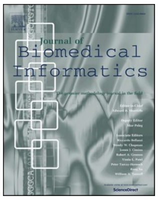
Journal of Biomedical Informatics 119 (2021) 103838

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Learning Drug-Disease-Target Embedding (DDTE) from knowledge graphs to inform drug repurposing hypotheses

Changsung Moon [a], Chunming Jin [b], Xialan Dong [b], Saad Abrar [a], Weifan Zheng [b,c,*], Rada Y. Chirkova [a,*], Alexander Tropsha [c,*]

[a] Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
[b] BRITE Institute and Department of Pharmaceutical Sciences, College of Health and Sciences, North Carolina Central University, Durham, NC 27707, USA
[c] UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, NC 27599, USA

ARTICLE INFO

Keywords:
Data mining
Graph embedding
Knowledge graph
Drug repurposing

ABSTRACT

We aimed to develop and validate a new graph embedding algorithm for embedding drug-disease-target networks to generate novel drug repurposing hypotheses. Our model denotes drugs, diseases and targets as subjects, predicates and objects, respectively. Each entity is represented by a multidimensional vector and the predicate is regarded as a translation vector from a subject to an object vectors. These vectors are optimized so that when a *subject-predicate-object* triple represents a known drug-disease-target relationship, the summed vector between the subject and the predicate is to be close to that of the object; otherwise, the summed vector is distant from the object. The DTINet dataset was utilized to test this algorithm and discover unknown links between drugs and diseases. In cross-validation experiments, this new algorithm outperformed the original DTINet model. The MRR (Mean Reciprocal Rank) values of our models were around 0.80 while those of the original model were about 0.70. In addition, we have identified and verified several pairs of new therapeutic relations as well as adverse effect relations that were not recorded in the original DTINet dataset. This approach showed excellent performance, and the predicted drug-disease and drug-side-effect relationships were found to be consistent with literature reports. This novel method can be used to analyze diverse types of emerging biomedical and healthcare-related knowledge graphs (KG).

1. Introduction

Drug repurposing (also known as drug repositioning) is the process of applying a known drug with specific therapeutic indication to another disease [1]. It has advantages over traditional drug discovery methods in that it can significantly reduce the cost and time required for drug development since the known drugs have already demonstrated safety in humans. Recently, we have conducted a bibliometric review of drug repurposing [2] by examining over 25 million articles in PubMed. We found that over 60% of the ~35,000 drugs or drug candidates identified in the study have been tested in more than one disease. Close to 200 drugs have been tested in over 300 diseases each. Some efforts have yielded unexpectedly good results for novel therapeutic targets. Thus, drug repurposing is a very powerful strategy in starting new drug discovery programs especially for rare or understudied diseases.

Various experimental and computational techniques have been employed for drug repurposing. One common approach is the experimental high throughput screening (HTS) of FDA-approved drugs, e.g., those from Prestwick Chemical Library® (Prestwick Chemical, Illkirch-Graffenstaden, France) against a novel biological target of interest that is considered relevant to the new disease. Conversely, computer-aided drug discovery/design (CADD) techniques have been employed to conduct virtual screening of known drugs or drug candidates. For example, if the 3D X-ray structure is available for the target implicated in the disease under investigation, docking of existing drugs to the target structure can be used [3]; alternatively, if a new target is identified for a disease and it is similar to the targets of known drugs, these known drugs can then be experimentally tested for use against the new disease. The target similarity can be based on protein sequences or 3D structure alignment [4]. Ligand-based similarity search as well as QSAR (Quantitative Structure-Activity Relationship) modeling were also reported to be of value in identifying FDA-approved drugs for repurposing [5,6].

* Corresponding authors at: BRITE Institute and Department of Pharmaceutical Sciences, College of Health and Sciences, North Carolina Central University, Durham, NC 27707, USA (W. Zheng).

E-mail addresses: wzheng@nccu.edu (W. Zheng), rychirko@ncsu.edu (R.Y. Chirkova), alex_tropsha@unc.edu (A. Tropsha).

https://doi.org/10.1016/j.jbi.2021.103838
Received 17 December 2020; Received in revised form 10 May 2021; Accepted 8 June 2021
Available online 11 June 2021

Another systematic strategy for drug repurposing is based on comprehensive analysis of structured chemical biology databases with known relationships among drugs, targets and diseases. Examples of these databases are DrugBank [7], Chem2Bio2RDF [8], Pharos [9] and ROBOKOP [10]. These databases are either stored as knowledge graphs (KG) [10] or as semantic triples [8] that can be systematically explored to discover hidden relationships and predict new ones. Graph embedding technologies such as Node2Vec [11] and Graph2Vec [12] have been developed in recent years and some of them have been employed in biomedical research [13]. Most of these technologies were developed to handle homogeneous networks and have limitations to be extended to heterogeneous network analysis. For example, a support vector machine (SVM) was applied [14] for finding new interactions in drug-target networks on a bipartite graph. Other representative approaches that integrate heterogeneous network information for drug repurposing can be found in [15–18]. These methods base their predictions on established drug-drug similarity, target-target similarity as well as known drug-target associations. Technically, these are not node/graph embedding methods in that they do not derive node embedding from the network itself; rather, most methods are based on structures of the drugs, and sequence-based embedding of the targets.

DTINet, a unique method for drug repurposing based on integrated heterogeneous network data, has been recently published [19]. It integrates a variety of drug-related information to construct a heterogeneous network, and then employs a compact feature learning algorithm to obtain a low-dimensional vector representation of nodes (drugs and proteins). It utilizes a set of known drug-target associations as the reference to find the best projection from drug space onto the protein space, such that the projected feature vectors of drugs are close to the feature vectors of their known targets. During prediction, DTINet infers new drug-target relationships for a drug by sorting its target candidates based on their proximity to the projected feature vector of this drug in the projected space. The predicted drug–target links can be further analyzed and experimentally validated. The authors have compared DTINet with other contemporary methods for drug-target interactions. In a ten-fold cross-validation, performance of each method was assessed; and DTINet has performed better than other methods based on the chosen metrics. Several predicted drug-target interaction pairs have been experimentally tested and validated.

