Gene Embeddings of Complex network (GECo) and hypertension disease gene classification

Cagatay Dursun[†]
Department of Biomedical Engineering
Marquette University – Medical College of Wisconsin
Milwaukee WI USA
cdursun@mcw.edu

Jennifer R. Smith
Rat Genome Database
Department of Biomedical Engineering
Medical College of Wisconsin
Milwaukee WI USA
jrsmith@mcw.edu

G. Thomas Hayman
Rat Genome Database
Department of Biomedical Engineering
Medical College of Wisconsin
Milwaukee WI USA
gthayman@mcw.edu

Serdar Bozdag[†]
Department of Computer Science
Marquette University
Milwaukee WI USA
serdar.bozdag@mu.edu

ABSTRACT

Diseases such as hypertension, cancer, and diabetes are the causes of nearly 70% of the deaths in the U.S. Such complex diseases involve multiple genes and their interactions with environmental factors. Therefore, identification of genetic factors to understand and decrease the morbidity rates of those complex diseases is an important and challenging task. With the generation of an unprecedented amount of multi-omics datasets, network-based methods have become popular to represent the multilayered complex molecular interactions. Particularly network embeddings, the low-dimensional representations of nodes in a network are utilized for gene function prediction. Most of the network embedding methods, however, could not integrate multiple types of datasets from genes and phenotypes. This is an important limitation as multi-omics data integration alleviates the issues due to missing data and lack of context-specific data. To address this limitation, we developed a network embedding algorithm named GECo that can utilize multilayered heterogeneous networks of genes and phenotypes. We evaluated the performance of GECo using genotypic and phenotypic datasets of the model organism *Rattus norvegicus* to classify hypertension disease-related genes. Our method significantly outperformed the state-of-the-art network embedding methods by 94.97% AUC in prediction where the second-best performer achieved 85.98% AUC.

CCS CONCEPTS

Computing methodologies→Learning latent representations; Applied computing→Biological networks; **Computational genomics**; **Bioinformatics**; Systems biology; Transcriptomics

KEYWORDS

Network integration, random walk with restart, multiplex heterogeneous networks, network propagation, graph representation, node embedding, feature learning, multi-omics data integration, genotype to phenotype mapping, disease gene prediction, hypertension, rat

Availability and implementation: The source code is available on GitHub at https://github.com/bozdaglab/GECo under Creative Commons Attribution-NonCommercial 4.0 International Public License.

[†]Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

1. Introduction

Almost two-thirds of the deaths in the U.S. are caused by diseases such as cardiovascular disease, including hypertension, cancer and diabetes [8]. Complex diseases such as these involve interactions of multiple genes with each other and with environmental factors [19]. Identification of these genes in order to dissect these complex diseases and reduce death rates is therefore a vital yet difficult endeavor. Integrative network analysis methods are important to facilitate the understanding of the complexity of multilayered molecular interactions and elucidate the genotype-phenotype relations [4, 5, 13, 35]. Avalilable datasets, such as existing multi-omics datasets might have patterns of missing data across the different datasets that cannot be eliminated by integration [25] and often particular datasets are not specific for the particular complex trait in question, such as protein-protein interactions (PPI) [34]. Supervised learning algorithms allow adjusting the contribution of non-condition specific information in the datasets. Thereore, overcoming those challenges requires supervised learning algorithms to be applied on integrated network data. Learning latent representation of networks [10, 23], allows flexible downstream analysis such as link prediction, community detection and node classification for social and biological networks. Several

approaches to latent representation of networks have been implemented in the social networks domains [10, 21, 23, 32]. Many of the recent studies utilize simple networks for latent representation, and recently many tools have emerged to learn the latent representation of more complex networks to capture the richer information content [1, 7, 36, 37]. Some of these methods rely on random walks to learn the latent features of the nodes in the network inspired from word2vec developed for the natural language processing domain [18]. Random walks is a type of network propagation algorithm where the information disseminates starting from a node to other nodes through the edges of the underlying network. Random walks are effective to capture proximity of nodes to each other since they use the topological structure of networks, and they are scalable for large networks. Word2vec uses the Skip-gram algorithm to learn the word embeddings; its objective function is to minimize the cross-entropy loss of word sequences (sentences). The Skip-gram algorithm relies on the hypothesis that the words that appear more frequently in same contexts are similar to each other [23].

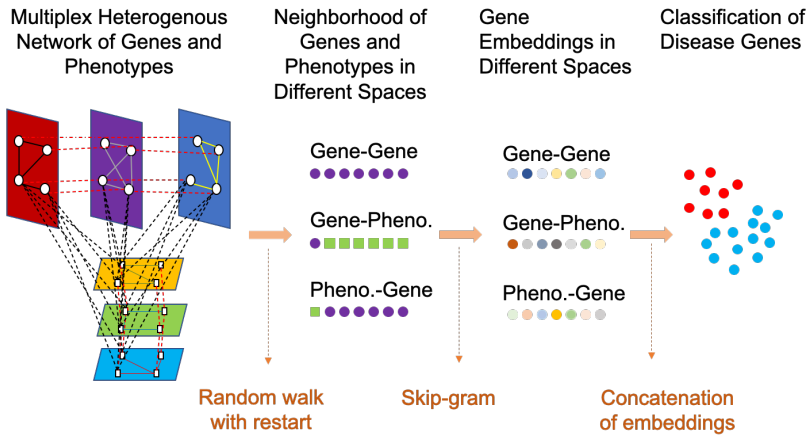

Figure 1. Overview of the GECo framework. GECo first generates the complex multiplex heterogeneous undirected weighted network. Then, it obtains the neighborhood of Gene-Gene, Gene-Phenotype and Phenotype-Gene by random walk with restart. Thirdly, we take the top *N* of those neighborhood and learn the latent representation of nodes using the Skip-gram algorithm. The learned genes embeddings of different spaces then are concatenated and used for classification using a supervised classification algorithm such as GLM.

