Title	Synergy between embedding and protein functional association networks for drug label prediction using harmonic function
Author(s)	Timilsina, Mohan; Mc Kernan, Declan Patrick; Yang, Haixuan; d'Aquin, Mathieu
Publication Date	2020-10-16
Publication Information	Timilsina, Mohan, Mc Kernan, Declan Patrick, Yang, Haixuan, & d'Aquin, Mathieu. (2020). Synergy between embedding and protein functional association networks for drug label prediction using harmonic function. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). doi:10.1109/TCBB.2020.3031696
Publisher	ACM and IEEE
Link to publisher's version	https://dx.doi.org/10.1109/TCBB.2020.3031696
Item record	http://hdl.handle.net/10379/16447
DOI	http://dx.doi.org/10.1109/TCBB.2020.3031696

Synergy Between Embedding and Protein Functional Association Networks for Drug Label Prediction using Harmonic Function

Mohan Timilsina, Declan Patrick Mc Kernan, Haixuan Yang, and Mathieu d'Aquin

Abstract—Semi-Supervised Learning (SSL) is an approach to machine learning that makes use of unlabeled data for training with a small amount of labeled data. In the context of molecular biology and pharmacology, one can take advantage of unlabeled data. For instance, to identify drugs and targets where a few genes are known to be associated with a specific target for drugs and considered as labeled data. Labeling the genes requires laboratory verification and validation. This process is usually very time consuming and expensive. Thus, it is useful to estimate the functional role of drugs from unlabeled data using computational methods. To develop such a model, we used openly available data resources to create (i) drugs and genes, (ii) genes and disease, bipartite graphs. We constructed the genetic embedding graph from the two bipartite graphs using Tensor Factorization methods. We integrated the genetic embedding graph with the publicly available protein functional association network. Our results show the usefulness of the integration by effectively predicting drug labels.

Index Terms—Label Propagation, Networks, Prediction, Embeddings, Harmonic

1 Introduction

The genome-wide identification of all target proteins of drug candidate compounds is a demanding issue in drug discovery. Researchers in pharmaceutical science assessed a tremendous amount of protein groups and developed methods for analyzing essential targets. Understanding the molecular biology of the protein is crucial for designing specific and selective inhibitors or ligands to adjust protein activity. The early recognition of protein activity, and active site information through the identification of selective drug targets can be cost-effective measures in the drug discovery process.

Currently, recognizing drug-target interactions has dramatically escalated in drug development. The publicly available drug databases, such as DrugBank and KEGG, contain experimentally verified information about drug-target interactions [1]. This known information to identify the drugs and the targets using in silico method, which reduces the time and cost of drug development. In recent years network-based analysis has brought considerable attention in drug repositioning to decrease the cost of new drug development [2]. The network-based methods are well explored to understand the network of drugs, disease, genes, and drug side-effects [3].

The functional classification or node classification on networks, also known as a collective classification, has been one of the most active and influential research fields in Artificial Intelligence (AI) [4], [5]. It is due to semi-supervised learning require less human interference and gives higher certainty. There are different variants of graph-based label propagation algorithms proposed that can be applied for node classification problems [4], [5], [6], [7], [8], [9].

Similarly, in recent years, along with the topology-based methods, network embedding methods [10], [11] has drawn significant attention in node feature learning from the graphs. These methods are successfully applied in pharmacological studies. Such methods have shown promising results in polypharmacy [12] and drug side-effect prediction [3]. While both the topology-based methods and embedding methods claim encouraging performances in some applications, a combination of both methods has drawn little attention. Such a combination is especially useful when there are heterogeneous data available that can provide complementary information for a given task. In this paper, we focus on the problem of drug label prediction by integrating three heterogeneous networks of two bipartite graphs of drug-gene interactions and tumor samples-gene association and one protein functional association graph. We expect to achieve such a task by a novel integrative method by employing both topology-based methods and network embedding methods. Thereby the two bipartite graphs are transformed into a "homogeneous" graph for a natural combination with the protein functional association graph.

2 The present study

The focus of our study is on a prediction of the "Mechanism of Action (MOA)" of the drugs. MOA refers to the drug-

- *Mohan Timilsina and Mathieu d'Aquin are with the Data Science Institute, National University of Ireland Galway, Ireland.*
  E-mail: mohan.timilsina@insight-centre.org,mathieu.daquin@nuigalway.ie.
- *Declan Patrick Mc Kernan is with the Department of Pharmacology and Therapeutics, National University of Ireland Galway,Ireland.*
  E-mail: declan.mckernan@nuigalway.ie.
- *Haixuan Yang is with the School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway,Ireland*
  E-mail: haixuan.yang@nuigalway.ie.

binding capacity or interaction to the same biological targets [13] proteins. Every drug has molecular or biological targets to which the drug binds, such as receptors or enzymes. Receptors, activities comprise of *agonist, antagonist, inverse agonist, or modulator*, while for enzyme includes *activator or inhibitor*. Ion channel modulators consist of *opener or blocker*. The prediction of these activities is crucial because it can guide better drug development and help to prevent late-stage drug failures [14].

MOA of drugs can be determined by (i) Microscopy (ii) Biochemical and (iii) Computational methods [15]. The first two methods are expensive and time-consuming as it is tedious to conduct experiments and interpret data manually. Thus the computational method can be useful to systematically and quickly generate a few hypotheses. Therefore these hypotheses can be tested for later laboratory validation and verifications. With the application of machine learning techniques, the computational model learns patterns of drugs and target genes from data and then predicts target genes of existing or new drugs.

The potential drug targets that the pharmaceutical industry can exploit are apprehended in the intersection between the druggable genome and those genes related to disease [16]. The encoding of the proteins in a shared space between drugs and disease can be extracted using embedding methods. The embedding graphs and protein functional association graph are two different types of information complement each other, which we have investigated using the following research questions (RQ):

- *RQ1: Does an integrative approach of combining genetic embeddings with the combined protein interactions network improve prediction of the Drug MOAs?*
- *RQ2: Do all protein interactions network provide similar accuracy for the prediction of the Drug MOAs?*
- *RQ3: Can genetic embeddings graph perform better than the individual protein functional association graph for predicting MOA?*

3 Datasets

3.1 Tumor-Gene data

The tumor is a disease caused by abnormal cell cycle and linked to a series of changes in the activity of genes. We used COSMIC (Catalogue Of Somatic Mutations In Cancer) Methylation Data in our analysis because the DNA methylation is considered an excellent target for anticancer therapies and the drugs which are targeted for DNA methylated gene have been developed to increase efficacy, stability and to decrease toxicity [17]. The epigenetic modifications such as DNA methylation alter gene expression at the level of transcription by upregulating, downregulating, or silencing genes completely. Therefore, recognizing drugs MOA's for epigenetic modifications are of great clinical interest [18].

The COSMIC[1] database uses the expert-curated information of somatic mutations in human cancers and is freely available. The processed data has the fields: **id, sample name, location**, and **gene names**. Each sample name is a tumor sample of the patient extracted from a particular location of the body; for instance, ”TCGA-CV-A6JN-01” is a tumor sample and anatomical position is ”Upper Aerodigestive Tract.” From this data, we are interested in tumor samples and gene names. The tumor samples are taken from ten different anatomical locations. The edges between the tumor samples and genes are based on the fact reported in the cosmic differential methylation data. The gene names used in the methylation data are the accepted HGNC[2] (HUGO Nomenclature Committee) identifier that gives the unique gene symbols and names for the human loci. We labeled this relationship as ”hasGene” for example, Tumor-[”hasGene”]-Genes where tumor and genes are the nodes, and ”hasGene” is the edge type. Hence, we constructed a Tumor sample and Gene bipartite graph.