The success of DTINet in modeling and predicting new drug-target interactions has inspired our work in developing a graph embedding algorithm for drug repurposing. For this purpose, we have utilized the basic framework of TransE [20] and adapted it to analyzing drug-disease-target network data. Here, entity relationships are represented as vector translations in the embedding space: if a relationship is true, the embedding of the object entity should be in close proximity to the embedding of the subject entity plus the relationship vector. It takes the heterogeneous network (i.e. knowledge graph) data, generates node embedding and conducts link prediction in a self-consistent fashion without the need to generate drug embedding and target embedding prior to learning the mapping between the drug space and the target space in separate steps [19]. Thus, the embedding in our method is customized to the link prediction itself, and affords the potential to improve the predictive power of this approach.

The aims of this paper are four-fold: (1) devise and describe a new algorithm for embedding knowledge graphs that capture drug-disease-target relationships; (2) assess the performance of the new algorithm compared to previously published results; (3) employ the new embedding models to predict unknown relationships among drugs and diseases; and finally (4) validate the predictions via retrospective literature and web search to determine if the new relationships predicted by the model constitute viable hypotheses for drug repurposing or adverse side effects.

2. Materials and methods

In Fig. 1, we have outlined the overall workflow of this study. First, it extracts true triples from the DTINet dataset. By definition, in a true triple all three associations among a drug, a disease and a target are known in the DTINet dataset. Missing any one association would render the triple an unknown triple. Second, the DDTE algorithm would generate embedding vectors for the drugs, the targets and the diseases so that overall the vector of a drug plus the vector of the associated disease is in proximity of the vector of the target. Finally, the output of DDTE includes the optimized embedding vectors for the drugs, the diseases and the targets as well as a set of predicted triples ranked according to their predicted scores. Details are described below.

2.1. Dataset

The DTINet data was originally from DrugBank [7], CTD [21], HPRD [22], and SIDER [23], where the drug nodes were extracted from DrugBank [7], the protein/target nodes from the HPRD database [22] and the disease nodes from the Comparative Toxicogenomics Database [21]. Side effect nodes and drug-side effect relations were extracted from SIDER [23]. The dataset captures known target-disease, disease-drug, target-drug, protein-protein, disease-disease and drug-drug associations as (0,1)-matrices [19]. These data matrices (available at https://github.com/luoyunan/DTINet) have been downloaded. We have created semantic (drug-disease-target) triples such that the subject is a drug, the predicate is a disease, and the object is a protein/target. The criterion for generating such a triple is as follows: for a drug *Di*, a disease *D'j*, and a target *Tk*, when and only when all three associations are known in the above data matrices, the triple *Di-D'j-Tk* is considered to be a true known triple; missing any connection in the triple renders it an unknown triple. The whole set of triples {*Di-D'j-Tk*} constitutes the positive samples for DDTE.

2.2. The embedding algorithm

The main motivation of our work is to create a self-contained embedding and link prediction method that takes as its input the heterogeneous network (i.e. knowledge graph) and optimizes a custom designed loss function to ensure that the resultant node embedding vectors (for drugs, proteins and diseases) can be used to effectively model the known relationships among the triples of drugs, diseases and targets. The algorithm does not require pre-calculating node features for drugs, proteins and diseases. The node features (embedding vectors) are the results learned directly from the network/graph structure itself. This approach can afford better prediction of missing links. We also aim to avoid the requirement of conducting a separate step of supervised

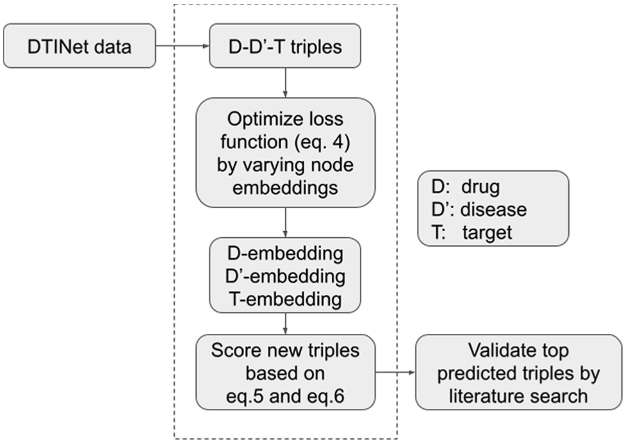

DTINet data
D-D'-T triples
Optimize loss function (eq. 4) by varying node embeddings
D-embedding
D'-embedding
T-embedding
Score new triples based on eq.5 and eq.6
Validate top predicted triples by literature search
D: drug
D': disease
T: target

Fig. 1. Overall workflow of this study.

machine learning. The dynamic nature of the graph embedding algorithm implies that node embedding is not intrinsic to the nodes. Other methods [15–18] need to calculate drug features based on their molecular structures and protein features by their sequences, while our embedding approach can adapt to the network structure itself.

The DTINet dataset can be considered as a knowledge graph (KG), which is formally defined as follows: Let $E = \{e_1, e_2, \cdots\}$ be a set of entities including drugs and proteins. They are the nodes of the graph, and $R = \{r_1, r_2, \cdots\}$ be a set of all relation types, which are the edge labels in the graph. The graph can also be equivalently viewed as a set of triples. The format of a triple is (s,p,o), which is (*subject*, *predicate*, *object*), where the subject (drug) and the object (target) are entities (i.e., nodes), and the predicate (disease) is the relation type.

The set of all known triples in the dataset, denoted as $X_{(s,p,o)}$ or simply X , is considered to be the positive samples by this algorithm. Since an added set of negative samples has been successfully used to help optimize a model to efficiently reduce the cost function (cf. equation (4)), we have adopted a similar negative sampling approach. The set of negative samples $X'_{(s,p,o)}$ is defined as follows:

$$X'_{(s,p,o)} = \{(s', p, o) | s' \in E\} \cup \{(s, p, o') | o' \in E\} \tag{1}$$

where E is the set of all the entities in the dataset; *s*, *p*, *o* are the subject, predicate and object of a triple, respectively. For each known triple in the original set, we create a set of negative samples by replacing the subject *s* with an entity in E as the new subject s'; or replacing object *o* with a random entity in E as the new object *o'*. These negative samples have not been in the original data set; in other words, the intersection of the original set of positive samples $X_{(s,p,o)}$and the generated set of negative samples $X'_{(s,p,o)}$ is empty.

