential to construct holistic maps of disease. Here, Data Integration with Deep Learning (DIDL), a novel nonlinear deep learning method is proposed to predict inter-omics interactions. It consists of an encoder that automatically extracts features for biomolecules according to existing interactions, and a decoder that predicts novel interactions. The applicability of DIDL is assessed with different networks namely drug-target protein, transcription factor-DNA element, and miRNA-mRNA. Also, the validity of novel predictions is assessed by literature surveys. Furthermore, DIDL outperformed state-of-the-art methods. Area under the curve, and area under the precision-recall curve for all three networks were more than 0.85 and 0.83, respectively. DIDL has several advantages like automatic feature extraction from raw data, end-to-end training, and robustness to sparsity. In addition, tensor decomposition structure, predictions solely based on existing interactions and biochemical data independence makes DIDL applicable for a variety of biological networks. DIDL paves the way to understand the underlying mechanisms of complex disorders through constructing integrative networks.

Introduction

The recent emergence of high throughput technologies has allowed the generation of previously unbelievable amounts of big biological data. The speed of data generation has surpassed data analysis, providing biomedical scientists with tremendous datasets of a size that they have not been encountering before. Hence, big data analysis is a major challenge in modern biology. Although a variety of methods for omics data analysis have been developed in recent years, inter-omics data integration remains a major challenge. Indeed, it is now commonly believed that the description of biomedical phenomena cannot be reduced to alterations of a single type of biomolecules. Indeed, it is pivotal to consider not only the interactions between one layer of omics data but also complex inter-layer communications to identify the flow of biological information and generate a through holistic view of the underlying events. Such investigations have rarely been done due to the lack of appropriate computational techniques.

A number of methods have been developed for omics data integration in order to predict inter-omics interactions. Some of these methods rely on correlation of gene expressions[1,2]. However, it is limited to certain layers and some others such as genomics and epigenomics data cannot be incorporated. Network embedding, also known as network representation learning, has been recently proposed as a method to embed network nodes into a low-dimensional vector space named latent features, by capturing topological properties of networks and other side information. In another word, this method calculates the similarity between pairwise nodes to find a low-dimensional manifold structure hidden in the high-dimensional data[3,4]. One of the methods developed for interaction prediction based on network embedding is matrix factorization in which latent features are detected from network topology[5]. The Data Fusion by Matrix Factorization (DFMF)[6] is a method to predict direct and indirect interactions between heterogeneous nodes. However, these methods are not able to extract highly non-linear patterns from data. A more recent interaction prediction method, node2vec, learns low-dimensional representations for nodes of the network by maximizing the probability of the occurrence of subsequent nodes in random walks over a network. This method was applied for homogeneous[7] and heterogeneous interaction prediction[8].

Deep learning is a kind of machine learning technique that automatically extracts high level features from the raw data of very large, heterogeneous, high-dimensional data sets. Hence, the deep learning methods could be assumed a good fit in considering the complexity of big data in biology[9–11]. The advantage of deep learning has been exploited in network

embedding to find complex structural features and learn deep, highly nonlinear node representations[4]. The idea of combining matrix factorization and deep learning is known as deep matrix factorization. This method extracts representations with two deep neural networks (DNN) and calculates similarity of representations through a cosin function as a decoder that is not trainable. The deep matrix factorization is used for recommender systems and has been shown to be superior to traditional matrix factorization[12]. This strategy has recently been used in the prediction of Drug-Target interactions[13]. Taken together, most present studies are focused on only one type of biological network and need specific biochemical features or cannot handle big and sparse data of all types of heterogeneous multi-layer networks.

Tensor decomposition is a powerful tool for a variety of heterogeneous, sparse, and big data of multi-layer networks[14]. Here, due to the advantage of deep learning and tensor decomposition, it is attempted to develop the application of deep learning in big biological data integration through employing tensor decomposition by an end-to-end strategy in order to be applicable in multi-layer networks without lying on a specific biochemical feature. In this paper, Data Integration with Deep Learning (DIDL) method, is proposed for inter-omics interaction prediction. At first, a network embedding method that is based on deep learning and tensor decomposition techniques is applied. This method consists of an encoder with two DNNs, that extracts representation for biological entities and considers heterogeneity, and a tensor factorization decoder, that predicts the probability of interactions. Then, to demonstrate the applicability of this method, it is evaluated with three different biological data sets, drug-target protein, transcription factor (TF)-DNA element, and miRNA-mRNA. Overall, this study shows, for the first time, the ability to model multi-omics networks for interaction prediction and consider heterogeneity of biomolecules in different omics layers.

Method

Interactions between heterogeneous biomolecules are based on biological principles, for instance, a miRNA targets a group of genes which are functionally related[15,16] and a TF regulates a bundle of genes which incorporate a specific sequence in their sequence upstream[17]. Hence, the probability of interaction between two given nodes in two different network layers can be estimated based on known interactions between each of these two nodes with other elements in the opposite layer. Indeed, unknown interactions can be predicted by employing network topology characteristics instead of biochemical features. In this regard, DIDL can be known as a homolog of recommender systems or completion matrix task.