3.2 Drug-Gene data

For the drug-gene data, we used the Drug-Gene Interaction Database (DGIdb). The interaction types describe the MOA between a small molecule and a protein. The term ”MOA” and ”interaction” are interchangeably used in this study. In DGIdb, a drug-gene interaction is defined as ”a known interaction (e.g., inhibition) between a known drug compound (e.g., lapatinib) and a target gene (e.g., EGFR).” This database is a publicly available druggable genome resource [19]. DGIdb has improved its usefulness as a resource for mining clinically actionable drug targets using expert curation and mined from multiple resources such as DrugBank, therapeutic target database (TTD), PharmGKB, and ClinicalTrials.gov. DGIdb acts as resources to generate hypotheses for the mutated genes that might be therapeutic targets or prioritized for anticancer drug development [20]. From this database, we queried drugs for the genes that have an association with Tumors from our tumor-gene bipartite graph using HGNC gene symbols and DGIdb API[3]. The term *interaction type* between genes and drugs, used by DGIdb are based largely on literature mining and obtained from existing publicly available reviews and databases [20]. We extracted the *drugs name* and *interaction type* that specify how the drug interacts with the gene. We labeled each interaction type of drugs as the drug functions. There are seven different types of interaction we observed in our graph, namely, *Blocker, Antagonist, Agonist, Activator, Inhibitor, Channel Blocker*, and *Binder*. The drug and gene nodes are connected by the ”target” relationship. The set of known *interaction type* is noted as the ’gold standard’ data in this study, and is used for evaluating the performance of the semi-supervised machine learning algorithm in the cross-validation experiments as well as training data in the prediction. The detailed summary of the nodes and edges after the construction of the graph is in Table 1.

3.3 Gene-Gene interaction data

For the Gene-Gene interaction data, we used publicly available STRING[4] version 10.5 protein-protein interaction database. The interaction links are for the **homosapiens** class. STRING provides uniquely comprehensive coverage and ease of access to both experimental as well as predicted

[1] http://cancer.sanger.ac.uk/cosmic/analyses

[2] https://www.genenames.org/

[3] http://dgidb.org/api/

[4] https://string-db.org/

Property	Value
Number of tumor samples	3397
Number of genes	1048
Number of drugs	3884
Number of relations between drugs and genes (actions)	10301
Number of relations between tumor samples and genes (hasGene)	58079
blocker (Label)	186
antagonist (Label)	528
agonist (Label)	525
activator (Label)	216
inhibitor (Label)	688
channel blocker (Label)	183
binder (Label)	202
gene-gene embeddings network number of edges	9872

TABLE 1
Summary of the tumor samples-gene, drug-gene and gene-gene embeddings networks.

genetic interaction information. To convert the protein interaction network into a gene interaction network for STRING, we performed the following steps: (i) Protein names were mapped to their encoding genes by parsing of EnsEMBL files. (ii) In the case of genes encoding multiple proteins, we took the edge of maximum (integrated) weight connecting any pair of proteins encoded by such genes. A similar technique for protein to gene mapping has also used in the prior studies [21].

There are eight different variants of interaction channels available in STRING, which are as follows: *co-expression, co-occurrence, database, experimental, fusion, neighborhood, textmining*, and *combined*. The *combined* protein functional association network is based on combining the probabilities from the different interaction channels. The brief description of the interaction channel is given in Supplementary Table 3. The links in the channels are all weighted and modeled as an undirected network. Due to very few links in the Fusion interaction channel, we omitted the fusion genetic interaction in our studies. The number of edges in the interaction channels is demonstrated in Table 2.

Interaction Channels	Value
co expression	208,470
cooccurence	1,166
textmining	322,883
database	23,169
neighborhood	18,929
fusion	20
experimental	170,642
combined	358,627

TABLE 2
The number of protein functional association network extracted from the STRING Database.

The edge weight between the genes means confidence scores, which are scaled between zero and one. They refer to the estimated likelihood that a given interaction is biologically meaningful, specific, and reproducible, given the supporting evidence [22].

4 Solution Approach

The main aim of this study is to classify the MOA's of drugs by combining genetic interaction and genetic embeddings network. For this, we need the graph as input. The input graph is the genes with few labeled information about the drugs. For this purpose, we combined the embedding gene graph constructed from the tumor-gene and drug-gene bipartite graph with a real protein functional association graph. It is shown in the input process in Figure 1. Once we have the input graph, then we propagate the drugs label information in these networks using the harmonic function. Harmonic function propagates the drug label in the unlabeled nodes. Those nodes which are unlabeled in the beginning are now labeled after the propagation. After the propagation is over, we get the label propagation scores of every drug for the respective genes. Using these scores, we evaluated the efficiency of the harmonic function. The whole input, process, and output are shown in Figure 1.

5 Methods

5.1 Construction of Gene-Gene Embeddings Graph

We have two bipartite graphs (i) tumor samples and gene graph (ii) drugs and gene graph, as shown in Figure 1 first input graph. These two graphs with two different relationships can also be viewed as a multi-relational graph. The multi-relational graph is a tuple $G := (V, E, L)$ where V is a set of nodes, L is a set of relationships and $E \subseteq VXVXL$ set of edges. The set of nodes and edge label in our graphs are $V = \{tumor\ samples, drugs, genes\}$ and $L = \{actions, hasGene\}$. The multi-relational graph can be modeled as tensors, which are n-modal generalizations of matrices. The features of the nodes in the multi-relational graph can be extracted using tensor factorization. We extracted the features of genes using the tensor factorization method. To do so, we employed the RESCAL framework [10]. RESCAL can embed multiple types of edges and perform collective learning through latent components of the model. In our graph, the genes are shared between the tumor samples and the drug. The shared nodes representation in RESCAL captures the similarity of the nodes in the relational domain. Thus, the genes with many similar observed relationships have similar latent representations. In matrix notations, RESCAL tensor factorization can be expressed as: $F_k = EW_kE^T$, where $F_k \in \mathbb{R}^{N_eXH_e}$. The symbol E is the Entity embedding matrix of size N_eXH_e, N_e is the number of entities or nodes, H_e is the number of latent feature for entities. Similarly, symbol W_k is the asymmetric weight matrix for relation k of size H_eXH_e. The matrix F_k captures all scores for the k relationship and the *i*-th row of $E \in \mathbb{R}^{N_eXH_e}$ captures the latent representation of e_i which is the latent feature representations of entity e_i. Once the latent representation of gene nodes are extracted, then it is used to construct gene-gene similarity graphs using different machine learning kernel method.

In this work, we used the K-NN method to construct the gene-gene graph from the extracted feature vectors. The K-NN method to construct graphs is considered as the established data structure in data mining [23] and machine learning. Moreover, in the situation with datasets without explicit graph structure, it is desirable to use the K-NN graph construction for the network analysis method [24]. Thus in the context of our work two genes (g_i, g_j), are connected if one of the genes is among the other gene's nearest neighbor and the edge weight is 1, i.e, $w_{i,j} = 1$ else the edge weight is 0, i.e, $w_{i,j} = 0$. We used 'euclidean' distance metric