Note: negative sampling is a widely used technology to help with optimizing the loss function in this class of algorithms [20]. In a typical machine learning study, one explicitly creates positive and negative samples in roughly equal proportions. In DDTE (as well as TransE) algorithm, however, this balance is already ensured in the calculation of the loss function (equation (4)). A close analysis of the loss function reveals that for each synthetic negative sample, the algorithm includes its contribution together with the corresponding positive sample from which the negative sample is generated. In other words, each positive sample is used as many times as there are negative samples introduced. Thus, the apparent imbalance issue is dealt with in the loss function.

The algorithm DDTE was inspired from knowledge graph embedding models named TransE [20] and ContE [24]. In TransE, a predicate is regarded as a translation from a subject vector to an object vector. When a triple (*subject*, *predicate*, *object*) holds true, TransE trains the model to represent features of the entities (drugs, proteins) and the predicate in a low dimensional vector space and to make the summed vector of the subject *s* and the predicate *p* to be as close as possible to the vector of the object *o*. DDTE has been developed similarly in the embedding space for drug-disease-target prediction as follows.

The model is trained so as to make $v_s + v_p \approx v_o$ for each known triple (*s*, *p*, *o*) where $v_s, v_p, v_o \in R^k$ are *k*-dimensional vector embedding of the subject *s,* the predicate *p*, and the object *o*, respectively. When a triple is (*s*, *p*, *o*) in the positive set *X,* v_o should be close to $v_s + v_p$; otherwise, $v_s + v_p$ should be distant from v_o. To calculate the distance (dissimilarity) $\eta()$between two embedding vectors, we use the L_1-norm (equation (2) and (3)):

$$t = v_s + v_p \tag{2}$$

$$\eta(s,p,o) = \sum_{i=1}^{k} |t[i] - v_o[i]| \tag{3}$$

Here, *t* is sum of the two vectors (subject and predicate), and *k* indicates the dimensionality of the vectors. The following margin-based loss function is employed as the objective function in this algorithm:

$$L = \sum_{(s,p,o) \in X} \Big(\sum_{(s',p,o) \in X'_{(s,p,o)}} \max\big(0, \gamma + \eta_{(s,p,o)} - \eta_{(s',p,o)}\big) + \sum_{(s,p,o') \in X'(s,p,o)} \max(0, \gamma + \eta_{(s,p,o)} - \eta_{(s,p,o')})\Big) \tag{4}$$

where, $\gamma > 0$ is the margin parameter, $\eta(s,p,o)$ is the dissimilarity between $v_s + v_p$ and v_o for the positive set X , $\eta_{(s',p,o)}$ is the dissimilarity between $v_{s'} + v_p$and v_o for the corresponding negative set of triples; and $\eta_{(s,p,o')}$ is the dissimilarity values between $v_s + v_p$and $v_{o'}$. The mini-batch gradient descent method with AdaGrad [25] is used as the optimization algorithm to minimize the loss function (equation (4)), resulting in the optimized embedding vectors for all subjects, predicates and objects in the dataset. Our chief reason for choosing the L1-norm for distance calculation is its known robustness against outliers in a dataset, which is the case for typical systems chemical biology datasets that are often pulled together from different data sources.

2.3. Model performance statistics

To validate and test the predictive performance of our embedding model, we shall define a scoring function for any given model-predicted triple (*s*, *p*, *o*). The prediction score is $\psi(s = e_i|p,o)$, for all $e_i \in E$ or $\psi(o = e_i|s,p)$, for all $e_i \in E$, where the score roughly corresponds to the probability of each missing subject or missing object. There are two kinds of predictions: (1) subject predictions: given a known link between a predicate *p* and an object *o* and infer missing subjects *s*; and (2) object predictions: given a known link between a subject *s* and a predicate *p*, infer missing objects *o*.

For subject predictions, the following score function is used:

$$\psi(s = e_i|p,o) = -\eta_{(e_i,p,o)} \ \ for \ \forall_{e_i} \in E \tag{5}$$

For object predictions, the following score function is used:

$$\psi(o = e_i|s,p) = -\eta_{(s,p,e_i)} \ \ for \ \forall_{e_i} \in E \tag{6}$$

All predictions are ranked according to the scores to find the *N* highest scoring entities, among which the correct predictions are defined as the hits. Thus, *Hits@N*values are a good measure for the model performance and were calculated in this study. Another more robust statistical measure for model performance is obtained from the Mean Reciprocal Rank (*MRR*), which is computed as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{7}$$

where, *Q* is a set of test triples, and $rank_i$ is the rank position of the true answer for the *i-th* triple. A higher *MRR* value indicates a better model.

Hits@N is the hit percentage of true samples in a test set being ranked by a model within the top N positions against a set of decoy (i.e. negative) samples. This metric reflects how well a model ranks a true sample relative to decoy samples.

To calculate Hits@N for a predictive model (either a DTINet model or a DDTE model), a set of decoy samples is generated from each of the true samples in a test set. For DDTE, the true samples are those DDT (drug-disease-target) triples in the test set; for DTINet, the true samples are the unique drug-target (DT) pairs extracted from those triples. In both cases, the score of a true sample is compared with the scores of respective decoy samples to obtain the ranks of the true sample relative to decoy samples. The subtlety is that DT-pairs are scored in the case of DTINet and DDT triples are scored in the case of DDTE. However, the actual calculations of the Hits@N for DDTE and DTINet are in fact consistent.