Consider a two-layer network in which two omics layers are connected by inter-omics interactions between heterogeneous biomolecules. If the number of first and second omics layer biomolecules are n_1 and n_2 respectively, the structure of network is represented by an adjacency matrix R_{12} of shape $n_1 \times n_2$ as follows:

$$R_{12}(i,j) = \begin{cases} 1 & \text{if there is interaction between } i_{th} \text{ node of the first and } j_{th} \text{ node of the second omics layer} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $1 \leq i \leq n_1$, $1 \leq j \leq n_2$ and $R_{12}(i,j) = 0$ corresponds to non-interaction or unknown interaction which means the existence of interaction has not been investigated yet. Although both of non-interactions and unknown interactions are zero in R_{12}, they are different. The DIDL method uses the information of inter-omics interactions to predict new interactions. Remarkably, it is not necessary to know homogeneous interactions in a specific omics layer.

This method has two main components (figure 1):

- An encoder: two DNNs operating on adjacency matrix and producing latent features for biomolecules of first and second omics layers
- A decoder: a tensor factorization model using these latent features to predict the probability of interactions.

The DIDL method seeks to find the best latent features for representing each biomolecule according to existing interactions. The following are the details of the network structure and model training.

Encoder of DIDL

The first part of method is encoder, which extracts the best latent feature for representing each biomolecule. For a given biomolecule, its latent feature capture information of all associated kinds of interactions. Two DNNs are proposed as an encoder that extracted high level features from adjacency matrix R_{12} for inter-omics interaction prediction. As biomolecules of first and second omics layers are heterogeneous and different from each other, one DNN is used for biomolecules of first omics layer and another DNN is used for biomolecules of second omics layer. To extract feature vectors of first omics layer biomolecules, DNN for first layer, DNN_1, takes as input rows of R_{12}, which are $n_2 \times 1$ vectors, and produces $k \times 1$ latent feature vectors of first omics layer biomolecules. As well to extract feature vectors of second omics layer biomolecules and considering heterogeneity of nodes, DNN for the second layer, DNN_2, takes as input columns of R_{12}, which are $n_1 \times 1$ vectors,

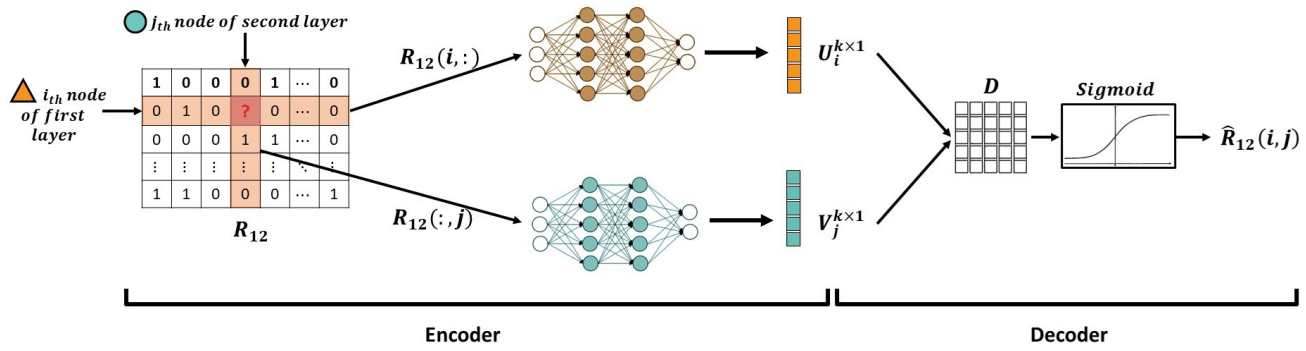

Figure 1. Overview of DIDL method framework. On the left, rows and columns of R_{12} are feed to an encoder which consists of two DNNs. The outputs of the encoder are latent features of the biomolecules. Finally, these latent features are transformed into a tensor factorization decoder and its output indicates the probability of interaction between heterogeneous biomolecules.

and produces $k \times 1$ feature vectors of second layer nodes. These features have less dimension in comparison with rows and columns of R_{12}, therefore $k < n_1, n_2$.

For the investigation of the interaction between i_{th} biomolecule of the first omics layer and j_{th} biomolecule of the second omics layer, pair (i_{th}, j_{th}), their latent feature vectors are calculated. The latent feature vector of i_{th} biomolecule of the first omics layer and the latent feature vector of j_{th} biomolecule of the second omics layer are represented by U_i and V_j, respectively. To find U_i, i_{th} row of R_{12} which consist of inter-omics interactions of i_{th} biomolecule of the first layer with biomolecules of the second omics layer is fed to DNN_1 and the output is U_i. As well to find V_j, j_{th} column of R_{12} which consist of inter-omics interactions of j_{th} biomolecules of second omics layer with biomolecules of the first omics layer is fed to DNN_2 and the output is V_j. So, the outputs of these DNNs take the following form:

$$\begin{aligned} U_i &= f_{DNN_1}(R_{12}(i,:)) \\ V_j &= f_{DNN_2}(R_{12}(:,j)) \end{aligned} \tag{2}$$

where f_{DNN_1}, f_{DNN_2}, $R_{12}(i,:)$ and $R_{12}(:,j)$ are total function of DNN_1, total function of DNN_2, i_{th} row of R_{12} and j_{th} column of R_{12}, respectively. Notably, the heterogeneity of biomolecules are considered by designing a separate DNN for each kind omics layer.

Decoder of DIDL

After calculating feature vectors by the encoder, decoder applies these feature vectors to investigate existence of interactions. The decoder objective is to calculate the probability of interacting heterogeneous biomolecules. Therefore, the decoder utilizes latent feature vectors U_i and V_j to assign a score that represents how likely it is that there is interaction between i_{th} biomolecules of the first omics layer and j_{th} biomolecules of second omics layer.

