Leveraging hierarchical tree data as guide for random walk on heterogeneous network for drug repurposing

Dongmin Bang[1], Sangsoo Lim[2] and Sun Kim[1,3,4,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University
[2]BioInformatics Institute, Seoul National University
[3]Department of Computer Science and Engineering, Seoul National University
[4]AIgenDrug, Co., Ltd

Background

- To solve the problem of escalating cost and time required for new drug development, drug repurposing drew attention as a new paradigm. Drug repurposing focuses on using 'old' drugs to treat both common and rare diseases outside the scope of the original indication.
- Various tools leverage large-scale heterogeneous biomedical network for drug-repurposing.
  I. Classical machine learning-based method : Degree Weighted Path Count
  II. word2vec-based method : edge2vec
  III. Network propagation-based method : multiscale-interactome
  IV. Graph Convolutional Neural Network : LAGCN

Motivation

- Handling heterogeneous biomedical network is a very difficult and yet unsolved problem. There exists two major challenges for current drug repositioning tools :
  I. Most of the existing node embedding tools do not consider multiple node-type and edge-type features of biomedical network.
  II. Biomedical networks are highly biased to genes and gene-gene interactions, which cover up to 81% of nodes and 89% of edges of the whole network.
- To handle these challenges, we introduce a new concept for guiding random walker with biological prior knowledge when generating sequences of nodes prior to node-embedding generation.
  I. Teleportation, a concept borrowed from Google's PageRank algorithm[1], guides random walker to teleport to a hierarchically similar compound and disease, instead of randomly following the network topology. the pathway and the context.

Method

- As illustrated in Figure 1, 'RandomTeleporter' consists of three modules : Teleport-guided random walk, Skip-gram model, and logistic regression for link prediction. After training, model predicts the treatment probability of a given drug-disease pair.
- We implemented "RandomTeleporter" by integrating word2vec model with teleportation from PageRank algorithm and edge type-transition matrix from edge2vec model[2].
  I. Edge-type transition matrix is trained before the random-walk process, which implies edge-type distribution of the heterogeneous network. The random walker then chooses the next edge-type according to the prior edge-type and its transition matrix.
  II. Teleportation occurs when random walker arrives at any disease or drug node, following the user-given teleportation probability. Our model teleports the walker to a hierarchically similar drug/disease node, rather than any other random node.
- After performing knowledge-guided random walk, node sequences are then passed onto the famous skip-gram model for node embedding generation.
- Finally, two embedded nodes, each from drug and disease, are subtracted and the outcome vector is used as input for logistic classifier which performs binary classification and outputs a treatment probability score.

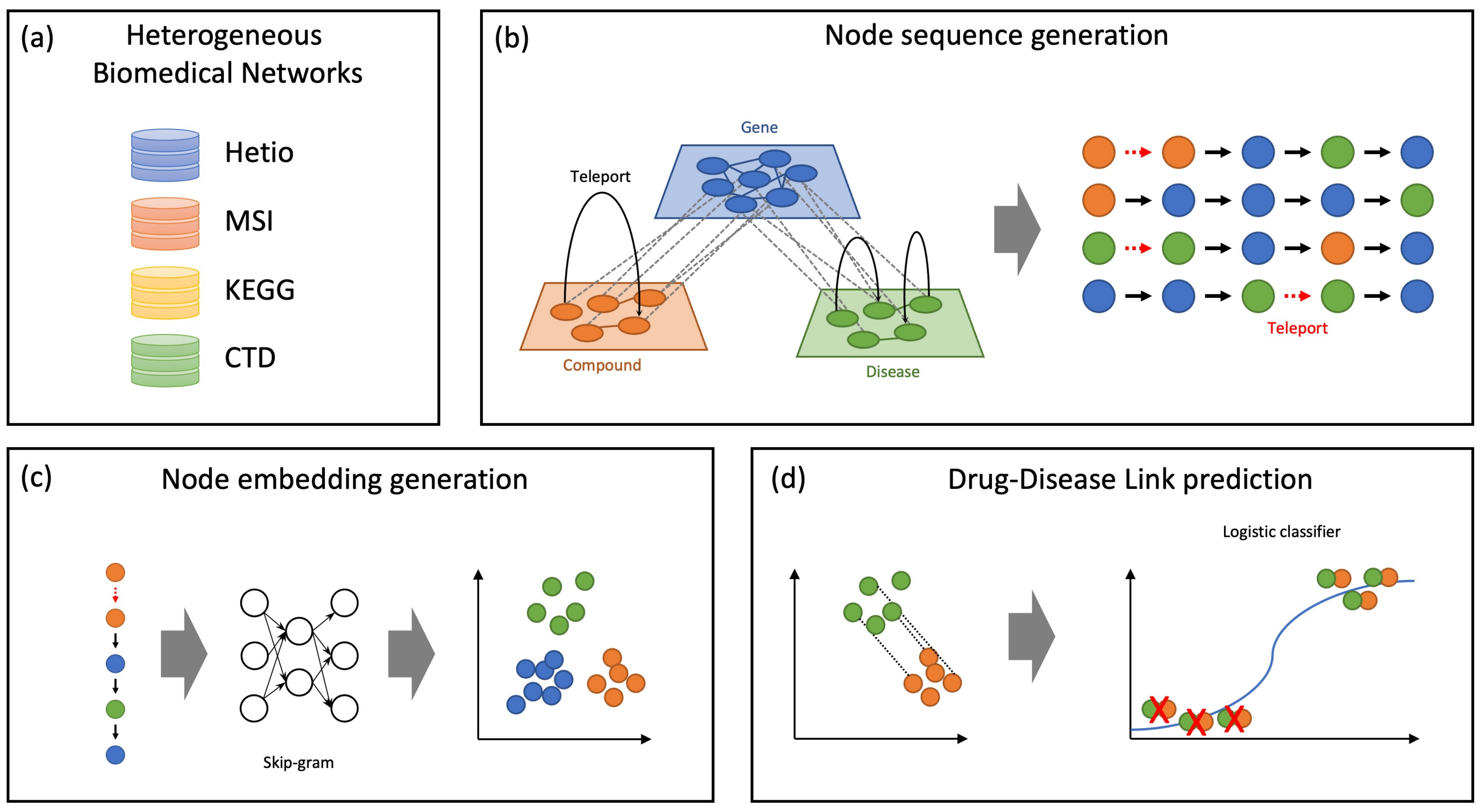
(a) Heterogeneous Biomedical Networks
Hetio
MSI
KEGG
CTD
(b) Node sequence generation
Gene
Teleport
Compound
Disease
Teleport
(c) Node embedding generation
Skip-gram
(d) Drug-Disease Link prediction
Logistic classifier