DeepWalk is a random walk-based node embedding algorithm which utilizes homogeneous networks [23]. It generates node sequences using truncated random walks then utilizes the Skip-gram algorithm to learn the node embeddings. DeepWalk employs a hierarchical softmax as a normalization factor to speed up the process of the Skip-gram. node2vec is another random walk based node embedding algorithm for homogeneous networks [10]. There are two key differences between node2vec and DeepWalk; first, node2vec uses negative sampling as a normalization factor, second, it uses a biased random walk to capture both structural similarities of nodes as well as homophily [11]. Metapath2vec, another random walk-based node embedding algorithm, was developed to address the need for heterogeneous networks [7]. It has two key features; it utilizes metapaths for random walks to guide the random walks and it can apply heterogeneous negative sampling based on the number of node types in the heterogeneous network. These methods do not utilize the rich structure of multiplex heterogeneous networks.

To address those limitations, we developed a network embedding algorithm called Gene Embeddings of a Complex network (GECo) that can utilize multiplex heterogeneous networks. GECo utilizes multiplex gene and phenotype networks to learn the latent feautures of genes for a downstream analysis (Figure 1). First, GECo creates the complex network structure using multiple gene/phenotype layers and a bipartite network of genes and phenotypes. GECo uses a random walk with restart strategy to generate the node neighborhoods. In random walk with restart (RWR), an imaginative particle starting from initial node(s) moves to its immediate neighbor in the network with $1 - r$ probability or restarts from the initial node with probability of $r \in [0, 1]$. GECo generates three types of neighborhoods: i) gene-gene: gene rankings starting from each gene node, ii) gene-phenotype: phenotype rankings starting from each gene node, and iii) phenotype-gene: gene rankings starting from each phenotype node. Then, it learns the embeddings of genes based on each of these neighborhoods utilizing the Skip-gram algorithm. Finally, those different gene embeddings are concatenated and utilized by a supervised classification algorithm for disease-associated gene prediction.

We compare GECo's performance with other approaches on node learning [7, 10, 23]. GECo outperforms the other approaches by about 9% margin. For the top 20 novel predictions we find supporting evidence in the literature, suggesting that GECo performs especially well at identifying disease related gene predictions. Source code of GECo can be accessed on GitHub at https://github.com/bozdaglab/GECo.

2. Materials and Methods

2.1. Random Walk on Multiplex Heterogeneous Network

GECo utilizes a random walk with restart (RWR) algorithm on undirected multiplex heterogeneous networks to compute a node ranking starting from each node in the network and applies the Skip-gram algorithm to learn the latent features of genes in the network based on these node rankings which represent the neighborhood of nodes in the network (Figure 1). Unlike several network embedding algorithms [7, 10, 23] that utilizes simulated

random walks, GECo utilizes the steady state distribution of RWR to generate the neighborhood of the nodes in the network Eq. (1).

$$p_{t+1} = (1 - \mathrm{r})Wp_t + rp_0 \quad (1)$$

The probability distribution vector of nodes at time t is represented by p_t. W is the transition (walk) matrix of the network Eq. (2) calculated by the multiplication of degree (strength) diagonal matrix (D) of the unweighted (weighted) network and A is the adjacency matrix of the network.

$$W = D^{-1}A \quad (2)$$

After a number of steps Eq. (1) reaches a steady state for undirected networks [5]. The magnitude of r affects the convergence rate of the RWR algorithm, large r leads to fast convergence to steady state [15], and to limit the diffusion of the random walk. The steady state distribution (p_s) can be used as a proximity vector for the nodes in the network for a given set of initial nodes. GECo sets $r = 0.7$ by default as in other RWR algorithms [14, 17, 27, 31].

2.2. Node Embedding

The novelty of GECo is both its capability to handle a multiplex heteronegenous network as input, and its utilization of the node proximity of different neighborhood spaces to generate the embeddings of nodes in the network. Unlike metapath2vec GECo handles heterogeneous nodes by dividing them into different neighborhood spaces and then applies regular negative sampling without differentiating the node types in the Skip-gram algorithm. Moreover unlike other random walk based node embedding algorithms, GECo utilizes the stationary distribution of random walks instead of truncated random walks.

GECo utilizes top N nodes of RWR ranks of different neighborhood spaces as a proximity measure of the node sequences (Figure 1). To generate Gene-Gene and Gene-Phenotype embeddings, GECo seperates the node rankings where the starting node is a gene node in the network. Then, GECo uses the top N nodes in these separate neighborhood spaces and utilizes the Skip-gram algorithm to learn the gene embeddings. Similary GECo generates Phenotype-Gene and Phenotype-Phenotype neighborhood spaces by sepearating the node ranks where the start node is a phenotype node. Since the aim of the generated node embeddings is to learn the gene embeddings GECo does not generate phenotype embeddings.

2.3. Complex Network of Rat

To predict hypertension disease related genes in rats, we applied GECo on a complete multiplex heterogeneous rat network. We created a three-layer gene interaction network, namely co-expression, protein-protein interaction and pathway layers. We created a three-layer phenotype network using rat strains, namely mammalian phenotype ontology (MPO) term-based similarity, disease ontology (DO) term-based similarity and quantitative phenotype (QP) measurements-based similarity of strains. We connected the gene multiplex network to the phenotype network based on the MP annotation-based similarity of genes and strains. All the layers were composed of undirected and weighted edges. GECo learns the gene embeddings utilizing this multiplex heterogeneous network.