For this purpose, a decoder based on tensor factorization[18] is suggested. This decoder applies latent feature vectors U_i and V_j and predict score for interaction existence through an operation based on tensor factorization like this:

$$Score(i,j) = U_i^T D V_j \tag{3}$$

that $Score(i,j)$ is the score of interaction existence between pair (i_{th}, j_{th}) and D is a trainable parameter matrix of shape $k \times k$ that models interactions between heterogeneous biomolecules according to the latent feature vectors. As probability of interaction is between 0 and 1, a sigmoid function is applied on $score(i,j)$ to calculate probability of interactions (equation 4).

$$\hat{R}_{12}(i,j) = sigmoid(Score(i,j)) = \frac{1}{1+e^{-U_i^T D V_j}} \tag{4}$$

where $\hat{R}_{12}(i,j)$ is the probability of interaction between pair (i_{th}, j_{th}). Next, it is described how to train neural network weights and biases of the model in order to predict interactions.

Model training and optimizing

The encoder maps all biomolecules to latent feature vectors. Then, the decoder predicts probability of interactions. The encoder trains network structure in order to find the best feature vectors that could be used for inter-omics interaction prediction. In order to this aim, actuality $R_{12}(i,j)$ and prediction $\hat{R}_{12}(i,j)$ are compared to calculate error and try to decrease it. Thus, DNNs of encoder and tensor factorization decoder are trained by optimizing encoder parameters and matrix D using the cross-entropy loss, which takes the following form:

$$loss = \frac{1}{m} \sum_{(i,j) \in train\ set} R_{12}(i,j) \log \hat{R}_{12}(i,j) + (1 - R_{12}(i,j)) \log(1 - \hat{R}_{12}(i,j)) \tag{5}$$

that forces the model to obtain high probability for interactions (positive samples) and low probability to non-interactions (negative samples) and m is the number of samples. As encoder and decoder parameters are trained simultaneously, the DIDL become an end-to-end trainable model for inter-omics interaction prediction.

For the multi-layer network, there is an interactions list which consist of triples like (biomolecule in first omics layer, biomolecule in second omics layer, kind of interaction between biomolecules). The kind of interaction is one (positive sample), if there is interaction between the pair of biomolecules and the kind of interaction is zero (negative sample), if there is non-interaction between the pair of biomolecules. The train set in equation 5, must be included triples with positive and negative interactions. The adjacency matrix R_{12} includes information of the interactions between heterogeneous biomolecules, but does not include non-interacting biomolecules. Indeed, $R(i,j) = 0$, it indicates two stat that there is no interaction or the interaction is not yet discovered. This ambiguity is a challenge in deep learning methods which relies on both positive and negative interactions for training steps. In order to solve this challenge, we have applied negative sampling. It means that some pairs of biomolecules, which we are actually unaware of the interaction existence are chosen as negative examples[19].

According to equation 5, the DIDL considers the first-order proximity that means local pairwise proximity between two connected biomolecules[4] in biological networks. In addition, heterogeneous biomolecules which have high second-order proximity, share many common neighbors[4], i.e. the rows or columns of the adjacency matrix are similar to each other. Because these rows or columns are the DNNs inputs, biomolecules with high second-order proximity have similar encoder inputs. Consequently, latent features of biomolecules with high second-order proximity become similar. Therefore, the DIDL capture first-order and second-order proximity simultaneously to preserve the biological network structure.

Experimental setup

The DIDL method is implemented based on Tensorflow. The encoder is pairs of 4-layer neural network architectures with the Relu activation functions and two 64 and 32 hidden units in the first and second hidden layers, respectively. The latent feature vector dimension, k is 20 for all three data sets and the batch size are 32, 32, and 1024 for the miRNA-mRNA data set, the Drug-Target data set, and the TF-DNA data set, respectively. The model parameters are randomly initialized with a Gaussian distribution with zero mean and a standard deviation of 0.01. To optimize the model, the Adam optimizer[20] is utilized with a learning rate of 0.0001. In order to improve the generalization of the model for the prediction of unobserved inter-omics interactions the drop out and L_2 normalization are implemented and set to 0.5 and 0.08, respectively.

The performance of the DIDL is evaluated against the node2vec method. The node2vec learns the embeddings of the nodes in networks by applying the Skip-gram model to node sequences generated by a biased random walk. The window size, walk length, walks per vertex and dimensions have been set 10, 40, 10, and 25, respectively.

Results

The tremendous generation of omics data provides a unique opportunity to construct holistic maps for complex disorders. However, the construction of integrative networks is limited due to the lack of sufficient data about the interactions between heterogeneous biological entities. The emergence of machine learning methods allows to approach this problem. In this study, we have developed DIDL, which is a deep learning-based method for big biological data integration. This method consists of an encoder, that extracts representation for biomolecules, and a decoder, that predicts the probability of interactions. These representation vectors are extracted according to existing interactions and negative samples which are chosen randomly. To assess the performance of this method, it has been applied for three different heterogeneous biological datasets: drug-target protein, TF-DNA element, and miRNA-mRNA.