Figure 1. Workflow of RandomTeleport

Results

- 4 Datasets are used to evalutate the performance of "RandomTeleport".
  - *Hetio Network (Hetio), Multi-scale interactome (MSI), The Comparative Toxicogenomics Database (CTD), Kyoto Encyclopedia of Genes and Genomes (KEGG)*
- For comparison, word2vec and edge2vec model were used as baseline.

Evaluation of prediction performance

- To compare with other drug repurposing tools, we select two tools from four categories introduced in Background.
- Proposed model, "RandomTeleport" outperformed other models in three of the four dataset in accuracy, AUROC, AUPR and F1 score of the predicted drug-disease treatment pair (Figure 2, Table 1).

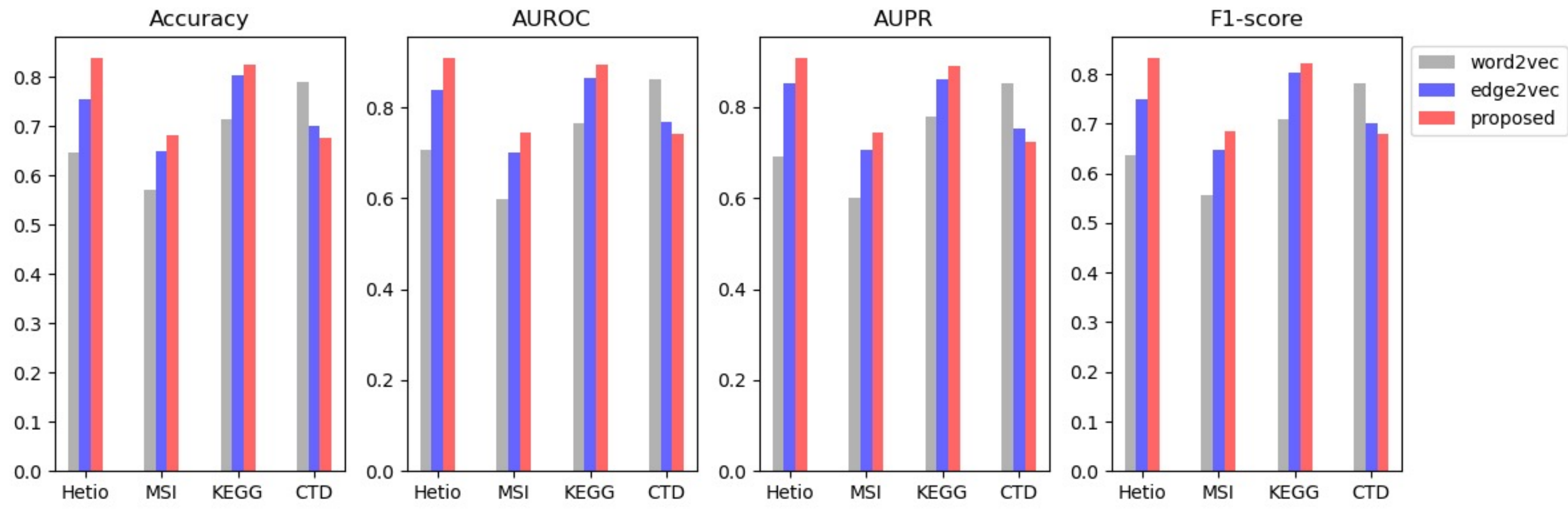

Figure 2. Accuracy, AUROC, AUPR and F1 score of each model on 4 datasets (Hetio, MSI, KEGG, CTD). Proposed model, shown in red, outperformed others in three of four datasets.