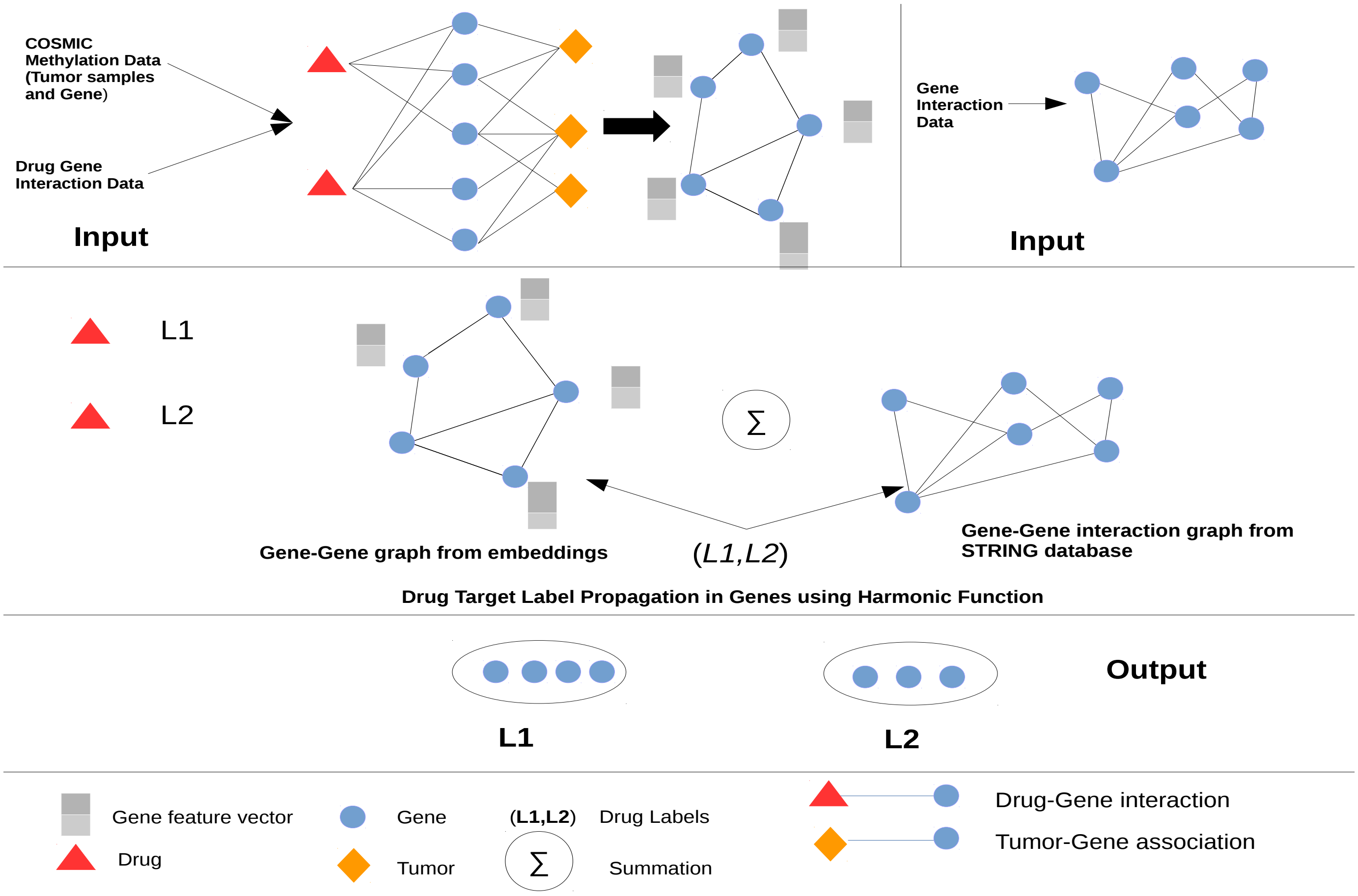
COSMIC Methylation Data (Tumor samples and Gene)
Drug Gene Interaction Data
Input
Gene Interaction Data
Input
L1
L2
Gene-Gene graph from embeddings
(L1,L2)
Gene-Gene interaction graph from STRING database
Drug Target Label Propagation in Genes using Harmonic Function
Output
L1
L2
Gene feature vector
Drug
Gene
Tumor
(L1,L2) Drug Labels
Σ Summation
Drug-Gene interaction
Tumor-Gene association

Fig. 1. The two bipartite graph is constructed using two different data source (i) Cosmic methylation data for tumor samples and genes (ii)Drug Gene interaction from DGIdb database. From these 2 bipartite graphs, the gene encodes the features shared between tumor samples and drugs. From the encoded features, Gene-Gene graph is constructed. The third data is from the STRING database for genetic interaction. The harmonic function is applied to the Gene-Gene graph for the genes which have a drug label. The output is the genes that are classified as the for drugs function.

to calculate the distance between the data points because it is considered as the best method for continuous feature vectors [25], [26]. The optimum K is chosen from the 5 fold cross validation in training sets.

5.2 Semi Supervised Learning Using Harmonic Function

Semi-Supervised Learning (SSL) is halfway between supervised and unsupervised learning. A semi-supervised learning algorithm is exposed to both unlabeled and labeled data. SSL using harmonic functions is a method of classifying data by considering the data group as a graph. Let us assume a weighted graph G with n nodes indexed as $1, ..., n$. A symmetric weight matrix, denoted as W, represents the strength of linkage. All weights are non-negative ($w_{ij} \geq 0$), and if $w_{ij} = 0$, there is no edge between nodes i and j. We assume that the first l training nodes have binary label, $y_1, y_2, ..., y_l$, where $y_i \in \{-1, 1\}$. The remaining is the unlabeled nodes given as $u = n - l$ also known as test nodes. Thus the goal here is to predict the label for the unlabeled nodes for $y_{l+1}, y_{l+2}, ..., y_n$. The underlying assumption used here is that the label of an unlabeled node is likely to be similar to the label of its neighboring nodes. Mathematically, to find a function $f(x) \in \{-1, 1\}$ on the vertices, such that $f(x_i) = y_i$. In the graph context, a harmonic function is a function that has the same values as given label on the labeled data, and satisfies the weighted average property on the unlabeled data: $f(x_i) = y_i, i = 1, ...l;$

$$f(x_j) = \frac{\sum_{k=1}^{l+u} w_{jk} f(x_k)}{\sum_{k=1}^{l+u} w_{jk}}, j = l + 1...l + u.$$

This iterative procedure will converge to a harmonic function, regardless of the initial values on the unlabeled vertices. The unnormalized graph Laplacian matrix L is defined as:

$$L = D - W \tag{1}$$

where D is the degree matrix and W is the weight of edges between the nodes. The normalized graph Laplacian is given by [27]:

$$\mathcal{L} = D^{-1} L \tag{2}$$

$\mathcal{L}$ has close connection to random walk processes on graphs [28]. The normalized Laplacian matrix can be subdivided into 4 submatrix as $\mathcal{L}$ is an $(l + u)$ x $(l + u)$ matrix with labeled ones are listed first.

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}_{ll} & \mathcal{L}_{lu} \\ \mathcal{L}_{ul} & \mathcal{L}_{uu} \end{bmatrix} \quad 276$$

The function f can be partitioned into functions of labeled and unlabeled nodes (fl, fu), and let $y_l = (y_1, ..., y_l)$. Then solving the constrained optimization problem using Lagrange multipliers with matrix algebra, the harmonic solution is $f_l = yl$ and,

$$f_u = -\mathcal{L}_{uu}^{-1}\mathcal{L}_{ul}y_l \tag{3}$$

5.3 Zoom-in View of the Label Propagation

To demonstrate the label propagation mechanism in a real genetic interaction, we took the co-expression protein functional association graph. The red node is gene *PIK3CB* which is the seed node in Figure 2. This node is labeled as the target of ″inhibitor″ drugs. We extracted the one-hop ego network around the *PIK3CB* seed gene limiting only ten genes for demonstration purposes. We set the status vector for the *PIK3CB* gene as 1 and other genes as 0 and apply the harmonic function.

From Figure 4, we can see that as the time passes the weight of the seed nodes starts to decrease whereas the other neighboring node starts to increase. After time $t > 1$ all the nodes reach to a uniform distribution of the weight. It means the neighboring nodes will adopt the same label as seed node.