Drug-Target

Drug repositioning or repurposing is a promising approach in drug discovery. In recent years, a few strategies have been developed for drug repurposing. There are some disadvantages with the available methods such as demanding a considerable amount of biological information to be retrieved from literature or existing databases[21]. To show the capability of our model

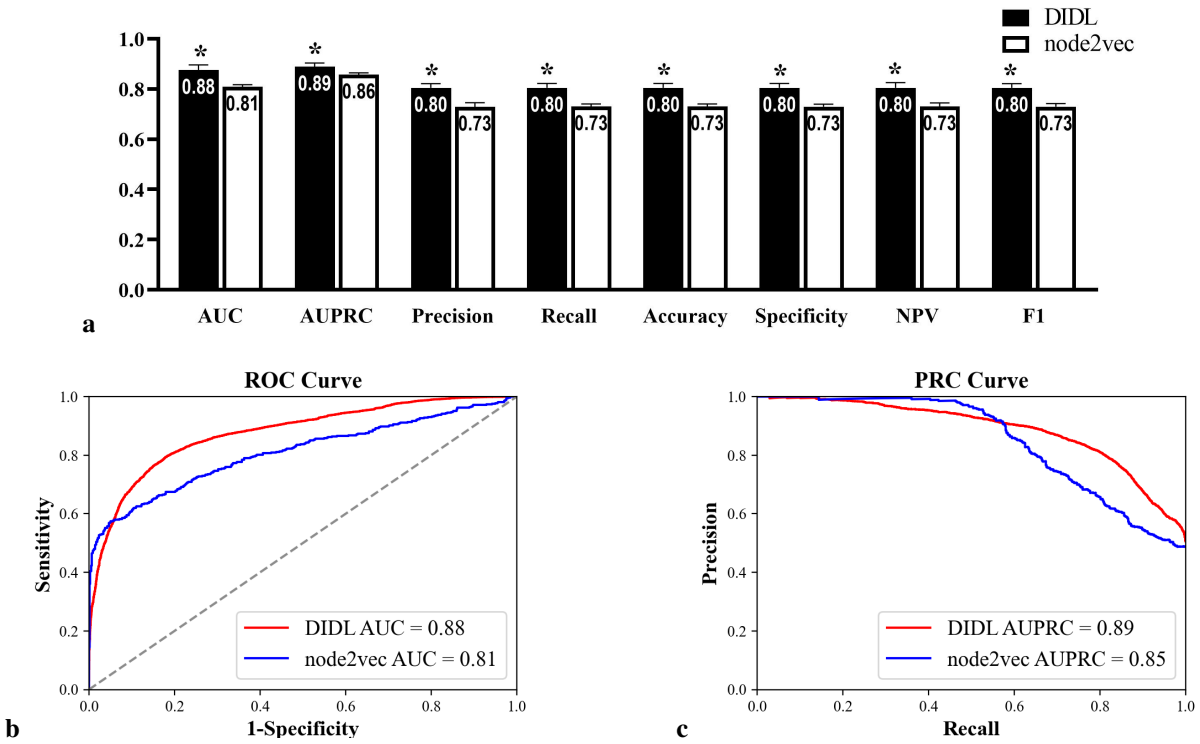

Figure 2. Evaluation of DIDL using 10 fold cross-validation and comparison with node2vec for Drug-Target prediction. (**a**): Results are evaluated in AUC, AUPRC, Precision, Recall, Accuracy, Specificity, negative predictive value (NPV), and F1 measures and also compared with node2vec. The * indicates P value<0.05. (**b**): Roc curves and (**c**) PRC curves for DIDL and node2vec.

for drug repositioning, we employed the DIDL method to predict new links between drugs and proteins. Small-molecule Drug-Target interactions known Drug-Target interactions were extracted from DrugBank database[22]. This data set consists of 1507 drugs, 1642 target proteins, and 6439 interactions. To evaluate the application of the DIDL method for drug target prediction, a 10-fold cross-validation was performed and different indices including area under receiver operating characteristic curve (AUC), area under precision-recall curve (AUPRC), precision, recall and accuracy measures, which are described previously[23], are obtained for DIDL method. The performance of the DIDL was further assessed using comparison with node2vec[7] which is a method based on topology to predict interactions. The performance of DIDL was significantly better than node2vec as revealed by T-test analysis (P value<0.05, figure 2).

TF-DNA

Transcriptional regulation of gene expression is a result of the interactions of TFs to specific DNA sequence elements named transcription factor binding sites (TFBSs) which is a critical step to control cell behaviors. However, the current knowledge about these interactions is preliminary. Current algorithms use data derived from chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) or sequencing (ChIP-Seq) techniques or apply a combination of in-silico sequence motif detection and experimental data for prediction of TF-TFBSs interactions. However, their performances are limited due to insufficient data[24]. To assess the performance of the DIDL on the link prediction between TFs and TFBSs, known human TF-TFBSs experimental data were extracted from the Enrichr database using ChEA 2016[25]. This dataset consists of 175 TFs, 35116 genes, and 407245 interactions. This data on known TF-DNA interactions was exploited by DIDL to predict novel interactions. A 10-fold cross-validation scheme is used for this dataset and performance indices are measured. Furthermore, the node2vec is also applied for this dataset and the AUC, AUPRC, precision, recall and accuracy measures of the node2vec for this data set are obtained. Notably, DIDL outperformed node2vec according to all indices and revealed by T-test analysis (P value<0.05, figure 3).

miRNA-mRNA

The complexity of the RNA world has been increasingly appreciated in recent decades. miRNA is a key regulator of a variety of cellular processes and identification or prediction of its interaction with mRNA is yet a main challenge. Despite