2.3.1. Gene Network. We created a multiplex gene network using PPI, pathway and co-expression layers.

2.3.1.1. *PPI Layer*. To create the PPI layer, whole physical interactome data for *Rattus norvegicus* were downloaded from the STRING V11 database [30]. For protein pairs having multiple interactions between them, we aggregated their interactions by taking the arithmetic mean of the interaction weights. Protein IDs were mapped to gene IDs by using an alias file in the STRING database. Proteins that were mapped to the same genes were merged into a single node, and the mean of their interaction weights were used as the merged interaction weight.

2.3.1.2. *Pathway Layer*. To create the pathway layer, *Rattus norvegicus* pathway annotation of genes and the pathway ontology tree were downloaded from the Rat Genome Database (RGD) [16]. Semantic similarity of genes was calculated using the ontologyX R package [9]. Ontology based demantic similarity measures the degree of relatedness between two entities by the similarity in meaning of their annotations over a set of ontology terms [24]. We applied a best-match-average approach for term sets. To increase the computational efficiency and decrease the number of uninformative correlations, a maximum possible hard threshold value of 0.5668 was applied to shrink the number of edges without losing any nodes in the layer.

2.3.1.3. *Co-expression Layer* The co-expression layer was based on the RNA-seq expression dataset (GSE50027) [33] of liver for 6 Lyon Hypertensive (LH) and 6 Lyon Normotensive (LN) rat models in the Gene Expression Omnibus (GEO) database [2]. We applied a filtering metric of zFPKM following [12] to filter the active genes by eliminating the background genes in the dataset. We used zFPKM > -3 to select expressed genes as recommended in [12]. We removed the genes having undefined values for four or more samples after zFPKM calculation. We calculated the coexpression weights of remaining genes for normotensive and hypertensive samples using the Pearson correlation. Then, we calculated the differential coexpression weight of genes by calculating the natural log of the ratio of coexpression weights for hypertensive samples to normotensive samples. Finally, we applied a maximum possible hard threshold value of 7.2393 to shrink the number of edges while keeping the network connected.

2.3.2. Strain Network. We created a multiplex phenotype network using MPO, DO and QP measurements-based strain similarity layers.

2.3.2.1. *MP-Based Strain Similarity Layer* The MP-based strain similarity layer represents the strain similarity based on MP annotations of strains provided by RGD [29]. To make this layer more context specific, we computed similarity scores based on hypertension-related MP terms. We determined a list of hypertension-related MP terms (Supplementary Table 1) and calculated a vector that represents each strain based on their MP annotation semantic similarity to the hypertension related MP term vector (See 2.3.1.2). Then we calculated the similarity of strains by dot-product of those semantic similarity vectors. Finally, we applied a maximum possible hard threshold value of 0.1032 to minimize the number of edges while keeping the network connected.

2.3.2.2. *DO-Based Strain Similarity Layer* The DO-based strain similarity layer represents the strain similarity based on DO annotations of strains provided by RGD. Similar to the MP-based strain similarity layer (See 2.3.1.2), to make this layer more context specific, we computed similarity scores based on hypertension-related DO terms. We determined a list of hypertension related DO terms (Supplementary Table 2) and generated the DO-based strain similarity layer using DO annotations of strains. Finally, we applied a maximum possible hard threshold value of 0.0740 to minimize the number of edges while keeping the network connected.

2.3.2.3. *QP Measurements- Based Strain Similarity Layer* To create the QP layer, we downloaded the quantitative phenotype measurements namely systolic blood pressure, heart rate and heart weight annotated to strains in RGD. We used studies having males samples only, to prevent measurement bias between different sexes, since number of females samples were very low. The range of the age of samples we filtered to be 28 to 410 days representing adult rats. We utilized only *in vivo* measurements for systolic blood pressure to prevent measurement bias. We calculated Euclidiean distance of the strains using these three measurements. We took the multiplicative inverse of the distance and applied a maximum possible hard threshold value of 6.1 to minimize the number of edges while keeping the network connected.

2.3.2.4. *Gene-Strain Bipartite Layer* The gene-strain bipartite layer represents the similarity of genes to strains and connects the gene nodes to the strain nodes based on their semantic similarity of MP annotations. We downloaded the gene MP annotations from RGD. To increase the number of covered genes in the bipartite layer we utilized MP annotations of mouse orthologs of rat genes, by downloading mouse MP annotations from MGI [28]. We utilized RGD MP annotation for strains. Then, we calculated the semantic similarity following a strategy similar to the gene pathway layer (See 2.3.1.2).

2.4. Disease Gene Classification Using Gene Embeddings

To evaluate the performance of GECo embeddings, we employed the Generalized Linear Model (GLM) to classify hypertension related genes using gene embeddings. The feature set of each gene was composed of concatenation of embeddings based on Gene-Gene and Gene-Phenotype neighborhood spaces. We employed 10-fold 10-repeat stratified cross validation for performace measurement. We used stratified sampling since the known hypertension disease genes are about 1% of the the number of genes in the network. We used the rat gene disease annotations in RGD to determine the set of "ground truth" hypertension disease related rat genes. We included the hypertension disease-annotated genes which have only non-expression based experimental evidence codes.

2.5. Comparison with Existing Node Embedding Algorithms

We compared GECo with DeepWalk [23], node2vec [10] and metapath2vec [7] node embedding algorithms.

For DeepWalk, we used the node2vec implementation with paramaters for p=1 and q=1 for assessing DeepWalk performance as it is stated the performance difference between hierarchical softmax and negative sampling normalizations in the Skip-gram algorithm is not significant [7]. Since DeepWalk can only utilize a homogeneous network, we generated an aggregated gene network of three gene layers by taking the geometric mean of the edge weights and keeping all the genes in the network. DeepWalk has several hyperparameters; the number of walks controls the number of runs of random walk per node, walk length controls the number of time that random walker will hop per random walk, context size controls the window size that the Skip-gram algorithm maximizes the co-occurrence probability among the nodes that appear within, embedding size controls the latent feature vector of a node that will be learned by Skip-gram algorithm. We ran DeepWalk for all combinations of key hyperparameter value choices (i.e., number of walks per node: {10, 20}, walk length: {80, 100}, context size: {5, 20} and embedding size: {128, 256, 512}).