	Hetio	MSI	KEGG	CTD
word2vec	0.65	0.57	0.71	**0.79**
edge2vec	0.75	0.65	0.80	0.70
proposed	**0.84**	**0.68**	**0.82**	0.68

Table 1. Accuracy of three drug repurposing models in 4 datasets

Visualization of constructed embedding spaces

- For learning an embedding space consisting of multiple node types, in contrast to baseline word2vec model (Figure 3(a)), proposed model (Figure 3(b)) showed well-clustered and well-separated embedding space according to node's node type (Disease, Compound)
  - Dimension Reduction : Principal Component Analysis (PCA), t-Stochastic Neighbor Embedding (t-SNE)

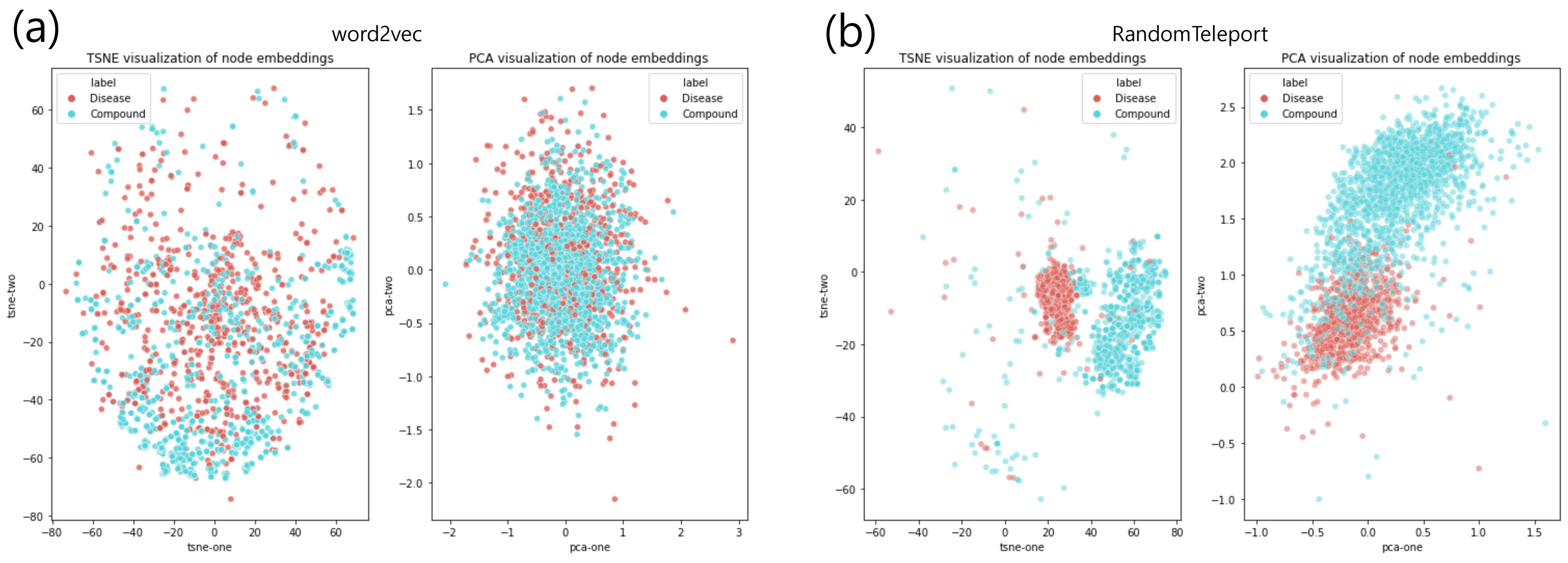

Figure 3. Visualization of embedding space constructed by each model. (a) Embedding space constructed by word2vec model. (b) Embedding space constructed by proposed model, "RandomTeleport"

Conclusion

- By guiding random walk process with tree-structured hierarchy data as teleport probability, the walker generates more biologically meaningful node sequences, enabling efficient node embedding and, in the end, a higher performance in disease-drug link prediction task.
- Technical contribution lies in providing a methology for utilizing tree-structured hierarchy data for guiding machine learning tasks. In RandomTeleporter model, each drug/disease nodes

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117 (1998)
[2] Gao, Z., Fu, G., Ouyang, C. *et al.* edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. *BMC Bioinformatics* 20, 306 (2019).

Bio & Health Informatics Lab, Seoul National University, Seoul 151-744, Korea (http://biohealth.snu.ac.kr)