5.4 Combining Multiple Graphs

As we have multiple graphs, it is natural to incorporate them as supplementary information sources. For instance, protein interactions can be represented as various types of graphs according to their co-expression, co-occurrence, fusion, or other relationships. To incorporate all the graphs, one can straightforwardly combine the normalized Laplacian matrix of the various graphs [29], [30]:

$$\mathcal{L}_{comb} = \sum_{k=1}^{m} \mathcal{L}_k$$

where m is the number of graphs to incorporate, $\mathcal{L}_{comb}$ is the combined normalized Laplacian matrix. Thus using $\mathcal{L}_{comb}$ in Equation 3 we get the score for unlabeled nodes as,

$$f_u = -\mathcal{L}_{uu_{comb}}^{-1}\mathcal{L}_{ul_{cobm}}y_{l_{comb}} \tag{4}$$

We used Equation 4 from here onwards in all our experiment.

6 Experiments

The experiment was conducted on the combined and individual protein functional association network. The brief description of the label or MOAs is shown in Supplementary Table 3. First of all, the 1048 genes were labeled as a target for drug functions using DGIdb, which is a database that annotates the genes for drug-gene interactions and potential druggability. This database allows the search of interaction for drugs-genes by gene or drug names. As this is a multi-label classification problem, we took the strategy of One Vs. Rest approach that comprises of training a single classifier for each class, with the samples of that class as positive samples and all other samples as negatives. Using this strategy, we computed the accuracy of the model. The detailed summary of all the networks used in the experiment is in Table 1.

We used ten-fold cross-validation to evaluate our approach. We randomly partitioned the nodes into training and testing sets. The ROC (receiver operating characteristic) score is calculated and then averaged over all the ten partitions. ROC score measures the overall quality of the ranking induced by the classifier, rather than the quality of a single value of threshold in that ranking [31]. ROC score of 0.5 corresponds to random guessing, and a ROC score of 1.0 implies that the algorithm succeeded in putting all of the positive examples ahead of all of the negatives. The value of parameter k for the nearest neighbor was determined by ten-fold cross-validation in a training set of the data.

Reproducibility The datasets and the codes used in this study is available in:
https://github.com/timilsinamohan/closed_form_harmonic_function

6.1 Comparison with Embeddings, Combined Genetic Interaction and Combined Genetic Interaction + Embeddings Graphs

In this experiment, we demonstrated the prediction performance of harmonic function using (i) Embeddings (EMB), (ii) Combined Genetic Interaction (CGI) (iii) Combined genetic interaction with Embedding (CGI+EMB) graph. The accuracy of label prediction is reported in terms of mean AUC-ROC scores.

The result of the experiment is demonstrated in Figure 5. We used 30% randomly picked labeled data and 70% unlabeled data. Of the seven drug functions prediction, we observed that the (CGI+EMB) method outperformed the individual graph in predicting **Blocker, Antagonist, Inhibitor, Channel Blocker**, and **Binder** label. We performed the paired t-test for ten-fold cross-validations results between (i) CGI+EMB versus EMB (ii) CGI+EMB versus CGI. The result of the test is shown in Table 3. From the test, we observed a significant difference between CGI+EMB with CGI (P = 2.329e-4) and CGI+EMB with EMB (P = 1.59e-5) for predicting Blocker label. We observed a similar prediction for the Antagonist label (P = 1.544e-2) by CGI+EMB versus CGI and (P = 2.927e-3) by CGI+EMB vs. EMB. The predictions performed by CGI+EMB graph is significantly different from EMB graphs in the prediction of all the label whereas CGI+EMB has non-significant P-values with CGI graphs in predicting all the label except Antagonist and Blocker label.

6.2 Comparison with Individual protein functional association network

We have six individual protein functional association network. For each graph, we applied the harmonic functions using 30% randomly picked labeled data. The performance of the label prediction by each graph is in Figure 6. The results showed that the text mining, database, and experimental protein functional association network have mean AUC-ROC score of more than 0.6 for predicting the (i) Binder (ii) Blocker, (iii) Channel blocker, (iv) Agonist and (v) Antagonist versus all drug label.

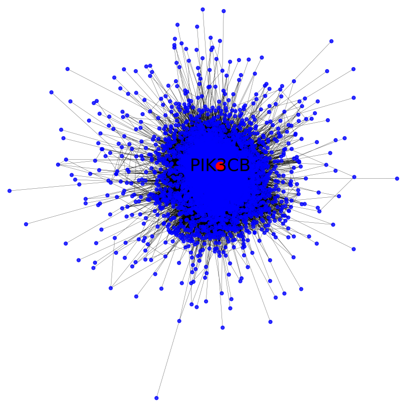

Fig. 2. A co-expression Network, with one seed node *PIK3CB* labeled as red.

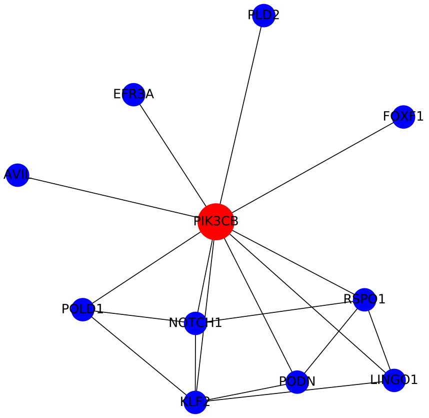

Fig. 3. Zoom-in view of the seed node *PIK3CB* with 10 neighbors

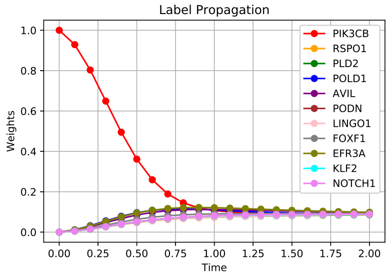

Fig. 4. Label propagation around the neighbors of gene *PIK3CB*

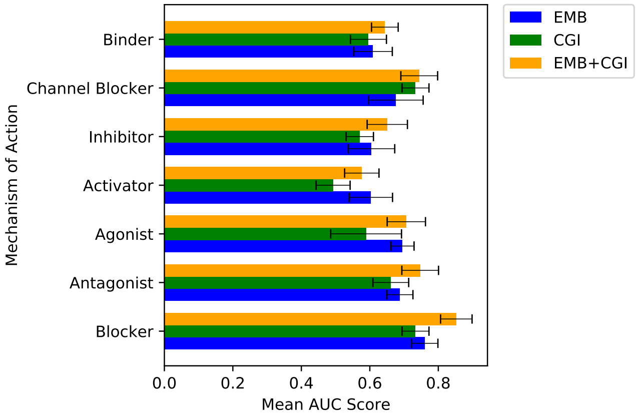

Fig. 5. The bar chart shows the Mean AUC-ROC score of 10 fold cross-validation to predict different drug functions. Each barchart shows the mean AUC-ROC score using different graphs for predicting one drug function versus all. The error bar is the standard deviation obtained from the 10-fold cross-validation for the AUC-ROC score.

	CGI	EMB
CGI+EMB (Blocker)	**2.329e-4****	**1.591e-5 *****
CGI+EMB (Antagonist)	**1.544e-2 ***	**2.927e-3 ****
CGI+EMB (Agonist)	6.149e-1	**9.686e-3 ****
CGI+EMB (Channel Blocker)	2.975e-1	8.101e-1
CGI+EMB (Inhibitor)	1.412e-1	**3.783e-3 ****
CGI+EMB (Activator)	3.408e-1	**2.229e-3 ****
CGI+EMB (Binder)	1.498e-1	**4.226e-2 ***

TABLE 3
Result of the paired t-test between CGI+EMB Vs CGI and EMB graphs predicting the label. The figure displayed in the table is the P-value at α = 0.05, *** $p < 0.05$ refers highly significant. ***: Highly Significant, **: Moderately Significant, *:Lowly Significant.