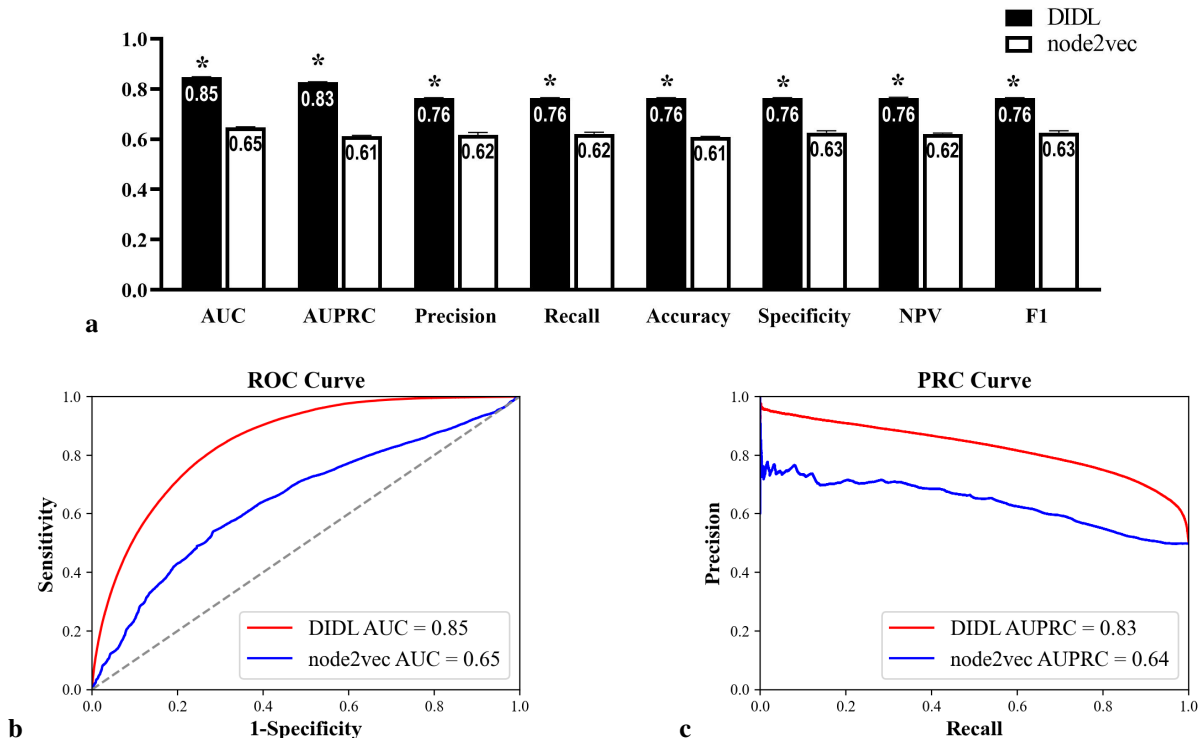

Figure 3. Evaluation of DIDL using 10 fold cross-validation and comparison with node2vec for TF-DNA prediction. (**a**): Results are evaluated in AUC, AUPRC, Precision, Recall, Accuracy, Specificity, NPV, and F1 measures and also compared with node2vec. The * indicates P value<0.05. (**b**): Roc curves and (**c**) PRC curves for DIDL and node2vec.

huge efforts, current tools still have suboptimal performance and even the best available algorithms have low accuracy and sensitivity[24,26]. In order to assess the performance of DIDL in miRNA target prediction, the experimentally validated human miRNA-mRNA interactions were retrieved from miRTarBase 7.0 and 6488 interactions with strong evidence for 704 miRNAs and 2524 mRNAs were chosen. Next, DIDL was employed to predict further interactions which revealed to have a good performance as all measures are equal to or greater than 0.8. Remarkably, the DIDL method also outperforms node2vec as revealed by T-test analysis (P value < 0.05, figure 4).

In addition, this method is compared with some state-of-the-art methods. The miRAW dataset has been harvested from Pla et al study[11] to compare DIDL with TargetScan (conserved)[27], miRAW (7-2:10 AE)[11] and DIANA microT[28]. This dataset consists of 449 miRNA, 6318 mRNA, 33142 positive samples, and 32248 negative samples. In this dataset, non-functional interactions have been assumed as negative samples. The comparison of the DIDL method with the existing algorithms indicates that the presented method has improved performance over other state-of-the-art target prediction methods (Figure 5). The measures of TargetScan, miRAW and DIANA microT for miRAW data set are harvested from Pla et al study[11].

To further assess our method, we have performed a literature-based evaluation of novel predicted interactions. The model makes a prediction for every heterogeneous pair of miRNA and mRNA. Then, miRNA-mRNA interactions were ranked based on probability of interaction. Interestingly, seven out of the ten top interactions have recently been confirmed in experimental investigations which are not yet incorporated in miRTarBase (Table 1). These validation strategies underscore the applicability of DIDL for miRNA target prediction.

Impact of network sparsity

Big biomedical data are often high-dimensional but sparse[39]. As the presented method is based on adjacency matrix of the biological networks, sparsity of the adjacency matrix is potentially an important factor in modeling performance. Hence, to assess the robustness of the method performance to network sparsity, 10% of the interactions were held out as the test set and then the sparsity of the remaining network was gradually increased by random removal of a portion of the remaining interactions in the train set. As expected by increasing network sparsity, the performance of the model decreases. However, by removing until around 50% of interactions, the model performance is robustly acceptable, especially for miRNA-mRNA and Drug-Target data sets (figure 6).