Calculation of Hits@N for a DTINet model. (1) For each true DT-pair, a set of decoy drug-target (DT) pairs is generated by combining the drugs and targets available in the test set being considered. The drugs and targets are crossed-coupled, with known DT-pairs removed, to obtain the decoy samples. (2) The known DT-pair and the decoy DT-pairs are then scored by a DTINet model and ranked from best to worst. (3) The ranking of the true DT-pair is recorded. (4) Steps (1) to (3) are conducted for every true DT-pair in the test set. Finally, (5) The percentage of times when the true DT-pair is ranked within top 1, top 3 and top 10 are calculated for Hits@1, Hits@3 and Hits@10, respectively.

Calculation of Hits@N for a DDTE model. The procedure to calculate Hits@N is fundamentally the same as that for DTINet. (1) We generate a set of random combinations of DDT-triples for each true triple in the test set by *holding constant* the disease node in the triple while varying the drugs and targets, replacing them with drugs and targets in the test set. These triples, minus true triples, are used as the decoy samples. (2) For each true triple as well as the decoys, we calculate their scores and all triples are ranked from best to worst. (3) Then the ranking of the true triple is recorded. (4) repeat (1) to (3) for all true triples in the test set. Again, (5) the percentage of times when the true triples are ranked within top 1, top 3 and top 10 are calculated as Hits@1, Hits@3 and Hits@10, respectively.

In addition to MRR and Hits@N metrics, we have also calculated AUROC (Area Under Receiver Operating Characteristic curve) and AUPRC (Area Under Precision-Recall Curve). These two standard metrics reflect overall or averaged performance of a model while MRR and Hits@N assess the algorithm's ability for "early recognition" of interesting predictions. This issue has been recognized in both bioinformatics and cheminformatics applications [26,27]. PRC curves are especially good for showing the trade-offs between precision and recall rates. Together, these metrics afford more comprehensive assessment of a predictive model.

2.4. Hyperparameters optimization

The effectiveness of our algorithm depends mainly on four hyper-parameters described below. Random search approach was adopted to find optimal values for them covering the following ranges:

κ –the dimensionality of the embedding space; the range is$\{60, \cdots, 200\}$
γ – the margin parameter in loss function (equation (4)); the range is $\{0.1, \cdots, 2.0\}$
λ –the learning rate; the range is $\{0.001, \cdots, 0.2\}$
b = the batch size; the range is $\{60, \cdots 200\}$

Note that λ (the learning rate) and b (the batch size) are parameters of the mini-batch gradient descent optimization algorithm adopted for this study.

2.5. Empirical evaluation protocol

The characteristics of the DTINet dataset [19] are given in Table 1. It was converted into triples, which were then randomly split into 60% as the training set, 20% as the validation set used to determine the optimal hyper-parameters, and 20% as the hold-out test set. For 5-fold validation, we shuffled the set of all triples randomly five times, and then each of the shuffled sets is split into training (60%), validation (20%) and test (20%) sets with the same ratios. Thus, five training, validation and test sets have been created for the 5-fold cross-validation experiments. Each training set was used to build a DDTE embedding model; each corresponding validation set was used to optimize the parameter setting; and each test set was used to test the model and generate the performance statistics reported in the results section.

Table 1
Characteristics of the DTINet Dataset.

Protein	1,493
Disease	5,603
Drug	708
Training triples	118,203
Validation triples	39,400
Test triples	39,400

3. Results

3.1. Optimized hyper-parameters

Before conducting the model performance experiments, we first select the optimal values for the four hyper-parameters (cf. METHODS). The validation sets were used to do so. The optimal values were found based on *MRR* results obtained with the validation sets. For this dataset, the optimal hyperparameters were $(\kappa : 140, \gamma : 1.5, \lambda : 0.082, b : 140)$, and they were used throughout all models reported in this paper.

3.2. Model performance statistics

5-fold cross-evaluation experiments were conducted as follows. The whole DTINet data set was first shuffled five times to form five randomly ordered data sets, each of which was divided into training (60%), validation (20%), and test sets (20%). Each training set had 118,203 known triples. There were 39,400 known triples in the validation set and 39,400 known triples in the test set.

To evaluate DDTE embedding model, we computed the scores of triples (s', p, o) for $\forall_{s'} \in E$, and then rank all of these triples by the scores in decreasing order (i.e., the rank of a triple that has the highest score is 1). Likewise, we computed the scores of triples (s, p, o') for $\forall_{o'} \in E$ and rank all of the scored triples similarly in decreasing order as well.

To measure the quality of the models, we use *MRR* as defined above (cf. METHODS). We also report $''Hits@1, 3, 10''$, which indicates the number of correct triples that appear in the top 1, 3, and 10 predictions.

Table 2 shows the 5-fold cross-validation experimental results obtained for this dataset. We ran 5-fold cross-validation and compared it with the original DTINet model [19] as the baseline. The table shows that for the prediction of the five test sets, our model outperformed the DTINet model's baseline. Our model's average *MRR* performance on DTINet data was around 0.80 while that of the DTINet model was about 0.70. Hits@N also indicated that DDTE performed better than corresponding DTINet models for the same datasets.

Since DDTE models score drug-disease-target triples and DTINet models score drug-target pairs, one may wonder how we can use Hits@N to compare the two different types of models. To better understand it, we should take a closer look at the way Hits@N is calculated. The fact that we **hold the disease constant** in a true triple when substituting the drugs and targets to generate decoys for a DDTE model (cf. Methods) allows us to derive the Hits@N that indicates where a true DT-pair (in a triple) ranks relative to the DT-pairs in the decoy triples. As a result, both DDTE model and DTINet model score DT-pairs when comparison is made to obtain the Hits@N (as well as MRR) metrics. The only

Table 2
Test Results on 5-Fold DTINet Data.