We utilized the same grid search for the common hyperparameters of node2vec and DeepWalk. We also input the same aggregated gene network to node2vec. For p and q values we used grid search for the interval of {0.25, 0.5, 1, 2}. p and q values controls the biased random walk; p controls the likelihood of the walk immediately revisiting a node, while q controls the likelihood of the walk revisiting local nodes.

For evaluating the performance of metapath2vec, we generated a metapath for the utilization of bipartite relations between genes and strains by guiding the random walks of metapath2vec using the gene-strain-gene scheme. We input the unweighted gene-strain bipartite network to the metapath2vec algorithm as it can only utilize unweighted bipartite networks. Number of walks was set to 1000, and walk length was set to 100. Embedding sizes used were

the same with DeepWalk and node2vec, for context size {5, 7, 10, 15} values were used as window sizes.

A grid search was made for all algorithms for the given hyperparameters.

Table 2. The numbers of nodes and edges in each layer of the multiplex rat network.

Layer	Number of Nodes	Number of Edges
PPI	16,244	762,198
Pathway	6,016	665,344
Differential co-expression	11,510	103,294
MP	805	323,610
DO	1,124	252,363
QP	244	5,697

3. Results

In this study, we developed a network representation learning algorithm, GECo that uses multiplex gene and phenotype networks to classify disease-associated genes. To assess GECo, we applied it on *Rattus norvegicus* datasets to predict hypertension disease-related genes. A multiplex gene network was generated using differential co-expression, PPI and pathway layers. A multiplex phenotype network was generated using MP, DO and QP strain layers. We applied a grid search for different hyperparameters (top *N*, embedding size, context size) for each embedding space across different network combinations.

We built the gene and phenotype network layers using publicly available expression, pathway, PPI and rat strain annotation datasets (see 2.3). The number of nodes and edges of each layer in the final multiplex heterogeneous network is shown in Table 2.

3.1. GECo Outperforms Existing Network Embedding Algorithms

We ran DeepWalk, node2vec and metapath2vec node embedding algorithms on rat data to compute gene embeddings and compared the hypertension disease gene classification performances by GLM.

Table 1 shows the comparison of the other algorithms' best performing results with 10 runs. GECo consistently outperformed the closest performer algorithm by generating about 9% higher area under the receiving operating characteristic curve (AUC).

3.2. GECo is Robust to Hyperparameters

We investigated the effects of top *N*, embedding size and window size parameters of GECo on the gene disease classification performance. We studied the parameter sensitivity of GECo as measured by the classification performance using AUC. We utilized a multiplex three-layer gene and three-layer phenotype network structure. Figure 2 shows the classification performance as a function of each of the three parameters when fixing the other two for three different neibhorhoods namely Gene-Gene, Gene-Phenotype and Phenotype-Gene.

Table 1. Mean AUC values with standard deviations are given for 10 runs for each embedding. For the GLM classification 10-fold cross-validation was performed 10 times.

Gene	Mean AUC	Std. Dev.
GECo	94.97	0.24
Node2vec	85.98	0.54
DeepWalk	85.21	0.85
Metapath2vec	82.33	0.05

Figure 2. A-B shows that top N smaller than or equal to 350 does not alter the gene classification performance much when Gene-Gene and Gene-Phenotype embedding spaces were utilized, and generated the highest performance. On the other hand for the Phenotype-Genotype embedding space there seems to be a slight variation in classification performance and top *N* 500 generates the highest classification performance (Figure 2. C).

We fixed embedding size at 150 for all neighborhoods, and fixed context size at 5 for Gene-Gene, and 50 for Phenotype-Gene, for the Gene-Phenotype neighborhood we set context size to top *N* value. We observed that the classification performance had a peak value for the Gene-Gene embedding space when the context size was 5 (Figure 2.D). For the Gene-Phenotype neighborhood, as we had one gene in the neighborhood, performance peaked when the context size converged to or equals to top *N* (Figure 2.E). This result was expected as we kept the gene (starting node for random walk) in the context where we learned the gene embedding with the phenotypes in the neigbhborhood. To learn gene embeddings using the Gene-Phenotype neighborhood requires the use of larger context sizes for the Skip-gram algorithm; the best performing context sizes happen to be the ones close to or equal to top *N*, which enables the Skip-gram to learn the gene embedding using more phenotypes in the neighborhood. For the Phenotype-Gene neighborhood when context size became 50, the classification performance started to converge (Figure 2.F).

We observed that different embedding sizes did not have much impact on classification performance (Figure 2.G). For the Gene-Phenotype embedding space embedding size of 200 and larger embedding sizes had similar performance results (Figure 2.H). For the Phenotype-Gene embedding space, embedding size of 150 had the top performance. We concluded that GECo is robust to hyperparameters for Gene-Gene and Gene-Phenotype embedding spaces, but for the Phenotype-Genetoype embedding space there is

a variation in classification performance with respect to different hyperparameter values.