From individual protein functional association network, the Coexpression, Experimental, Textmining, and Database graphs have performed better compared to Cooccurrence and Neighborhood graphs. For Blocker Vs. All Coexpression graph leads to a marginal improvement of the AUC-ROC score in comparison to Experimental and Textmining graphs. To perform the significant test among the protein functional association network, we chose the Coexpression graph with all the protein functional association network. It is because the coexpression network is constructed using similar mRNA expression data profiles; this makes the co-expression genes as the target for a particular drug function [32].

For predicting the Antagonist label, there is no significant difference observed in the prediction between Coexpression Vs. Experimental (P = 3.228e-1), Coexpression Vs. Textmining (P = 6.136e-1), and Coexpression Vs. Database (P = 7.27e-1) graphs. Similarly, a non significant difference is observed in predicting Agonist label using Coexpression Vs Experimental (P = 1.649e-1), Coexpression Vs Textmining (P = 2.307e-1) and Coexpression Vs Database (P = 8.01e-1) graphs. It also holds for predicting the Activator label where Coexpression Vs. Experimental (P = 5.5e-1) and Coexpression Vs. Textmining (P = 6.86e-1). In the case of predicting the Inhibitor label, Coexpression Vs. Experimental (P = 4.89e-2) has a weakly significant difference but no significant difference in Textmining and Database with Coexpression graphs. For Channel Blocker label prediction, we observed the significant difference in Coexpression Vs. Experimental (P = 6.336e-3) but no significant difference between Coexpression Vs. Textmining (P = 7.54e-1) graphs. Finally, in predicting the Binder label, we observed a non-significant difference between Coexpression Vs. Experimental (P = 5.8e-1) and Coexpression Vs. Textmining (P = 1.004e-1) graphs. We have provided detail results of the paired t-test in Supplementary Table 1.

6.3 Comparison between Embeddings and protein interactions Graphs

The embedding graph is quite different from the protein functional association network because they are constructed using only relational information of tumor samples, genes, and drugs. Therefore, we performed the paired t-test between the protein functional association network to find if there is a significant difference in the prediction of the label between the embeddings graph and other protein functional

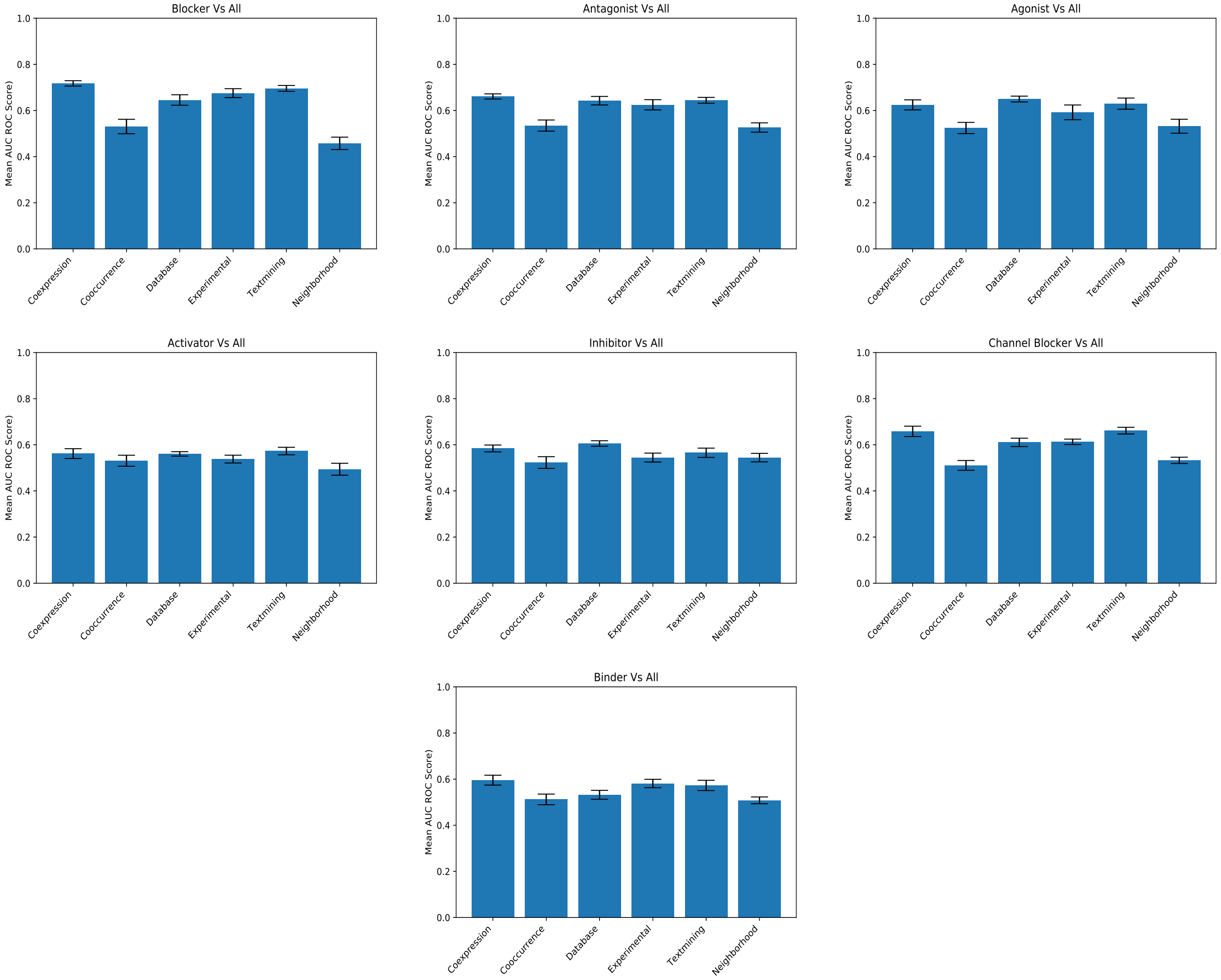

Fig. 6. The bar chart shows the Mean AUC-ROC score of 10 Fold cross validation to predict different drug functions. Each figure shows the mean AUC-ROC score using different protein interactions graphs for predicting one drug functions versus all. The error bar is the standard deviation obtained from the 10-fold cross validation for AUC-ROC score.

association network. For the prediction of the Blocker label, there was a significant difference between Cooccurrence (P= 8.28e-5), Neighborhood (P= 1.06e-5), and Database (P = 3.5e-2) with the Embedding graph.

For Channel Blocker, there is a significant difference between Cooccurrence (P= 1.315e-3) and Neighborhood (P= 4.624e-3) with the Embedding graph. For Binder label prediction a significant difference between Cooccurrence (P= 1.51e-2) and Neighborhood (P=1.82e-3) with Embedding graph. Similarly, the difference was significant with Cooccurrence (P= 2.117e-4) and Neighborhood (P= 3.332e-5) with the Embedding graph for the Antagonist label. For Agonist and Inhibitor Label prediction, the Embedding graph was borderline significant with Textmining (P= 5.09e-2) and Database (5.5e-2), respectively.

The Embedding graph showed significant difference in predicting Activator label with Coexpression (P = 7.32e-4), Experimental (P =3.45e5), Textmining (P = 3.576e-4) and Database (P = 6.596e-3) graphs. The detail results of all the paired t-test are in Supplementary Table 2.