	MRR	Hits @ 1	Hits @ 3	Hits @ 10
DDTE (fold 1)	0.80	68.50	89.70	98.55
DTINet (fold 1)	0.71	61.33	78.25	88.63
DDTE (fold 2)	0.79	67.04	89.48	98.56
DTINet (fold 2)	0.70	60.51	77.01	88.21
DDTE (fold 3)	0.79	66.77	89.61	98.58
DTINet (fold 3)	0.70	60.01	76.89	88.10
DDTE (fold 4)	0.79	66.52	89.32	98.46
DTINet (fold 4)	0.72	61.99	78.47	88.95
DDTE (fold 5)	0.80	67.86	90.18	98.66
DTINet (fold 5)	0.70	59.88	76.81	88.02

To compare DDTE and DTINet results, two sample *t*-tests have been performed on MRR, Hits@1, Hits@3 and Hits@10: the p-values are 1.52E-07, 1.01E-06, 3.10E-07, and 2.90E-07, respectively.

difference is how these two models make their predictions. By design, a DDTE model takes advantage of the known disease node involved in a triple when scoring DT-pairs while a DTINet model does not utilize this information in its prediction.

We have also calculated ROC curves and PRC curves (Fig. 2). For each test set, we have created a binary dataset that consists of the true triples in the test set and a set of unique synthetic negative samples. The proportions of true positives versus negative samples are 1:1, 1:5, 1:10 and 1: all possible negatives in 4 respective datasets. These different datasets simulate different levels of imbalance between positive and negative samples, posing increasing levels of difficulty for the models. For all five test sets, when the proportion of negative samples increase, the AUROC remain steady. This phenomenon is also reported in the DTINet paper [19], indicating that ROC curves and AUROC may not be a sensitive metric to evaluate model performance in imbalanced datasets. On the other hand, AUPRC gradually decreases when the proportions of negative samples rise. From 1:1 to 1:5 to 1:10, AUPRCs are 0.99, 0.96, and 0.94, respectively. When all the possible negative samples are considered, AUPRC drops to about 0.39 for DDTE models. This trend has also been reported in the DTINet paper - they have reported that AUPRC dropped to about 0.3 or less when all possible negative samples were used; and other methods in their report got even lower AUPRC (∼0.2 or less). Thus, DDTE models performed favorably based on AUPRC as well. Additional data regarding the 40 curves are included in the Supplementary data file.

DDTE appears to perform better than DTINet in terms of detecting true relationships among drugs and targets at top-ranking computational hits. The likely reasons for this could be as follows. (1) It utilizes data regarding the full connections among a triple of drug, disease and target. The information of true triples enables DDTE to capture important hidden information (i.e. the drug node) which has not been taken into account in other methods (including DTINet). (2) Since the node embedding vectors are generated dynamically based on the network structure as opposed to prior calculations based on chemical structures and protein sequences, DDTE could adapt the embedding vectors to better establish the relationships among drugs, diseases and targets in response to the structure of the network/knowledge graph of a training set. Thus, we believe DDTE represents a new type of network-based drug repurposing approach that is complementary to other published approaches [15–19].

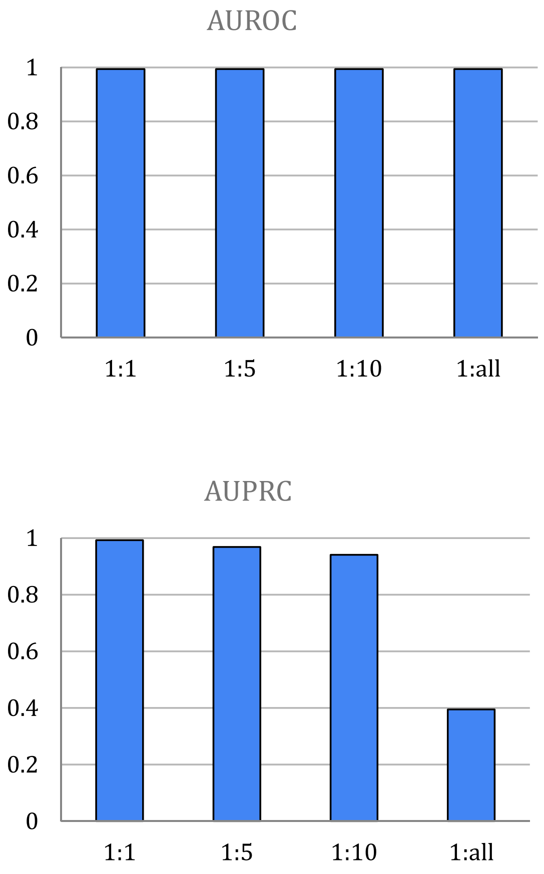


Fig. 2. Bar graphs of Areas Under ROC curves (AUROC) and Areas Under PRC curves (AUPRC). Note: AUROC and AUPRC data are based on the averages of five test sets. Standard deviations are small and not represented on the bar graphs.

3.3. *Identification of high scoring triples not recorded in the DTINet dataset*

As described above, five splits of the training data have been obtained by randomly dividing the dataset into portions of 60%, 20%, and 20% for the training, validation and testing sets, respectively. Thus, every such split led to a different model. We presented the statistics of all five models (Table 2); and it shows comparable performance among all them. Since one of the aims of this study was to demonstrate that some of the top scoring triples (that did not exist in the original DTINet dataset) could be found in the biomedical literature and/or the clinical trials reports, we believe that demonstrating this point using anyone of the models would serve this purpose. In a practical drug repurposing project, one may examine all the predictions by five (or even more) models and the top (consensus) predictions may be tested as drug repurposing candidates. Thus, we simply chose to use the first model to demonstrate the utility of our models in predicting new triples. To do so, we have scored the missing triples created by replacing either the subject or the object in each of the 39,400 test triples in the first test set. As a result, we generated two sets of predicted triples (from object prediction and subject prediction, respectively) and scored them using Equations (5) and (6). The highest scoring one hundred (100) predictions were carefully examined by researchers in PubMed and Web search to assess the viability of the predicted missing relationships. For example, we have taken each of the top-20 ranking computational hits between a drug and a disease, and searched it in the PubMed to see if evidence is found in the published work that shows the potential relationship between the drug-disease pair. We also searched ChemoText [28] for the predicted drug-disease pair or drug-side effect relationship. If they are found in either PubMed or ChemoText, the pair is labeled as a viable hypothesis (either a potential drug repurposing candidate or a potential side-effect prediction).