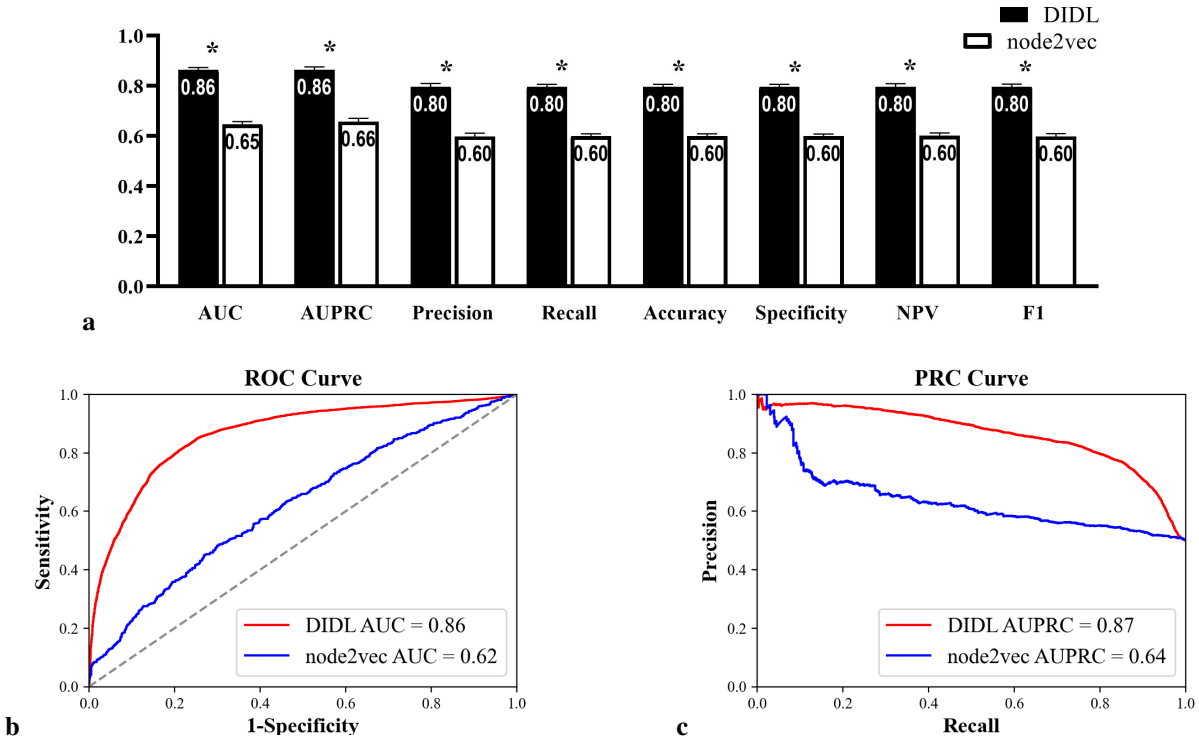

Figure 4. Evaluation of DIDL using 10 fold cross-validation and comparison with node2vec for miRNA-mRNA prediction. (**a**): Results are evaluated in AUC, AUPRC, Precision, Recall, Accuracy, Specificity, NPV, and F1 measures and also compared with node2vec. The * indicates P value<0.05. (**b**): Roc curves and (**c**) PRC curves for DIDL and node2vec.

Discussion

In order to achieve holistic views towards the complex mechanisms of physiological or pathological phenomena, it is imperative to construct multi-layer networks that consider the interactions of heterogeneous biomolecules. This study was aimed at developing a highly nonlinear mathematical data integration method based on deep learning for interaction prediction between any two layers of biological networks through considering known interactions. The encoder and decoder are simultaneously trained according to rows and columns of adjacent matrix of network interactions. The DIDL functionality was assessed for interaction prediction in Drug-Target, TF-DNA, and miRNA-mRNA networks and compared with alternative methods. Also, the validity of predictions was assessed by literature surveys.

An advantage of DIDL is that it is trained solely with known network interactions without any reliance on biological features of the nodes. This makes it easily applicable for a variety of biological networks especially in cases that sufficient knowledge on biological properties of interacting molecules or the mechanisms of interactions is not available. For instance, it is not yet comprehensively understood that how miRNAs select their mRNA targets. This makes the output of available prediction algorithms divergent as they are developed based on different assumptions[40]. Contrary, DIDL does not depend on such biological assumptions and operates on the basis of already experimentally known miRNA-mRNA interactions to predict further interactions. This method revealed to outperform even the best available prediction algorithms such as TargetScan, miRAW, and DIANA microT.

Large scale investigations on interactions between biomolecules including proteins have just recently begun and a majority of interactions are possibly yet undiscovered. Hence, considering the dependency of DIDL to recognized interactions, we were interested to know how robust this method is to network sparsity. We observed that DIDL has an acceptable function after removal of a considerable fraction of known interactions in train sets. This suggests that even in the current situation where intra- and inter-layer connections are not completely understood, DIDL can reliably be exploited.

Another advantage of the proposed method is that the process of feature selection and network representation is automatic. Although the logic of the method for predicting new interactions is the previous interactions, the tendency of interaction between neighboring nodes can vary depending on the network type. For example, in PPI network, the probability of interaction between two proteins sharing many common neighbors is actually low[41]. On the contrary, in Gene-Disease network, genes causing the same or similar diseases tend to interact with one another in the PPI network[42]. Therefore, manual feature

