# Deep mining heterogeneous networks of biomedical lined data to predict novel drug-target associations (DeepWalk)

* Abstract: - deep learning reveals feature of vertices of a large network that can be adapted in accommodating the similarity-based solution to provide of a large network that can be adapted in accommodating the similarity-based solutions to provide a flexible method of drug-target prediction.
* The researchers proposed a similarity-based drug-target prediction method that enhances existing association discovery methods by using a topology-based similarities measure.
* DeepWalk is a deep learning method, is adopted in the study to calculate the similarities within Linked Tripartite Network (LTN), a heterogeneous network generated from biomedical linked datasets.
* It shows promising results for drug-target association prediction: 98.96% AUC ROC score with a 10-fold cross-validation and 99.25% AUC ROC score with a Monte Carlo cross validation with LTN.
* After utilizing the DeepWalk they demonstrate that: (i) this method outperforms other existing topology-based similarity computation methods, (ii) the performance is better for tripartite than with bipartite networks and (iii) the measure of similarity using network topology outperforms the ones derived from chemical structure (drugs) or genomic sequence (targets).
* The researchers propose methodology proves that it can provide solution for drug-target prediction based on topological similarity with a heterogeneous network, and it may be readily re-purposed and adapted in the existing of similarity-based methodologies.
* Introduction: - drug-target prediction play an important role in drug discovery and drug repurposing. Because it is time and cost consuming researchers only investigate only few complete pharmacological profiles of desired target proteins and these small molecules are rarely systematically screened.
* The comprehensive understanding of drug-target associations, however, is relatively limited compared to the large number of chemical compounds and proteins discovered; this gap in knowledge is a strong incentive to predict associations between existing drugs and its targets.
* Computational methods can complement and guide in these experiments, early attempts of computational prediction, using docking simulations, are neither scalable nor adequate to handle the proteins missing 3-D structure information.
* The recent development in chemical databases present challenges to the researchers, to overcome these limitations, researchers have to adopt diverse machine learning methods, such as classification methods, and rule-based inference methods to predict drug-target associations.
* Similarity measure are fundamental to these methodologies. For example, the similarity measures of drug-drug and target-target pairs can be utilized for the weighting of potential associations, or to generate distinct kernel functions to train the different classification models.
* Recent researches show that the abundant topological interactions between biomedical entities in heterogeneous networks appears to be valuable for assisting in predictions.
* These topology-based methods are incapable of computing the topological similarities between biological entities; they cannot be reused and adapted in the existing similarity-based methods.
* Deep learning methods provide solution for extracting features of vertices, therefore, adopting deep learning methods for topological similarity measure provides tremendous value in drug-target prediction by reusing and adapting the existing similarity-based methods.
* In this paper researchers propose a similarity-based drug-target prediction method that adopts a deep learning algorithm, DeepWalk, to calculate the similarities for drug-drug and target-target pairs based on the topology of heterogeneous network named Tripartite Linked Network (TLN), derived from the existing linked open datasets in biomedical domain.
* The resulting similarity measure is used to infer drug-target association based on the ‘guilt-by-association’ principle that uses drug-drug and target-target similarities as the input for drug-target prediction.
* The researchers evaluate their method in following benchmark: -
  + Performance of a deep learning method compared to other topology-based similarity methods,
  + Value of multipartite (tripartite) network over bipartite networks and
  + Performance of topology-based similarity method over the ones relying on chemical structure and genomic sequence.
* The proposed method is proven to be capable of providing a flexible solution for drug-target prediction based on a heterogeneous network and can be easily reused and adapted in the excising similarity -based methods.
* Materials and methods: - Pipeline of similarity-based drug-target prediction with heterogeneous network: - the drug-target prediction method researchers propose is based on the topology of multipartite network of the existing drugs and protein targets.
* The association discovery pipeline can be separated into three steps: (i) Data preparation and benchmarking, (ii) similarity learning and (iii) association discovery.
* At start, a multipartite network which contain the topological interactions of the existing drugs and targets is constructed with the biomedical linked data.
* After that the similarity scores of the drug-drug and target-target pairs are learned based on the topology of the network.
* At, last new drugs-target associations are discovered and evaluated based on these similarities.
* Data preparation and benchmarking: - the researchers obtained the drugs, targets and drug-target associations from DrugBank which ascertains data-rich molecular biology content found in curated sequence databases, medicinal chemistry textbooks and chemical reference handbooks, and validates the collected data with the journal articles and textbooks.
* To create the LTN, they extracted the diseases, drug-disease and disease-gene associations from a human disease network named Diseasome, and merged these associations with the bipartite network they obtained from DrugBank.
* The disease-target association was created by mapping targets of DrugBank to the genes of the disease-gene associations in Diseasome based on Bio2rdf, Uniprot, HGNC and OMIM.
* Similarity Learning: - DeepWalk, a deep learning method, vectorizes the vertices in the network for similarity computation.
* In this method the local latent information of topology based on truncated random walks and maximizes the probability of a next vertex given the previous vertices in these walks.
* Two components are inherent in Deep Walk: (i) for each vertex , times of random walks with the length *t* are conducted with as the starting vertex, and (ii) for each walk, the SkipGram algorithm updates the vertex representation.
* SkipGram maximizes the co-occurrence probability among the vertices within a within *w* using the assumption as follows,