3.4. *Literature validation of predicted new relations*

Due to the nature of the training sets, the relationships between drugs and diseases are not restricted to therapeutic ones. In fact, drug side effects are also included in the training set, and thus they are in the model-predicted new relations as well. The following are examples of the predicted relationships verified through literature (PubMed) and web search (Table 3).

Table 3
Predicted New Relations.

Drug Name	Affected Condition	UniProt ID	Reference
New indications			
Propranolol	Child behavior disorders	P29317	[29]
Mepivacaine	Hypertension	P07359	[31]
Raltitrexed	Colonic disease	Q9UI32	[7]
Side effects			
Bortezomib	Gastroenteritis	Q9UIC8	[33]
Sulfasalazine	Respiratory distress syndrome	P00367	[34]

3.4.1. *Potentially new therapeutic relations*

Here we discuss several examples that illustrate the potential utility of our approach as a novel drug repurposing tool. The first prediction is the association between Propranolol and child behavior disorders. Propranolol was initially developed and used for the treatment of hypertension. Literature search has revealed that it could lead to significant improvements in cognitive performance in children with autism spectrum disorders [29]. Therefore, this newly predicted relation is of interest providing support to using our models for predicting potentially new therapeutic applications of an old drug. The associated target is P29317 (ephrin type-A receptor 2). Although it is not known how Propranolol affects child behavioral disorder, it is known that ephrins are involved in brain disorders with memory impairment symptoms, including Alzheimer's disease and anxiety. Ephrins may therefore induce cellular alterations mandatory for memory formation [30].

The second prediction is the association between Mepivacaine and hypertension. This drug is typically used as a topical anesthetic. It was known that anesthetics could cause cardiovascular side effects in hypertensive patients. A literature search has found that Mepivacaine solution without vasoconstrictor was safely used in hypertensive patients [31]. This fact was not included in the DTINet dataset, and again, this prediction appears to be supported by the literature data. The associated protein target is predicted to be P07359 (platelet glycoprotein Ib alpha chain); but it is not yet clear how this protein is involved in either topical anesthesia or hypertension.

Our model also predicted a link between Raltitrexed and colonic disease. Based on DrugBank information, Raltitrexed was used for the treatment of malignant neoplasm of colon and rectum, as well as pleural mesothelioma [7]. This prediction does not appear to be new, but it was not included in the original training set; and thus, we considered this a successful prediction for a "new" use of an old drug. The associated protein is Q9UI32 (glutaminase liver isoform, GLS2), which is known to be highly expressed in cancers of the bladder, breast, cervix, colon, kidney, liver, lung, ovary, prostate, rectum, thyroid and thymus [32].

3.4.2. *Potential side effects predictions*

We have predicted associations that imply side effects linked to specific drugs. For example, Bortezomib was predicted to be associated with gastroenteritis. This drug was originally used to treat multiple myeloma. According to literature search, it could indeed cause gastrointestinal problems [33]; however, this association was not present in our training set and thus considered as a new prediction. The associated protein target is predicted to be Q9UIC8 (leucine carboxyl methyltransferase 1); but no known information about this protein's involvement in gastrointestinal problems has been found.

Another case was Sulfasalazine, which was typically used for the treatment of Crohn's disease and rheumatoid arthritis. Our model predicted it to be associated with respiratory distress syndrome. In a literature search, it was found that this drug could indeed cause difficulty in breathing [34]. The associated protein is predicted to be P00367 (mitochondrial glutamate dehydrogenase1); but its connection to respiratory distress syndrome is not yet known.

3.5. *Verification of predictions by ChemoText*

Aforementioned manual searches in PubMed have supported the predictions by our algorithm in that the predicted cases implicate either new drug-indication pairs or drug-side-effect associations (Table 3). To further systematically verify these associations, we have examined a publicly available Web server (http://chemotext.mml.unc.edu) named ChemoText [28], which has captured all the MeSH (Medline Subject Heading) terms and their relationships. It is capable of identifying known drug-disease-target relationships and infer missing links between drug-disease-target triangles. All five of our predicted associations have been found in ChemoText as follows.

Propranolol is predicted by DDTE to be related to child behavior disorders while ChemoText has indeed documented the association between this drug and sleep behavior disorder. In another case, Mepivacaine has been predicted to be useful for hypertension, and ChemoText has also documented its association with hypertension as well. Finally, DDTE has predicted the link between Raltitrexed and colonic disease; yet again, ChemoText has documented its connection with colonic neoplasms.

For the two cases of side effects, Bortezomib was predicted by DDTE to be associated with gastroenteritis, and ChemoText has documented its relation with gastrointestinal diseases as well. Finally, DDTE predicted that Sulfasalazine may have side effect of respiratory distress syndrome; A search in ChemoText also revealed that Sulfasalazine is indeed associated with adult respiratory distress syndrome.

4. Discussion

We have conducted 5-fold cross-validation experiments to test the models built upon randomized training sets consisting of 60% of the total original set of triples in the DTINet dataset. The validation sets consisted of 20% of the total original set, which was used to find the optimal hyper-parameters to obtain the best models. These models were then used to predict the test sets consisting of another 20% of the original set of triples. However, once the optimal hyper-parameters were found, we could use 80% of the randomized dataset as the training and remaining 20% as the test set. This could further improve the performance over what was reported herein.

In addition to the above experiments, we have also employed the obtained optimal models to predict some of the original unknown (negative) samples to identify potential new relations among disease-drug-target triples. The identified drug-disease and drug-side-effect relationships were consistent with manual literature findings. These findings demonstrated that this new algorithm was effective in identifying potentially new relations. We have also searched in ChemoText, a web server that has captured algorithmically all known relations among drugs, targets and diseases published in PubMed based on MeSH terms. This has further supported our predictions.