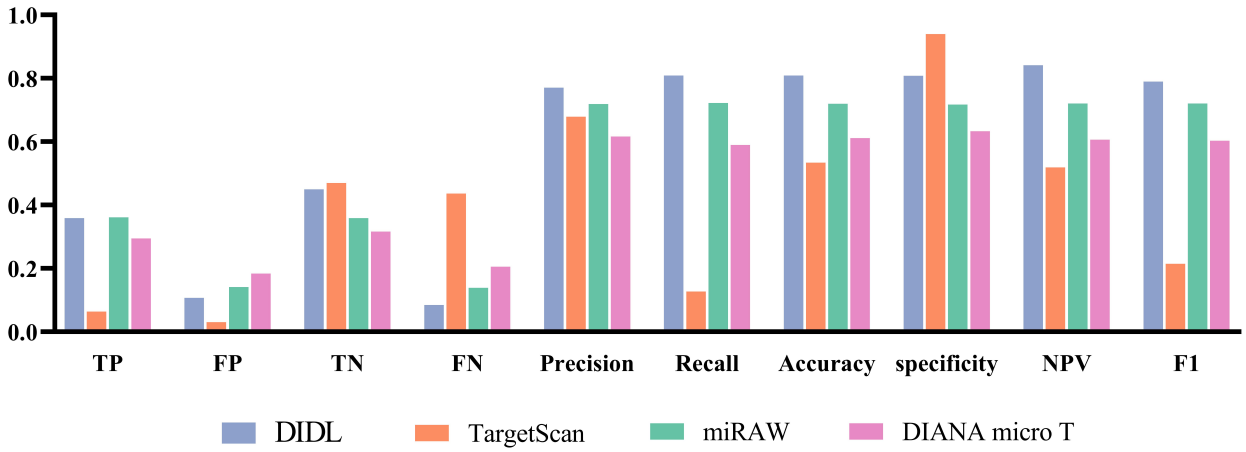

Figure 5. Comparison of predictive performance of the DIDL and miRNA target prediction methods. The measures of TargetScan, miRAW and DIANA microT for miRAW data set are harvested from Pla et al study[11].

Rank	miRNA	mRNA	Probability	Evidence
1	hsa-miR-15b-5p	ZEB1	0.9974	[29]
2	hsa-miR-34c-5p	EZH2	0.9973	[30]
3	hsa-miR-15b-5p	EZH2	0.9973	[31]
4	hsa-miR-34c-5p	ZEB1	0.9973	[32]
5	hsa-miR-34c-5p	TGFBR2	0.9972	*hsa-miR-34a, 34b[33]
6	hsa-miR-30c-5p	EZH2	0.9972	*hsa-miR-30d[34]
7	hsa-miR-15b-5p	RUNX2	0.9972	[35]
8	hsa-miR-15b-5p	TGFBR2	0.9971	[36]
9	hsa-miR-183-5p	SIRT1	0.9971	[37]
10	hsa-miR-34c-5p	FOXO1	0.9971	*hsa-mirR-34a, 34b[38]

Table 1. Novel miRNA target predictions with the highest probability scores by the DIDL. For each prediction, k is its rank in the ranked list of all predictions and literature evidence supporting the existence of the predicted interaction. The Pairs which are marked by * show that there is an interaction with other miRNA family member.

extraction is not a good choice especially when the network behavior is not known.

The DIDL is a novel auto encoder architecture that is capable of learning a joint representation of both first-order proximity and second-order proximity. This architecture is efficiently trained end-to-end in a single learning stage to simultaneously perform node representation and link prediction. Therefore, the decoder and encoder parameters jointly get optimized. Recent research indicates that with end-to-end learning, modeling graph-structured data can be considerably enhanced[43,44]. This can at least partly describe the superiority of DIDL performance compared to node2vec.

Taken together, we have here proposed DIDL, a novel method based on deep learning for omics data integration. It can be applied to construct multi-layer networks and generate comprehensive maps of the underlying mechanisms of complex disorders.

References

1. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, 418–429 (World Scientific, 1999).

2. Lê Cao, K.-A., González, I. & Déjean, S. integromics: an r package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855–2856 (2009).

3. Yue, X. *et al.* Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* **36**, 1241–1251 (2020).

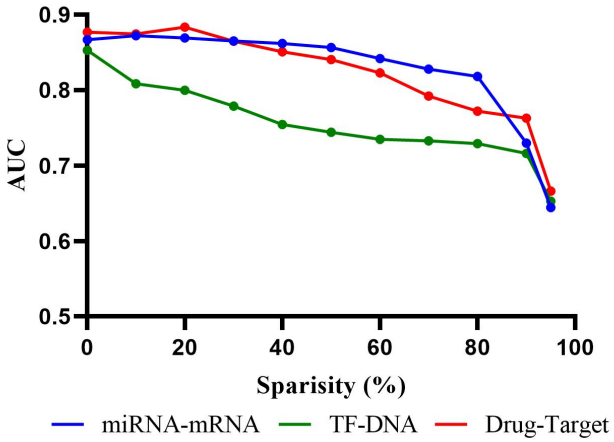

Figure 6. Impact of network sparsity

4. Zhang, D., Yin, J., Zhu, X. & Zhang, C. Network representation learning: A survey. *IEEE transactions on Big Data* (2018).

5. Menon, A. K. & Elkan, C. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, 437–452 (Springer, 2011).

6. Žitnik, M. & Zupan, B. Data fusion by matrix factorization. *IEEE transactions on pattern analysis machine intelligence* **37**, 41–53 (2014).

7. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864 (2016).

8. Crichton, G., Guo, Y., Pyysalo, S. & Korhonen, A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC bioinformatics* **19**, 176 (2018).

9. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. systems biology* **12**, 878 (2016).

10. Mamoshina, P., Vieira, A., Putin, E. & Zhavoronkov, A. Applications of deep learning in biomedicine. *Mol. pharmaceutics* **13**, 1445–1454 (2016).

11. Pla, A., Zhong, X. & Rayner, S. miraw: A deep learning-based approach to predict microrna targets by analyzing whole microrna transcripts. *PLoS computational biology* **14**, e1006185 (2018).