* Here, is the latent topological representation associated with each vertex , is modeled with a |*V*| d matrix, where |*V*| is the cardinality of vertex set V, and *d* is the dimension user input.
* *Pr*()) is approximated with Hierarchial Softmax by assigning the vertices to the leaves of a Huffman tree, and *Pr*()) can be computed as,



* Here and is the representation assigned to the vertex ’s parent.
* ) is a sequence of tree vertices to identify the vertex , where = root and = .
* The similarity of two vertices *u* and *v* is calculated as follows,



* Here, *d* is the dimension, and , are the components of vector *u* and *v* respectively.
* Association Discovery: - the researchers adapted two popular rule-based inference methods, drug-based similarity inference (DBSI) and target-based similarity inference (TBSI), to discover the drug-target associations with the similarities.
* DBSI predicts a drug-target association *s*() if a drug is similar with a drug that has an existing association with a target .
* For a pair of (), a confidence score of the pairs is calculated as,



* Here, *sim*(), is the similarity between and , and = 1 if there is an existing association between and otherwise = 0.
* Like this TBSI predicts a drug-target association *s*() if a drug is associated with a target that has a similar target .
* For a pair of (), a confidence score of the pair is calculated as,



* Here *sim*(), is the similarity between and , and = 1 if there is an existing association between and otherwise = 0.
* For a drug or a target as the input query, the DBSI and TBSI confidence are normalized as,