6.4 Ablation study

We conducted an ablation study to investigate the combination of different protein functional association network to predict the drug mechanism of action. For each label(MOA), we have 57 genetic interaction combinations. To demonstrate all the 399 (57 X 7) protein interactions for seven labels will be huge to report in the manuscript. Thus we have provided the results of all the combinations in Supplementary Table 4. However, in Table 4, we have provided the results of all the genetic interaction combinations along with the top 2 AUC ROC score for each label prediction.

We observed that for all the label predictions, there is an equal or very marginal improvement in the AUC-ROC score for the prediction of drug MOA's using ”All interaction” and other various genetic interaction combinations. For predicting ”blocker,” ”antagonist,” ”channel blocker,” MOA combining all the protein functional association network (”All interaction”) underperform slightly marginally, and the difference is not very significant. Whereas, for ”inhibitor” and ”binder,” combining all the protein functional association graph (”All interaction”) equals to the performance of other genetic interaction combinations. It provides us the information that even if we combine all the graphs, we do not lose significantly in terms of the AUC-ROC score.

Interaction Combination	Drug Actions	AUC_ROC
All interaction	blocker	0.701 ± 0.011
coexpression, cooccurence, database, experimental, textmining	blocker	**0.719** ± **0.011**
coexpression , cooccurence, experimental	blocker	0.718 ± 0.012
All interaction	antagonist	0.665 ± 0.010
coexpression, cooccurence, database	antagonist	**0.675** ± **0.091**
coexpression, database, textmining	antagonist	0.674 ± 0.006
All interaction	channel blocker	0.662 ± 0.01
coexpression, cooccurence, database, textmining	channel blocker	**0.669** ± **0.006**
coexpression, database	channel blocker	0.665 ± 0.013
All interaction	agonist	0.659 ± 0.015
coexpression, database, textmining, neighborhood	agonist	0.664 ± 0.013
coexpression, database, neighborhood	agonist	0.662 ± 0.013
All interaction	inhibitor	**0.618** ± **0.011**
cooccurence, database	inhibitor	**0.618** ± **0.012**
cooccurence, database, textmining	inhibitor	0.617 ± 0.018
All interaction	binder	**0.603** ± **0.013**
coexpression, database, experimental	binder	**0.603** ± **0.023**
coexpression, cooccurence, database, experimental	binder	0.602 ± 0.022
All interaction	activator	0.592 ± 0.023
cooccurence, database, textmining	activator	**0.599** ± **0.021**
cooccurence, database, experimental, textmining	activator	0.596 ± 0.027

TABLE 4
Ablation study to find the best combination of protein functional association network to predict drugs mechanism of action. The result reported is the AUC-ROC score for ten-fold cross-validation for predicting drug MOA. The figure behind ± sign is the standard deviation. All interaction means combining all the protein functional association network.

6.5 Performance of the Harmonic Function in an Embedding and Combined protein functional association graph using Different Label percentage

To demonstrate the robustness of the harmonic function, we used different percentages of the labeled data in training sets ranging from 10% to 90%. For each percentage of the labeled data, we ran 10 Fold cross-validation. For the optimum K, we estimated it from the training set in cross-validation and applied that K to construct the embedding graph. The performance of the algorithm is shown in Figure 7.

We observe that the algorithm performed better in predicting Blockers in comparison to other drug labels. The other key observation is that even if we used the different percentages of labeled data, there is not so much of a significant difference in the accuracy. For instance, in predicting the Blocker label, if we used only 30% of the labeled data, then the algorithm gives the mean AUC-ROC score of 0.81 and using 60% labeled data the AUC-ROC score is 0.84. The difference is minimal. Also, for predicting the Antagonist label, the prediction using 30% and 80% label is almost the same. It means that the harmonic function makes use of the graph structure to exploit the information of unlabeled data for the classifications problem.

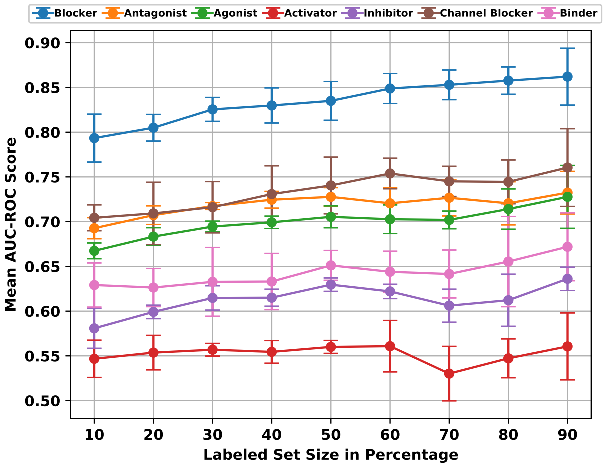

Fig. 7. Label Propagation using Harmonic Function in a EMB + CGI Graph with different label proportion.

6.6 Performance of the Harmonic Function with other Disease Gene Association Database.

In this study, we have extracted the embeddings of the genes shared between tumor samples and drugs. The tumor samples and gene associations are taken from the COSMIC database. We use this database because it is the most comprehensive source of information on somatic mutations and their frequencies in human cancers. Of course, we can apply this approach to any other disease and gene associations. We have implemented the harmonic function algorithm using 30% labeled data in popular OMIM, DisGeNet, and eDGAR database. The result of the experiment is shown in Table 5.

Label	DisGenet	eDGAR	OMIM
Blocker	0.80 ± 0.024	0.87 ± 0.014	0.89 ± 0.006
Antagonist	0.72 ± 0.056	0.80 ± 0.015	0.80 ± 0.008
Agonist	0.69 ± 0.012	0.81 ± 0.007	0.82 ± 0.011
Activator	0.60 ± 0.017	0.76 ± 0.012	0.77 ± 0.018
Inhibitor	0.61 ± 0.011	0.72 ± 0.015	0.73 ± 0.017
Channel blocker	0.68 ± 0.012	0.82 ± 0.013	0.83 ±0.019
Binder	0.64 ± 0.022	0.75 ± 0.019	0.77 ± 0.013

TABLE 5
Harmonic function label propagation experiments on general disease-gene association data sources, using 30% training labeled data. The scores are the AUC-ROC and the figure in the parenthesis is the standard deviation by 10 Trials.

In these datasets, we observed that the algorithms perform better in predicting most of the MOA than in the COSMIC cancer database. One of the reasons for that is cancer is a complicated disease that cannot be explained by individual pathways, but rather the interaction among multiple pathways [33]. Also, COSMIC data we used is a patient’s tumor sample and gene assosciation; therefore, different genetic backgrounds among different patients make it even more complicated to predict such MOA's with high accuracy.

7 Comparison with the state of art methods

We compared the result of Harmonic Function (HMN) with the state of the art graph-based label propagation algorithm namely: (i) Heat Diffusion (HD) [8], (ii) Local and Global Consistency Method (LGC) [6], (iii) OmniProp (OMNI) [5], (iv) Confidence Aware Modulated Label Propagation CAMLP [9] (v) Katz [7] and (vi) PageRank [34].

We divided 30% nodes into a training set and 70% nodes as testing sets and used precisely the same trails for all the state of the art algorithms. The results are compared by ten trials using randomly constructed 10 test sets for the networks to compute the AUC-ROC score using the "one versus all" strategy of multilabel classification.

Similarly we used non-graph based classification algorithms like Logistic, K-nearest neighbor and Support Vector Classification with linear and radial kernel. The embeddings of gene nodes are extracted from the RESCAL framework and trained these classifer in using the same setting as 30% data points into a training set and 70% data points as testing sets by using 10 trials.

From Table 6, we observed that HMN outperforms or equals to the state of the art algorithms. HMN method performs slightly better than the PageRank algorithm. Both the PageRank algorithm and HMN is based on the essence of a random walk on a graph, resting on the assumption that similar nodes are more likely to take similar labels. However, HMN has the best AUC-ROC score in Blocker and Binder labels. Note that HMN does not use any parameters and perform long-range or global diffusion on the graphs, and therefore, it is the simplest model with good AUR-ROC score in comparison to other algorithms. We also observed that the HMN outperforms the non-graph based supervised methods in predicting all the drug mechanism of actions.