The DDTE method has been designed to find missing links (a.k.a. link prediction) in an existing network/knowledge graph dataset rather than to predict links among external nodes (drugs, diseases and targets) in other databases. In fact, because of its dynamic embedding nature, DDTE is not able to predict the links among external nodes (nodes that do not already exist in the training set). This algorithmic characteristic can be viewed as a limitation of this method compared to other methods in that they require independent feature generation based on molecular structures and protein sequences, which are more appropriate for predicting relationships among external nodes. However, we believe that DDTE's network-specific dynamic embedding may have afforded it higher predictive power in detecting missing links among existing nodes because it explicitly takes into account the network/graph structure, while other methods do not do so. DDTE should be considered as complementary to previously published methods. It works well in finding missing links in a given dataset, but does not aim to predict other drug-target-disease relationships among novel nodes in external datasets.

In this presentation, we have studied only one of the test sets to predict potential unknown links among proteins, diseases and drugs. To get a more comprehensive evaluation, we could examine more test sets and find more unknown relationships that may be verified in the literature. This could find more verifiable relationships and afford additional repurposing hypotheses for future experimentation.

As the predicted unknown relationships could be verified in the literature, they formally represent retrospective analyses that can be regarded as instances that validate our models and can be viewed as proof-of-concept studies. Further, we have also identified additional hypotheses about either novel therapeutic uses of existing drugs or their possible side effects (data not reported herein); these hypotheses await experimental assessment.

It is interesting to make a comparison between DDTE and several contemporary methods that have been published that integrated heterogeneous network data and build predictive models for predicting new drug-target relationships. Chen et al. [15] integrated three different networks: protein–protein similarity network, drug–drug similarity network, and known drug–target interaction into a heterogeneous network. It uses a random walk algorithm to infer new connections between drugs and potential targets. No prior knowledge of drug-disease relationships is directly modeled in the algorithm. They seek to predict drug-target interaction, with known drug-target interaction as the training set to conduct a supervised learning of a transformation matrix used to predict potential drug-target interactions. Fu et al. [16] implemented a method that is similar to the above. Again, they employed known drug-target links from several data sources. Full connections among drug-disease-target triples were not explicitly employed in these methods. They do not predict drug-disease associations directly from the algorithm; rather, only drug-target interactions are directly predicted. Wang et al. [17] has reported a systematic approach to drug repositioning problem. A unique characteristic of their framework is that it automatically incorporates drug-target information into drug–disease association prediction. Their method does not require all connections (drug-disease, drug-target and disease-target) either. Similarly, Zheng et al. [18] developed a method of matrix factorization based on structure-structure similarity and target-target sequence similarity to establish relationships between drugs and targets. These methods rely on chemical structure similarity based on structural fingerprints and protein sequence-based similarity. The features of the nodes are precalculated based on structure and sequence information rather than derived from the network/knowledge graph structure.

Most recently, Fahimian et al. [35] has published a paper on a technology called RepCOOL. It represents a different class of technology. It constructed nine drug-disease networks from different data sources as the training sets. Each pair of drug-disease relationship is represented as a 9-dimensional feature vector, which is quite different from that of DDTE. In RepCOOL, several supervised machine learning technologies are used in the downstream to build classification models as a separate step.

DDTE, on the other hand, represents a very different class of methods and offers a complementary approach to the problem. In this method, (1) all node embedding vectors are NOT pre-calculated based on chemical structures and protein sequences; rather, node embedding vectors are derived from the network/graph structure that includes all connections among drugs, diseases and targets, which better captures the network structure in generating the embedding vectors/features; (2) DDTE includes explicitly all the connections among a drug, a disease and a target in the algorithm, laying the foundation for predicting drug-disease-target relationships directly as a result of the model prediction.

Finally, it is also interesting to mention a whole new class of graph data analysis method called Graph Neural Network (GNN) [36]. Our method (DDTE) adopts the assumption that majority of the known drug-disease-target relationships should satisfy $s + p = o$, where s is the drug vector, p is the disease vector and o is the target vector. The node embedding vectors are optimized so that this condition is satisfied for as many known triples as possible. This algorithm explicitly captures the triple relationships in the embedding space in a supervised learning fashion. On the other hand, GNN [36] and MPNN [37] aim to learn, in mostly unsupervised or semi-supervised fashion, the internal network/graph structure and map the node-to-node similarity relationships onto the embedding space. No known triple relationships are explicitly captured in these GNN (Graph Neural Network) models. The resultant node embedding vectors are often used in downstream machine learning modeling. Empirical comparisons between DDTE and GNN are currently beyond the scope of this paper but warrant future studies.

5. Conclusions

We have developed and validated a new biomedical knowledge graph embedding algorithm (DDTE). We found that this new approach enabled favorable performance in inferring novel drug-disease-target or drug-target-side effect relationships as compared to previously published results using the same knowledge graph. New links among drugs and diseases or drugs and side effects were found in our analysis of the DTINet dataset, demonstrating the potential use of this algorithm in drug repurposing or side effect prediction research campaigns.

We have analyzed only one chemical biology dataset, namely, the DNINet to demonstrate the value of the new graph embedding algorithm. More knowledge graph databases have appeared in the literature. For example, Pharos [9] and ROBOKOP [10] have been published in the past two years. They can be converted into semantic triples in the same format as (*s, p, o*) triples. As the algorithm developed herein is of general utility, it can be applied to study these larger databases, which can provide much larger sets of drug repurposing hypotheses and add significant values to these knowledge graph databases.

6. Contributors

AT, RYC and WZ conceived the project. CM designed and wrote the DDTE program. CM, CJ, XD, and SA carried out the computations. CM wrote the initial draft; and WZ, AT and RYC wrote and edited the final version. All authors approved the final manuscript.

Funding

This study was supported in part by NIH award to AT and WZ (U01CA207160), NIH award to AT (1OT2TR003441-01) and a joint ROI grant from the University of North Carolina System to AT, RYC and WZ.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

WZ and XD acknowledge the support from the Golden Leaf Foundation to the BRITE Institute.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103838.

References

[1] T.I. Oprea, J.E. Bauman, C.G. Bologa, et al., Drug Repurposing from an Academic Perspective, Drug Discov. Today Ther. Strateg. 8 (2011) 61–69.

[2] N.C. Baker, S. Ekins, A.J. Williams, et al., A bibliometric review of drug repurposing, Drug Discov. Today 23 (2018) 661–672.