* Here *Max*() is the maximum confidence and *Min*() is the minimum confidence for , and *Max*() is the maximum confidence and *Min*() is the minimum confidence for .
* Validation and evaluation metrics: - the researchers evaluated the predictions using three kinds of validation methods based on internal and external references.
* For internal validation, they implemented a 10-fold cross-validation, where conventionally a dataset is partitioned into 10 subsets: 1 for testing and 9 for training.
* They need to restrict the evaluation and benchmarking of their method to the similarity measure itself, and eliminate the impact of the inference-based method (TBSI or DBSI).
* They first randomly extracted a set of association , making sure that no isolated vertices are created. Then they derived , which is the complement of in the association space.
* The associations were randomly partitioned into 10 subsets {}.
* In each test of ten, a subset was used as a gold standard for testing while the nine remaining subsets of as well as were used as the training set.
* They also performed a Monte Carlo validation, where the set of associations were randomly partitioned into two parties, with the cardinality of M and , and used as gold standard predictions for the test and the rest subsets , and were used as the training set.
* For the validation with the external reference, they used the whole dataset as their training set and validated the predictions with the newly discovered drug-target associations.
* They calculated the Area Under the Receiver Operating Characteristic Curve (AUC) and Recovered Fraction (RF), to assess the quality of the predicted associations.
* They computed AUC with the ROC JAVA library, RF in top N was obtained from , where was the number of true positive predicted associations in *i*th query and was the number of missing associations in the gold standard.
* Results: - Comparison with topology-based similarity measures in bipartite network: - the researchers compute six association indices used for similarity computation in bipartite networks, which were Jaccard, Simpson, Geometric, Cosine, Pearson Correlation Coefficient (PCC) and Hypergeometric, and compared them with the DeepWalk results.
* They measured the top K percentage (5, 10, 15, 20, 30) and top N (10, 20, 50, 100, 500, 1000) predicted associations in the two validations.
* DeepWalk performed the best when the top K percentage of predicted associations were considered.
* They also observed that DeepWalk performed best when the top 500 and 1000 predicted associations were considered; other methods performed best with top 10, 20, 50, 100 predicted associations.
* Comparing the use of tripartite and bipartite network: - to determine whether the utilization of the additional disease-drug and disease-target associations improved the drug-target predictions, they compare the use of bipartite and tripartite networks.
* To check the performance of the DeepWalk the researchers applied both bi- and tripartite networks, SimRank and Line.
* SimRank computes vertex similarity with the structural context in a network based on a graph-theoretic mode and is applicable in any domain with object-to-object relationships.
* Line is a graph embedding method that represents the vertices in a network structure, local and global, to capture the first-order proximity and second-order proximity between the vertices.
* The researchers used two variants of Line obtained, Line (1st) and Line (2nd), which utilizes first-order and second-order proximity respectively.
* DeepWalk performed better than Line and SimRank in both types of validations with both DBSI and TBSI association models.
* Tripartite networks improved DeepWalk’s performance: from 84.23% to 87.83% with DBSI and from 85.75% to 90.31% with TSBI in the ten-fold validation; from 83.86% to 86.41% with DBSI and from 82.94% to 86.23% with TSBI in the external source-validation.
* They observed that DeepWalk achieved the best RF scores in all top K (except 5%) percentage and top N (500 and above) predicted associations.
* Not all drugs or targets could be associated with diseases in LTN. Therefore, in order to determine the respective influence of drug-disease and disease-target associations in the tripartite network, they partitioned the network into two main components: (i) fully connected, and (ii) partially connected component, to generate training and test sets of associations.
* The researchers compared the use of tripartite and bipartite network for the prediction on FDTA and PDTA using the same methods mentioned above.
* The same settings are used to conduct the analysis for DeepWalk, the researchers observe that the use of tripartite networks offer the largest improvement in prediction of drugs and targets that were associated with the disease.
* In the 10-fold validation, AUC scores of the FDTA were improved from 89.28% to 98.96% using DBSI model, and from 88.14% to 91.19% using the TSBI model.
* In the external source validation, scores were improved from 83.52% to 91.16% using DBSI model, and from 80.68% to 91.51% using the TBSI model.
* The AUC scores of the PDTA were hardly improved by using tripartite network, which indicated that the addition of new associations only improved the predictions of drugs and targets directly associated with diseases.
* The purpose of the similarity measures for DBSI and TBSI models is to make the drug-drug and target-target pairs, to infer drug-target associations by obtaining high similarity scores from drug-target associations that are true and low similarity scores for drug-target associations that are false.
* Two types of pairs were analyzed based on the contribution to DBSI and TBSI models in prediction: (i) positive pairs, which similarity is used to predict true drug-target associations, and (ii) negative pairs, which lack of similarity is used to confirm false drug-target associations.
* The researchers observed a notable improvement of the similarity calculations for FDTA by switching the input data from bipartite to tripartite network, and minor changes were sent for PDTA.
* The similarity distribution of the positive target-target pairs in the FDTA had a more remarkable improvement by using the tripartite network than with the drug-drug pairs, which was consistent with the experimental results the TBSI performed better than DBSI.
* They also showed that using tripartite network did not improve the similarity computations for the negative pairs for both FDTA and PDTA predictions.
* Influence of the number of available associations (i.e., data richness): - Testing the influence of the data richness required performing a Monte Carlo cross-validation, where M drug-target associations were randomly removed from the bipartite and tripartite networks as validation data and the remained data were used for training, respectively.
* Both AUC scores of the bipartite and tripartite networks decreased as more associations were removed – bipartite network suffered more than tripartite network.
* The three types of associations: drug-target, drug-disease and disease-target, existed in the LTN.
* The drug-target associations contributed both to the similarity computation and association discovery methods, while the drug-disease and disease-target associations contributed only to the similarity computation
* They designed three types of removal strategies: - (i) ‘drug-target removed’: removing drug-target associations while preserving all the vertices connected, thus without creating isolated drug vertices or isolated target vertices. (ii) ‘disease conserved’ which removed drug-disease or disease-target associations without isolating disease from the drugs and targets. (iii) ‘disease conserved’ which removed drug-disease or disease-target associations without consideration of keeping diseases connected after the removal.
* The researcher randomly removed P percentage from these 3 types to validate the results with external resources.
* Without hurting the discovery method, ‘disease conserved’ has a sharper drop than ‘drug-target removed’, which indicates that the enrichment of the drug-target bipartite network results in better predictions than improving drug-target associations.
* Topology-based V.S. Chemical structure- or genomic sequence-based: - the researchers compared topology-based DeepWalk to chemical structure- and genomic sequence-based methods across four experiments, predicting drug-target, association for four kinds of targets: ‘Enzyme’, ‘GPCR’, ‘Ion channels’ and ‘Nuclear receptor’.
* The DeepWalk using topology of the tripartite networks for similarity compuatations, outperforms the other two methods based on chemical structure and genomic sequence in terms of AUC.
* Discussion: - the DeepWalk has some limitations they are - the proposed method can predict the associations between the drugs and targets that exists within the network, but may not predict new drugs or targets in some practice use scenarios.
* Secondly, compared to DeepWalk, the traditional topology-based method, SimRank, shows potential for top N predictions, which can serve as an alternative method for these prediction efforts.
* The use of complex network may lead to new issues, like the effort of the pathway lengths or even the size and shape of networks, which can be caused by data mapping/integration in the network construction.
* The method presented by researchers is somewhat monotonous, it only considers the diseases in the network and is still not comprehensive enough to capture all the characteristics of drugs or targets that may not be represented on a network.
* Future studies could propose a hybrid similarity measure that includes both topological and non-topological features.
* In conclusion, we can say that the propose method assembles the similarity measure with the rule-based inference methods, DBSI and TBSI, for drug-target prediction.
* The deep learning methods for similarity measures can be associated with alternative classification models, which may lead to an improved performance overall in the future studies.