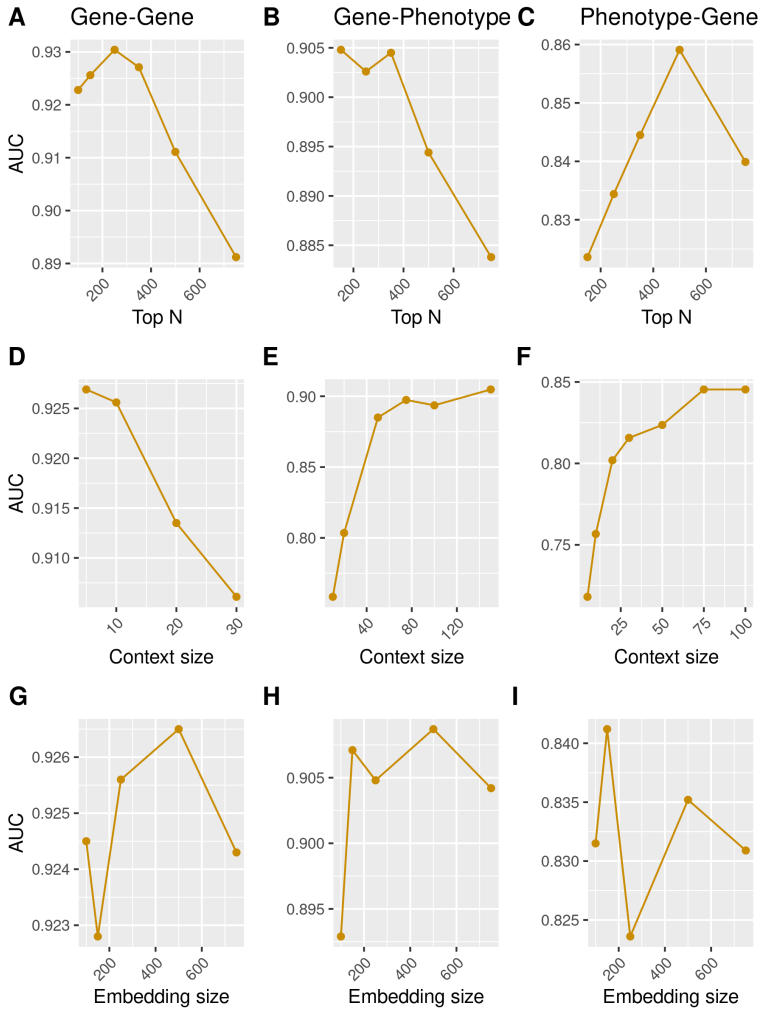

Figure 2. Parameter sensitivity of GECo. For each parameter evaluation the other two parameters are fixed. Panels by columns show the parmater plots by different embedding spaces.

3.3. Gene-Gene and Gene-Phenotype Embedding Spaces Generate Higher Gene Classification Performance Than Phenotype-Gene Embedding Spaces

Since we utilized multiple gene and strain datasets, we analyzed the contribution of single embedding spaces to the classification results across different network combinations (Figure 4). Gene-Gene embeddings performed better than Gene-Phenotype embeddings when the gene network was multiplex. When the gene and phenotype network were aggregated, then the Gene-Gene and Gene-Phenotype embeddings' performances were similar. Moreover, clearly Gene-Gene and Gene-Phenotype embedding spaces generated higher AUC values compared to Phenotype-Gene embedding space for gene classification.

3.4. Multiplex Gene Networks Generate Higher Gene Classification Performance Than Aggregate Networks

We investigated the effect of multiplex networks of genes and phenotypes on the classification performance (Figure 3 and Figure 5). We observed that single Gene-Gene embeddings performed better when the gene network was multiplex (Figure 3). On the other hand, for single Gene-Phenotype embeddings we did not observe a difference when the multiplex gene network was used. Overall, when we concatenated Gene-Gene and Gene-Phenotype embeddings we observed higher AUC scores for multiplex gene networks. The range of mean AUC for the four multiplex gene networks in Figure 3 was [92.69 - 94.85], while the mean AUC range for aggregated gene networks was [90.71 - 93.56].

We did not observe a similar trend for multiplex phenotype networks compared to aggregated phenotype networks. There is no statistically significant difference between Gene-Phenotype embeddings' performances (Figure 5. A, B, D). For one configuration Gene-Phenotype embeddings performed better where aggregated network's performance was higher than the multiplex (Figure 5. C). Gene-Gene embeddings showed mixed signals, for one case Gene-Gene embedding for aggregated phenotype network performed better (Figure 5. A), for another case mulplex phenotype network performed better (Figure 5. B), and for others there was not statisticall significant difference (Figure 5. C-D). Multiplex network embeddings performed better as is expected when the topology of the networks are different from each other [6]. While the gene layers that we utilized have different topological structures the phenotype networks were more similar to each other compared to gene layers. Moreover, we observed that the variation of layers

on the phenotype side did not have as large an impact on gene embeddings as the variation of layers on the gene side.

3.5. PPI Layer has the Largest Contribution to Classification Performance

We analyzed the effects of the gene and phenotype layers to evaluate their individual contributions to the gene classifications. When we omit the PPI layer from multiplex gene-phenotype networks, the classification AUC for the concatenated Gene-Gene and Gene-Phenotype embeddings dropped to 92.69 ($\pm$0.38) from 94.85 ($\pm$0.29). For the co-expression layer it dropped to 94.42 ($\pm$0.30), and for the pathway layer it dropped to 93.99 ($\pm$0.44), suggesting the contribution from the co-expression layer is the smallest. We observed a similar trend for the single embedding space classification performances (Figure 3). There is no clear sign for the contribution of phenotype layers to gene embeddings. We observed a drop in AUC for single Gene-Phenotype embeddings when we omit the single phenotype layers from the reference multiplex which have three phenotype layers (Figure 5). But, suprisingly we observed an increase in AUC for single Gene-Gene embeddings. Overall, GECo generated highest AUC score of 94.97 ($\pm$0.24) using an aggregated multiplex gene network of three layers and an aggregated phenotype network of MP and QP layers.