12. Xue, H.-J., Dai, X., Zhang, J., Huang, S. & Chen, J. Deep matrix factorization models for recommender systems. In *IJCAI*, vol. 17, 3203–3209 (Melbourne, Australia, 2017).

13. Manoochehri, H. E. & Nourani, M. Predicting drug-target interaction using deep matrix factorization. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–4 (IEEE, 2018).

14. Papalexakis, E. E., Faloutsos, C. & Sidiropoulos, N. D. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intell. Syst. Technol. (TIST)* **8**, 1–44 (2016).

15. Tseng, C.-W., Lin, C.-C., Chen, C.-N., Huang, H.-C. & Juan, H.-F. Integrative network analysis reveals active micrornas and their functions in gastric cancer. *BMC systems biology* **5**, 99 (2011).

16. Krishnan, K. *et al.* Microrna-182-5p targets a network of genes involved in dna repair. *Rna* **19**, 230–242 (2013).

17. Lis, M. & Walther, D. The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? *BMC genomics* **17**, 185 (2016).

18. Nickel, M., Tresp, V. & Kriegel, H.-P. A three-way model for collective learning on multi-relational data. In *Icml* (2011).

19. Demeester, T., Rocktäschel, T. & Riedel, S. Lifted rule injection for relation embeddings. *arXiv preprint arXiv:1606.08359* (2016).

20. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

21. Xue, H., Li, J., Xie, H. & Wang, Y. Review of drug repositioning approaches and resources. *Int. journal biological sciences* **14**, 1232 (2018).

22. Wishart, D. S. *et al.* Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34**, D668–D672 (2006).

23. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240 (2006).

24. Roopra, A. Magic: A tool for predicting transcription factors and cofactors driving gene sets using encode data. *PLoS computational biology* **16**, e1007800 (2020).

25. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016).

26. Quillet, A. *et al.* Improving bioinformatics prediction of microrna targets by ranks aggregation. *Front. Genet.* **10**, 1330 (2020).

27. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microrna target sites in mammalian mrnas. *elife* **4**, e05005 (2015).

28. Vlachos, I. S. *et al.* Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. *Nucleic acids research* **43**, D153–D159 (2015).

29. Zhang, W.-L., Zhang, J.-H., Wu, X.-Z., Yan, T. & Lv, W. mir-15b promotes epithelial-mesenchymal transition by inhibiting smurf2 in pancreatic cancer. *Int. journal oncology* **47**, 1043–1053 (2015).

30. Yu, Z. *et al.* Functional analysis of mir-34c as a putative tumor suppressor in high-grade serous ovarian cancer. *Biol. reproduction* **91**, 113–1 (2014).

31. Liu, C. *et al.* Screening of novel mirnas targeting ezh2 3'untranslated region using lentivirus mirnas library and their expressions in breast cancer cells and tissues. *Nan Fang yi ke da xue xue bao= J. South. Med. Univ.* **34**, 368–372 (2014).

32. Bissey, P.-A. *et al.* Mir-34c downregulation leads to sox4 overexpression and cisplatin resistance in nasopharyngeal carcinoma. (2020).

33. Ma, Z.-L. *et al.* Microrna-34a inhibits the proliferation and promotes the apoptosis of non-small cell lung cancer h1299 cell line by targeting tgfβr2. *Tumor Biol.* **36**, 2481–2490 (2015).

34. Yin, H. *et al.* Ezh2-mediated epigenetic silencing of mir-29/mir-30 targets loxl4 and contributes to tumorigenesis, metastasis, and immune microenvironment remodeling in breast cancer. *Theranostics* **10**, 8494 (2020).

35. Vimalraj, S., Partridge, N. C. & Selvamurugan, N. A positive role of microrna-15b on regulation of osteoblast differentiation. *J. cellular physiology* **229**, 1236–1244 (2014).

36. Tijsen, A. J. *et al.* The microrna-15 family inhibits the tgfβ-pathway in the heart. *Cardiovasc. research* **104**, 61–71 (2014).

37. Li, H. *et al.* Microrna-183 affects the development of gastric cancer by regulating autophagy via malat1-mir-183-sirt1 axis and pi3k/akt/mtor signals. *Artif. cells, nanomedicine, biotechnology* **47**, 3163–3171 (2019).

38. Li, C. *et al.* Micrornas as regulators and mediators of forkhead box transcription factors function in human cancers. *Oncotarget* **8**, 12433 (2017).

39. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).

40. Roberts, J. T. & Borchert, G. M. Computational prediction of microrna target genes, target prediction databases, and web resources. *Bioinforma. MicroRNA Res.* 109–122 (2017).

41. Zhang, M. & Chen, Y. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, 5165–5175 (2018).

42. Guo, X. *et al.* A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PloS one* **6**, e24171 (2011).

43. Tran, P. V. Learning to make predictions on graphs with autoencoders. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 237–245 (IEEE, 2018).

44. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. neural information processing systems* **29**, 3844–3852 (2016).

Author contributions

N.B., J.G., M.K., and Y.G. conceptualized the main idea, N.B. and M.A. harvested and analyzed the data. N.B., J.G., and M.K. proposed the method. N.B. performed simulations. N.B., J.G., and Y.G. contributed to the interpretation of the results. N.B. and M.A. drafted the manuscript and M.K., J.G., and Y.G. critically revised it. All authors approved the final draft and agreed to be responsible for the integrity of the entire work.

Competing interests

The authors declare no competing interests.