8 Investigation of the Harmonic Function's Drug MOA Prediction

We performed a literature-based evaluation of the prediction of drug MOA by harmonic function. Our task is to evaluate the quality of harmonic function's predictions about classifying genes based on drugs MOA. For this purpose, we trained harmonic function with 30% labeled data and ranked the top prediction based on the predicted harmonic scores 4. We explored the ten highest ranked predictions in the list. We searched the biomedical literature to see if we can find supporting evidence for these predictions

Table 7 shows harmonic function's predictions and literature evidence supporting these predictions. We note that the cited literature investigates interactions between the drug MOA and the target genes. For example, harmonic function classifies the gene *SCN11A* as the "antagonist" target for the drugs (Table 7, 5th highest ranked prediction for antagonist MOA). In fact, the work by Emery et al. [36] have found sodium channel "antagonists" involvement of *SCN11A* genes in treating most pain syndromes. Similarly, the gene *KCNH3* (Table 7, 1st highest ranked prediction for blocker MoA) are used for silent voltage gated "blockers" of K+ channels which may have potential benefit in diseases involving immune cell activation and proliferative diseases, such as cancer, fibrosis, atherosclerosis, and restenosis [35]. The analysis here shows the possibilities of harmonic functions predictions for gene classification.

9 Discussion

For the first research question, we found that the harmonic functions with combining CGI+EMB network predict Blocker with high accuracy in comparison to other drug functions in our datasets. The combined CGI+EMB network also performed better for Antagonist and Channel Blocker labels. Moreover, the Antagonist drugs are also called blockers, for instance, alpha-blockers, beta-blockers, and calcium channel blockers [42]. The studies also showed that Blocker drugs appear to have a beneficial clinical effect in cancer pathology. In our data, we incorporated the tumor's information; this might have helped for the higher prediction of the Blocker label. In clinical settings, blocker interactions are used to reduce the rates of progression of different solid tumors. The Blockers drug could potentially result in a 57% reduction in the risk of metastasis and a 71% reduction in the 10-year mortality rate in a breast cancer [43].

Similarly, another interesting observation from our study is the prediction of the Antagonist label. The prediction accuracy was third highest for Antagonist label prediction after Blocker and Channel Blocker using the harmonic function applied on CGI+EMB network. In drug design, Antagonistic drug combinations are mostly used to avoid the development of drug resistance. Furthermore, combination therapies are being used to combat drug resistance in cancer patients under chemotherapeutic agents [44]. This enhances the discovery of novel efficacious combinations of drugs and targets. Not only the drug resistance but also the antagonist drugs bindings can inhibit the specific cases like gastrointestinal cancer [45].

For the second research question, the results showed that using Coexpression graphs in the majority of the label prediction performed better than the other protein functional association network. The dynamic change of protein-protein interaction, such as co-expression networks, is the critical determinant of the disease state. Due to this, co-expression networks are richly targeted for drug design. Not only the Coexpression graphs but also the Textmining graphs showed similar performance in label prediction. The text mining methods are extensively used to extract genetic interaction form scientific literature to enrich drug-therapy networks [46] and disease studies.

For the third research question, we found that the EMB graph leads to a mean AUC-ROC score above 0.6 for predicting Blocker, Antagonist, Agonist and Channel Blocker label, as shown in Figure 5. The EMB graphs provided better prediction for Blocker, Antagonist, Agonist, Inhibitor, Channel Blocker, and Binder label than Cooccurrence and Neighborhood protein functional association network. By combining GGI+EMB graphs, the prediction performance has improved for the label Blocker, Antagonist, Channel Blocker, Inhibitor, and Binder. It means that the EMB graphs

	Blocker	**Antagonist**	**Agonist**	**Channel Blocker**	**Inhibitor**	**Activator**	**Binder**
LR	0.74 ± 0.016	0.64 ± 0.020	0.61 ± 0.012	0.64 ± 0.023	0.59 ± 0.019	0.55 ± 0.024	0.60 ± 0.019
KNN	0.65 ± 0.022	0.61 ± 0.018	0.57 ± 0.028	0.58 ± 0.024	0.54 ± 0.025	0.50 ± 0.020	0.55 ± 0.022
SVC (RBF)	0.75 ± 0.017	0.50 ± 0.134	0.50 ± 0.011	0.58 ± 0.111	0.53 ± 0.062	0.53 ± 0.049	0.62 ± 0.021
SVC (Linear)	0.74 ± 0.012	0.64 ± 0.017	0.58 ± 0.065	0.55 ± 0.128	0.54 ± 0.059	0.53 ± 0.043	0.62 ± 0.029
HD	0.76 ± 0.012	0.71 ± 0.013	0.68 ± 0.009	0.67 ± 0.020	**0.62 ± 0.015**	0.51 ± 0.010	0.56 ± 0.023
LGC	0.75 ± 0.012	0.71 ± 0.013	0.68 ± 0.009	0.67 ± 0.018	**0.62 ± 0.023**	0.51 ± 0.017	0.56 ± 0.022
OMNI	0.79 ± 0.015	0.70 ± 0.019	0.57 ± 0.015	**0.72 ± 0.017**	0.53 ± 0.009	0.54 ± 0.011	0.60 ± 0.023
CAMLP	0.65 ± 0.0211	0.52 ± 0.006	0.54 ± 0.089	0.65 ± 0.021	0.54 ± 0.012	0.51 ± 0.009	0.53 ± 0.017
Katz	0.73 ± 0.016	0.71 ± 0.007	0.68 ± 0.009	0.65 ± 0.018	0.61 ± 0.013	0.52 ± 0.021	0.54 ± 0.017
PageRank	0.81 ± 0.014	**0.72 ± 0.008**	**0.69 ± 0.012**	0.70 ± 0.03	0.60 ± 0.017	**0.55 ± 0.017**	0.61 ± 0.027
HMN	**0.82 ± 0.013**	**0.72 ± 0.009**	**0.69 ± 0.011**	**0.72 ± 0.024**	0.60 ± 0.015	**0.55 ± 0.015**	**0.63 ± 0.028**

TABLE 6

Comparison of Harmonic Function with state of the art Graph-Based Semi-Supervised machine learning algorithms. The result reported is the AUC-ROC score for ten trails for predicting drug MOA. The figure behind ± sign is the standard deviation.

k	Mechanism of Action	Gene names	Evidence
1	Blocker	*KCNH3*	Wickenden et al. [35]
5	Antagonist	*SCN11A*	Emery et al. [36]
1	Agonist	*GABRQ*	Li et al. [37]
1	Activator	*SCN9A*	Drenth et al. [38]
2	Inhibitor	*NMBR*	Zhao et al. [39]
3	Channel Blocker	*CACNA1F*	MCRory et al. [40]
2	Binder	*GABRA4*	Reddyet al. [41]

TABLE 7

Genes classified based on the drug's MOA's using harmonic function. The genes are assigned the highest scores by the harmonic function. For each prediction, we include its rank k in the ranked list of all predictions and literature evidence.

act as complementary information that enhanced the prediction performance.

We have used the RESCAL tensor factorization model for learning the node embeddings. This model's main caveat is that we do not know the number of latent components in advance. Thus, we need to do a grid search for the best parameter, which slows computation time. Another limitation in RESCAL based tensor embedding model is that the number of parameters grows linearly with the number of relationships in the graphs, making it difficult to scale in highly-relational graphs [47]. Thus, we consider assessing the quality of the state of the arts graph-based embedding model for drug label prediction for our future work.