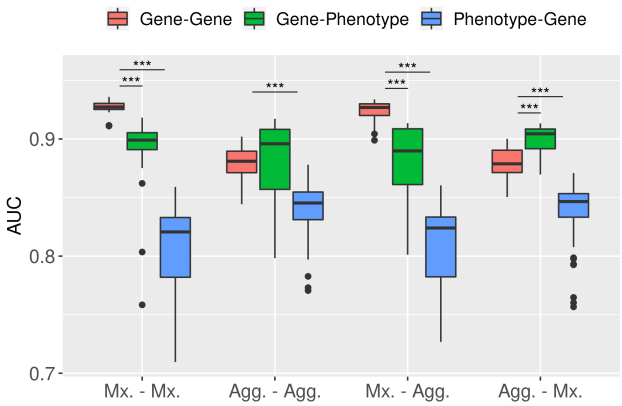

Figure 4. Contribution of gene-gene, gene-phenotype and phenotype-gene embedding spaces across different network combinations which utilize all three multiplex (Mx.) or aggregated (Agg.) gene and phenotype networks. Aggregated networks are generated by taking the geometric average of the overlapping edges. Groups are shown in Gene-Phenotype network format. ***: $p \leq 0.001$

3.6. GECo Predicts Novel Hypertension Genes

Our goal is to evaluate the quality of novel GECo's predictions of hypertension genes by performing a literature-based evaluation of new predictions. To do this we utilized the best seven results of different network combinations (Supplementary Table 3). We ranked the genes for ten-runs predictions of those seven network combinations by their prediction probability. We then calculated the final rank based on the median ranks of each gene across these seventy results. We filtered the known hypertension genes based on our ground truth set of hypertension disease related genes. We

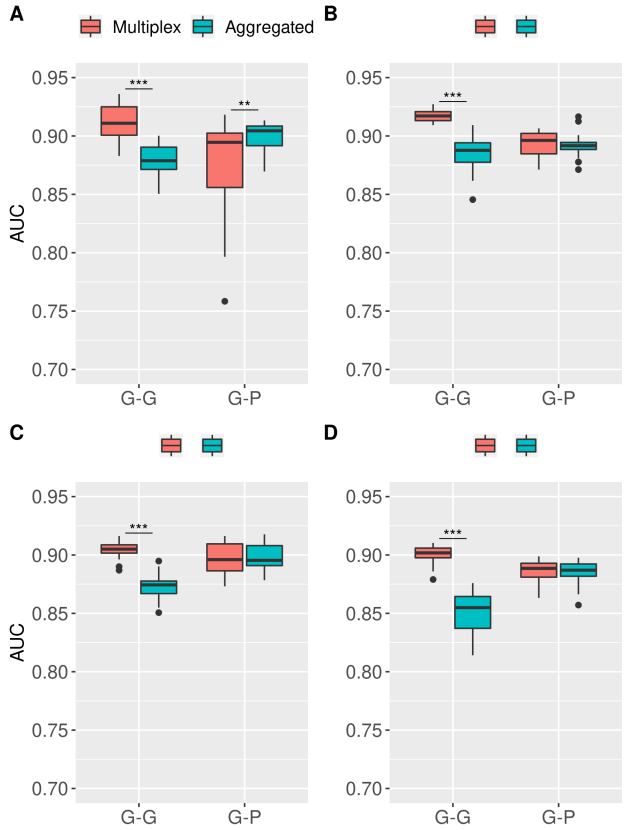

Figure 3. Effect of multiplex vs aggregated gene networks on the gene classification performance. All eight networks have multiplex networks of three phenotype layers and multiplex vs aggregated layers of A) (PPI, PWY, CO-EXPR), B) (PPI, PWY), C) (PPI, CO-EXPR), and D) (PWY, CO-EXPR). Aggregated networks are generated by taking the geometric average of the overlapping edges. PWY: Pathway, CO-EXPR: Co-expression. **: $p \leq 0.01$, ***: $p \leq 0.001$

investigated the top 20 novel predictions for genes that at the time were not annotated for hypertension-related disease at RGD. Performing new curation of these genes, we found that 18 of GECo's top predictions have published evidence, in most cases not merely expression-based, that the genes encode hypertension related proteins (Supplementary Table 4). We could not find adequate evidence for two of the top 20 predicted genes. It is thought that gene Pla2g10 may be involved in a signaling axis regulating blood pressure homeostasis [3], but we could not find any specific hypertension-related annotations for this protein. While we were unable to find evidence that gene Ckmt2 is related to hypertension, we did locate evidence that it is involved in cardiovascular disease [26].