[3] S. Kumar, S. Kumar, Molecular Docking: A Structure-Based Approach for Drug Repurposing, Silico Drug Des. (2019) 161–189.

[4] O.A. Gani, B. Thakkar, D. Narayanan, et al., Assessing protein kinase target similarity: Comparing sequence, structure, and cheminformatics approaches, BBA 1854 (2015) 1605–1616.

[5] J. Schuler, R. Samudrala, Fingerprinting CANDO: Increased Accuracy with Structure- and Ligand-Based Shotgun Drug Repurposing, ACS Omega 4 (2019) 17393–17403.

[6] G. Floresta, V. Patamia, D. Gentile, et al., Repurposing of FDA-Approved Drugs for Treating Iatrogenic Botulism: A Paired 3D-QSAR/Docking Approach, ChemMedChem 15 (2020) 256–262.

[7] D.S. Wishart, C. Knox, A.C. Guo, et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (2006) D668–D672.

[8] B. Chen, X. Dong, D. Jiao, et al., Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data, BMC Bioinf. 11 (2010) 255.

[9] D.-T. Nguyen, S. Mathias, C. Bologa, et al., Pharos: Collating protein information to shed light on the druggable genome, Nucleic Acids Res. 45 (2017) D995–D1002.

[10] C. Bizon, S. Cox, J. Balhoff, et al., ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources, J. Chem. Inf. Model. 59 (2019) 4968–4973.

[11] A. Grover, J. Leskovec, node2vec: Scalable Feature Learning for Networks, KDD 2016 (2016) 855–864.

[12] M. Grohe, word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data, in: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 2020, https://doi.org/10.1145/3375395.3387641.

[13] X. Yue, Z. Wang, J. Huang, et al., Graph embedding on biomedical networks: methods, applications and evaluations, Bioinformatics 36 (2020) 1241–1251.

[14] K. Bleakley, Y. Yamanishi, Supervised prediction of drug-target interactions using bipartite local models, Bioinformatics 25 (2009) 2397–2403.

[15] X. Chen, M.-X. Liu, G.-Y. Yan, et al., Drug-target interaction prediction by random walk on the heterogeneous network, Mol. BioSyst. 8 (7) (2012) 1970–1978.

[16] G. Fu, Y. Ding, A. Seal, et al., Predicting drug target interactions using meta-path-based semantic network analysis, BMC Bioinf. 17 (1) (2016) 160.

[17] W. Wang, S. Yang, X. Zhang, et al., Drug repositioning by integrating target information through a heterogeneous network model, Bioinformatics 30 (20) (2014) 2923–2930.

[18] X. Zheng, H. Ding, H. Mamitsuka, et al., Collaborative matrix factorization with multiple similarities for predicting drug-target interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1025–1033.

[19] Y. Luo, X. Zhao, J. Zhou, et al., A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, Nat. Commun. 8 (1) (2017) 1–13.

[20] A. Bordes, N. Usunier, A. Garcia-Duran, Translating Embeddings for Modeling Multi-relational Data, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013, pp. 2787–2795.

[21] A.P. Davis, C.J. Grondin, R.J. Johnson, et al., The Comparative Toxicogenomics Database: update 2019, Nucleic Acids Res. 47 (2019) D948–D954.

[22] T.S. Keshava Prasad, R. Goel, K. Kandasamy, et al., Human Protein Reference Database–2009 update, Nucleic Acids Res. 37 (2009) D767–D772.

[23] M. Kuhn, M. Campillos, I. Letunic, et al., A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (2010) 343.

[24] C. Moon, S. Harenberg, J. Slankas, et al., Learning Contextual Embeddings for Knowledge Graph Completion, Pacific Asia Conference on Information Systems (PACIS) vol. 10 (2017), 3132847.3133095.

[25] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (7) (2011).

[26] W. Zhao, K.E. Hevener, S.W. White, et al., A statistical framework to evaluate virtual screening, BMC Bioinf. 10 (2009) 225, https://doi.org/10.1186/1471-2105-10-225.

[27] J.-F. Truchon, C.I. Bayly, Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem, J. Chem. Inf. Model. 47 (2) (2007) 488–508.

[28] S.J. Capuzzi, T.E. Thornton, K. Liu, et al., Chemotext: A Publicly Available Web Server for Mining Drug-Target-Disease Relationships in PubMed, J. Chem. Inf. Model. 26 (2018) 212–218.

[29] I. Sagar-Ouriaghli, K. Lievesley, P.J. Santosh, Propranolol for treating emotional, behavioural, autonomic dysregulation in children and adolescents with autism spectrum disorders, J. Psychopharmacol. 32 (2018) 641–653.

[30] M. Dines, R. Lamprecht, The Role of Ephs and Ephrins in Memory Formation, Int. J. Neuropsychopharmacol. 19 (4) (2016) pyv106.

[31] B. Ezmek, A. Arslan, C. Delilbasi, et al., Comparison of hemodynamic effects of lidocaine, prilocaine and mepivacaine solutions without vasoconstrictor in hypertensive patients, J. Appl. Oral Sci. 18 (2010) 354–359.

[32] W.P. Katt, M.J. Lukey, R.A. Cerione, A tale of two glutaminases: homologous enzymes with distinct roles in tumorigenesis, Future Med. Chem. 9 (2) (2017) 223–243.

[33] A. Sharma, C.V. Preuss, Bortezomib. StatPearls, StatPearls Publishing, Treasure Island (FL), 2020.

[34] http://www.mayoclinic.org/drugs-supplements/sulfasalazine-oral-route/side-effects/drg-20066179 (accessed on October 15, 2020).

[35] G. Fahimian, J. Zahiri, S.S. Arab, et al., RepCOOL: computational drug repositioning via integrating heterogeneous biological networks, J. Translat. Med. 18 (2020) 375.

[36] S. Zhang, H. Tong, J. Xu, et al., Graph convolutional networks: a comprehensive review, Comput. Soc. Netw. 6 (2019) 11.

[37] M. Withnall, E. Lindelöf, O. Engkvist, et al., Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction, J. Cheminf. 12 (2020) 1.