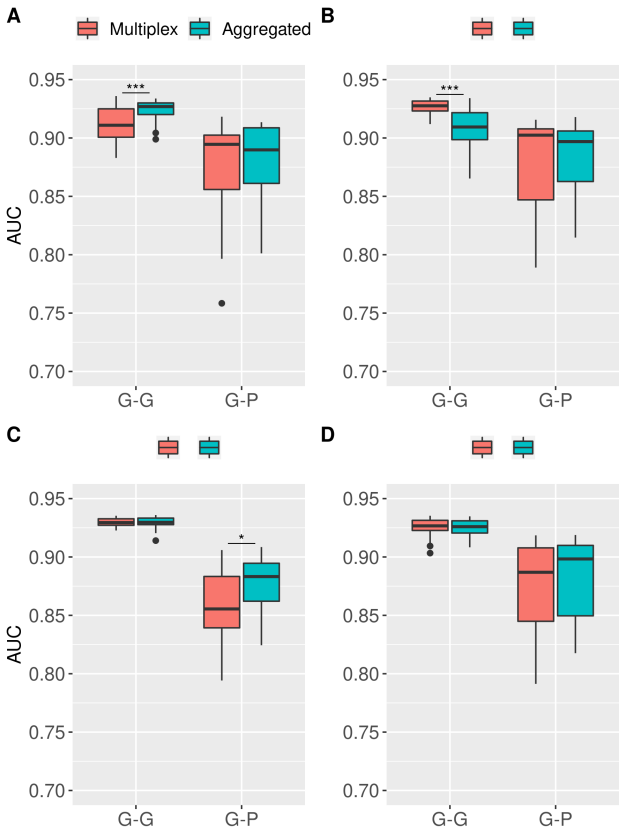

Figure 5. Effect of multiplex vs aggregated phenotype networks on the gene classification performance. All eight networks have multiplex network of three gene layers and multiplex vs. aggregated layers of A) (MPO, DO, QP), B) (MPO, QP), C) (DO, QP), D) (MPO, DO). Aggregated networks are generated by taking the geometric average of the overlapping edges. *: $p \leq 0.05$, ***: $p \leq 0.001$

4. Discussion

In this study, we developed a gene embedding tool called GECo that can utilize multiplex heterogeneous networks of genes and phenotypes. GECo uses stationary node ranks of RWR as proximity measure of nodes. GECo divides the node ranks into different neighborhood spaces, and then applies the Skip-gram algorithm to generate the node embeddings. GECo did not generate phenotype embeddings for the current experimental study, but it is straightforward to generate phenotype embeddings for a different problem setting.

To evaluate the performance of GECo, we applied it to a rat dataset to classify hypertension disease-related genes. We compared GECo to the state-of-the-art node embedding methods DeepWalk, node2vec and metapath2vec. GECo achieved a 94.97% AUC, and outperformed the closest performer by about 9% margin.

GECo has a number of hyperparameters; top *N* is the first *N* number of nodes in the ranked nodes after RWR generates a stationary node rank, context size is the window size for the Skip-gram algorithm, and embedding size is the length of the node embedding vector generated by the Skip-gram algorithm. We observed that GECo is robust to hyperparamaters especially for Gene-Gene and Gene-Phenotype embedding spaces (Figure 2).

We observed that Gene-Gene and Gene-Phenotype embedding spaces generated better gene disease classification performance compared to the Phenotype-Gene embedding space. We also analyzed the effect of multiplex network layers on classification performance. Multiplex gene networks performed better than the aggregated gene network. On the other hand, aggregated phenotype networks performed slightly better than the multiplex phenotype networks. This could be partly due to the incomplete nature of gene networks, and partly due to the use of gene embeddings where phenotype networks did not have higher impact.

As we integrated multiple gene and phenotype layers, we analyzed the contribution of each layer to the gene disease classsifcation performance. We used the multiplex three-layer gene and three-layer phenotype network as a reference network. The PPI layer had the largest effect, and the co-expression had the least contribution to the gene classification performance among gene layers in our experiments. When we removed single phenotype layers from the reference network, we observed a drop in gene classification performance when Gene-Phenotype embeddings were utilized, but this drop was suprisingly offset by an increase in Gene-Gene embeddings performance.

As a demonstration of the GECo's effectiveness, all genes in the network were ranked as to the likelihood of the genes being associated with hypertension-related diseases. Then, we filtered out the genes that were in the "ground truth" gene set from the rank list. Preliminary disease curation was then performed by RGD curators of the 20 highest ranking genes (See 3.6). For 18 of these genes, evidence for linkage to hypertension was found, usually from multiple sources and of a type not involving mere correlative gene expression change. The other two genes could be potentially novel hypertension-related genes. Of these two genes, one gene, Pla2g10, had suggestive evidence of an association with blood pressure regulation. And while the other gene, Ckmt2, could not be linked directly to hypertension based on current knowledge, there was evidence of an association with other cardiovascular diseases. Given that hypertension is well established as both a risk factor for and comorbidity of cardiovascular disease [20, 22], this suggests that Ckmt2 may be involved in processes that could influence either or both of these disease states The high level of success achieved by GECo in predicting linkage to the disease hypertension for a set of genes where no linkage had previously been curated in a model organism database is strong evidence of the tool's power and utility.

ACKNOWLEDGMENTS

This study was funded by grant HL64541 from the National Heart, Lung, and Blood Institute on behalf of the NIH (to RGD) and by grant R35GM133657 from National Institute of General Medical Sciences of the National Institutes of Health. This manuscript is dedicated to the memory of Prof. Mary Shimoyama who passed suddenly away suddenly on Feb. 19, 2020. The authors would like to express their appreciation to RGD curators for their contribution to novel predictions section, and Dr. Arjun Krishnan for fruitful discussions.

REFERENCES

[1] Bagavathi, A. and Krishnan, S. 2019. Multi-Net: A Scalable Multiplex Network Embedding Framework. *Complex Networks and Their Applications VII* (2019), 119–131.

[2] Barrett, T. et al. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*. 41, D1 (Jan. 2013), D991–D995. DOI:https://doi.org/10.1093/nar/gks1193.

[3] Berry, E. et al. 2015. Matrix metalloproteinase-2 negatively regulates cardiac secreted phospholipase A2 to modulate inflammation and fever. *Journal of the American Heart Association*. 4, 4 (Mar. 2015). DOI:https://doi.org/10.1161/JAHA.115.001868.

[4] Cho, D.-Y. et al. 2012. Chapter 5: Network Biology Approach to Complex Diseases. *PLoS Computational Biology*. 8, 12 (2012), e1002820. DOI:https://doi.org/10.1371/journal.pcbi.1002820.

[5] Cowen, L. et al. 2017. Network propagation: A universal amplifier of genetic associations. *Nature Reviews Genetics*. 18, 9 (2017), 551–562. DOI:https://doi.org/10.1038/nrg.2017.38.

[6] Didier, G. et al. 2015. Identifying communities from multiplex biological networks. *PeerJ*. 3, (Dec. 2015), e1525. DOI:https://doi.org/10.7717/peerj.1525.

[7] Dong, Y. et al. 2017. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (Halifax, NS, Canada, 2017), 135–144.

[8] Eyre, H. et al. 2004. Preventing Cancer, Cardiovascular Disease, and Diabetes. *Diabetes Care*. 27, 7 (Jul. 2004), 1812. DOI:https://doi.org/10.2337/diacare.27.7.1812.

[9] Greene, D. et al. 2016. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*. 33, 7 (Dec. 2016), 1104–1106. DOI:https://doi.org/10.1093/bioinformatics/btw763.

[10] Grover, A. and Leskovec, J. 2016. node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (San Francisco, California, USA, 2016), 855–864.

[11] Hamilton, W.L. et al. 2017. Inductive Representation Learning on Large Graphs. (2017), 19.

[12] Hart, T. et al. 2013. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*. 14, 1 (2013), 778. DOI:https://doi.org/10.1186/1471-2164-14-778.

[13] Kim, Y.-A. et al. 2016. Understanding Genotype-Phenotype Effects in Cancer via Network Approaches. *PLOS Computational Biology*. 12, 3 (2016), e1004747. DOI:https://doi.org/10.1371/journal.pcbi.1004747.

[14] Köhler, S. et al. 2008. Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*. 82, 4 (2008), 949–958. DOI:https://doi.org/10.1016/j.ajhg.2008.02.013.

[15] Langville, A.N. and Meyer, C.D. 2011. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

[16] Laulederkind, S.J.F. et al. 2018. A Primer for the Rat Genome Database (RGD). *Methods in Molecular Biology (Clifton, N.J.)*. 1757, (2018), 163–209. DOI:https://doi.org/10.1007/978-1-4939-7737-6_8.

[17] Li, Y. and Patra, J.C. 2010. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*. 26, 9 (2010), 1219–1224. DOI:https://doi.org/10.1093/bioinformatics/btq108.

[18] Mikolov, T. et al. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. (Sep. 2013).

[19] Motulsky, A.G. 2006. Genetics of complex diseases. *Journal of Zhejiang University. Science. B*. 7, 2 (Feb. 2006), 167–168. DOI:https://doi.org/10.1631/jzus.2006.B0167.

[20] Oparil, S. et al. 2018. Hypertension. *Nature Reviews Disease Primers*. 4, (Mar. 2018), 18014. DOI:https://doi.org/10.1038/nrdp.2018.14.

[21] Ou, M. et al. 2016. Asymmetric Transitivity Preserving Graph Embedding. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), 1105–1114.

[22] Patel, P. et al. 2018. Standardized Hypertension Management to Reduce Cardiovascular Disease Morbidity and Mortality Worldwide. *Southern Medical Journal*. 111, 3 (2018), 133–136. DOI:https://doi.org/10.14423/SMJ.0000000000000776.

[23] Perozzi, B. et al. 2014. DeepWalk: online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (New York, New York, USA, 2014), 701–710.

[24] Pesquita, C. et al. 2009. Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*. 5, 7 (Jul. 2009), e1000443. DOI:https://doi.org/10.1371/journal.pcbi.1000443.

[25] Ritchie, M.D. et al. 2015. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*. 16, 2 (2015), 85–97. DOI:https://doi.org/10.1038/nrg3868.

[26] Shendre, A. et al. 2017. Admixture Mapping of Subclinical Atherosclerosis and Subsequent Clinical Events Among African Americans in 2 Large Cohort Studies. *Circulation. Cardiovascular Genetics*. 10, 2 (Apr. 2017). DOI:https://doi.org/10.1161/CIRCGENETICS.116.001569.

[27] Smedley, D. et al. 2014. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*. 30, 22 (2014), 3215–3222. DOI:https://doi.org/10.1093/bioinformatics/btu508.

[28] Smith, C.L. and Eppig, J.T. 2009. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*. 1, 3 (2009), 390–399. DOI:https://doi.org/10.1002/wsbm.44.

[29] Smith, J.R. et al. 2020. The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic acids research*. 48, D1 (Jan. 2020), D731–D742. DOI:https://doi.org/10.1093/nar/gkz1041.

[30] Szklarczyk, D. et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 47, D1 (Jan. 2019), D607–D613. DOI:https://doi.org/10.1093/nar/gky1131.

[31] Valdeolivas, A. et al. 2018. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*. 35, 3 (Jul. 2018), 497–505. DOI:https://doi.org/10.1093/bioinformatics/bty637.

[32] Wang, D. et al. 2016. Structural Deep Network Embedding. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), 1225–1234.

[33] Wang, J. et al. 2015. Systems biology with high-throughput sequencing reveals genetic mechanisms underlying the metabolic syndrome in the Lyon hypertensive rat. *Circulation. Cardiovascular Genetics*. 8, 2 (Apr. 2015), 316–326. DOI:https://doi.org/10.1161/CIRCGENETICS.114.000520.

[34] Wang, S. et al. 2018. Typing tumors using pathways selected by somatic evolution. *Nature Communications*. 9, 1 (Oct. 2018), 4159. DOI:https://doi.org/10.1038/s41467-018-06464-y.

[35] Yan, J. et al. 2018. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*. 19, 6 (Nov. 2018), 1370–1381. DOI:https://doi.org/10.1093/bib/bbx066.

[36] Zhang, H. et al. 2018. Scalable Multiplex Network Embedding. (2018), 3082–3088.

[37] Zitnik, M. and Leskovec, J. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*. 33, 14 (Jul. 2017), i190–i198. DOI:https://doi.org/10.1093/bioinformatics/btx252.