10 Conclusion

Our study used two different graphs (i) constructed from the feature extraction of a tumor, genes, and drugs from the multi-relational graph using the k-nearest neighbor approach and (ii) protein functional association graph from the protein-protein interaction STRING database. We combined these two graphs and applied harmonic function to classify seven different drug labels, namely Blocker, Antagonist, Agonist, Activator, Inhibitor, Channel Blocker, and Binder. The harmonic function predicted the highest AUC-ROC score for the Blocker, Channel Blocker, Agonist and Antagonist label using combined graph embedding and protein interactions. The graph-combining method showed better results on drug label prediction, performing significantly better than any single protein functional association graph such as coexpression, co-occurrence, database, experimental, neighborhood, and text mining, particularly for predicting Blocker, Channel Blocker, Antagonist, and Agonist label.

The graph-combining method provides a straightforward way of combining multiple graphs. However, work remains for the future. The harmonic functions assume homophily networks, which means nodes with similar characteristics tend to connect. The harmonic function propagates signals on the graph using the homophily principles, which sometimes leads to misclassification. For instance, different drug label targets the same genes. In the context of cancer, it is essential to use different drugs for combinational therapy because it targets critical pathways in a simply synergistic manner. It possibly reduces the performance of a harmonic function that assumes label smoothing. We have not looked at this perspective, which is an important feature to address.

Acknowledgments

We would like to thank the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P for the financial support to conduct this research.

References

[1] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu *et al.*, "Drugbank 4.0: shedding new light on drug metabolism," *Nucleic acids research*, vol. 42, no. D1, pp. D1091–D1097, 2013.

[2] F. Cheng, J. Zhao, M. Fooksa, and Z. Zhao, "A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes," *Journal of the American Medical Informatics Association*, vol. 23, no. 4, pp. 681–691, 2016.

[3] M. Timilsina, M. Tandan, M. d'Aquin, and H. Yang, "Discovering links between side effects and drugs using a diffusion based method," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[4] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.

[5] Y. Yamaguchi, C. Faloutsos, and H. Kitagawa, "Omni-prop: Seamless node classification on arbitrary label correlation," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[6] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in neural information processing systems*, 2004, pp. 321–328.

[7] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[8] H. Yang, I. King, and M. R. Lyu, "Diffusionrank: a possible penicillin for web spamming," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 431–438.

[9] Y. Yamaguchi, C. Faloutsos, and H. Kitagawa, "Camlp: Confidence-aware modulated label propagation," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 513–521.

[10] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data." in *ICML*, vol. 11, 2011, pp. 809–816.

[11] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[12] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *arXiv preprint arXiv:1802.00543*, 2018.

[13] P. Imming, C. Sinning, and A. Meyer, "Drugs, their targets and the nature and number of drug targets," *Nature reviews Drug discovery*, vol. 5, no. 10, p. 821, 2006.

[14] Editorial, "Mechanism matters," *Nature Medicine*, vol. 16, p. 347, 2010.

[15] M. Schenone, V. Dančík, B. K. Wagner, and P. A. Clemons, "Target identification and mechanism of action in chemical biology and drug discovery," *Nature chemical biology*, vol. 9, no. 4, p. 232, 2013.

[16] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature reviews Drug discovery*, vol. 1, no. 9, p. 727, 2002.

[17] X. Yang, F. Lay, H. Han, and P. A. Jones, "Targeting dna methylation for epigenetic therapy," *Trends in pharmacological sciences*, vol. 31, no. 11, pp. 536–546, 2010.

[18] S. Heerboth, K. Lapinska, N. Snyder, M. Leary, S. Rollinson, and S. Sarkar, "Use of epigenetic drugs in disease: an overview," *Genetics & epigenetics*, vol. 6, pp. GEG–S12 270, 2014.

[19] K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, "Dgidb 3.0: a redesign and expansion of the drug–gene interaction database," *Nucleic acids research*, vol. 46, no. D1, pp. D1068–D1073, 2017.

[20] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, I. Das, M. B. Callaway, J. M. Eldred *et al.*, "Dgidb: mining the druggable genome," *Nature methods*, vol. 10, no. 12, p. 1209, 2013.

[21] M. Timilsina, H. Yang, R. Sahay, and D. Rebholz-Schuhmann, "Predicting links between tumor samples and genes using 2-layered graph based diffusion approach," *BMC bioinformatics*, vol. 20, no. 1, p. 462, 2019.

[22] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork *et al.*, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research*, p. gkw937, 2016.

[23] M. Brito, E. Chavez, A. Quiroz, and J. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Statistics & Probability Letters*, vol. 35, no. 1, pp. 33–42, 1997.

[24] W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 577–586.

[25] K. Jonsson, J. Kittler, Y. Li, and J. Matas, "Support vector machines for face authentication," *Image and Vision Computing*, vol. 20, no. 5-6, pp. 369–375, 2002.

[26] M. Zhao and J. Chen, "A review of methods for detecting point anomalies on numerical dataset," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1. IEEE, 2020, pp. 559–565.

[27] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[28] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering with multi-layer graphs: A spectral perspective," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5820–5831, 2012.

[29] A. Argyriou, M. Herbster, and M. Pontil, "Combining graph laplacians for semi–supervised learning," in *Advances in Neural Information Processing Systems*, 2006, pp. 67–74.

[30] K. Tsuda, H. Shin, and B. Schölkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. suppl_2, pp. ii59–ii65, 2005.

[31] H. Shin, K. Tsuda, B. Schölkopf, A. Zien *et al.*, "Prediction of protein function from networks," in *Semi-supervised learning*. MIT press, 2006, pp. 361–376.

[32] I. Molineris, U. Ala, P. Provero, and F. Di Cunto, "Drug repositioning for orphan genetic diseases through conserved anticoexpressed gene clusters (cagcs)," *BMC bioinformatics*, vol. 14, no. 1, p. 288, 2013.

[33] D. C. Altieri, "Survivin, cancer networks and pathway-directed drug discovery," *Nature Reviews Cancer*, vol. 8, no. 1, pp. 61–70, 2008.

[34] B. Sergey, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[35] A. D. Wickenden, "K+ channels as therapeutic drug targets," *Pharmacology & therapeutics*, vol. 94, no. 1-2, pp. 157–182, 2002.

[36] E. C. Emery, A. P. Luiz, and J. N. Wood, "Nav1. 7 and other voltage-gated sodium channels as drug targets for pain relief," *Expert opinion on therapeutic targets*, vol. 20, no. 8, pp. 975–983, 2016.

[37] Y.-H. Li, Y. Liu, Y.-D. Li, Y.-H. Liu, F. Li, Q. Ju, P.-L. Xie, and G.-C. Li, "Gaba stimulates human hepatocellular carcinoma growth through overexpressed gabaa receptor theta subunit," *World Journal of Gastroenterology: WJG*, vol. 18, no. 21, p. 2704, 2012.

[38] J. P. Drenth, S. G. Waxman *et al.*, "Mutations in sodium-channel gene scn9a cause a spectrum of human genetic pain disorders," *The Journal of clinical investigation*, vol. 117, no. 12, pp. 3603–3609, 2007.

[39] Z.-Q. Zhao, L. Wan, X.-Y. Liu, F.-Q. Huo, H. Li, D. M. Barry, S. Krieger, S. Kim, Z.-C. Liu, J. Xu *et al.*, "Cross-inhibition of nmbr and grpr signaling maintains normal histaminergic itch transmission," *Journal of Neuroscience*, vol. 34, no. 37, pp. 12 402–12 414, 2014.

[40] J. E. McRory, J. Hamid, C. J. Doering, E. Garcia, R. Parker, K. Hamming, L. Chen, M. Hildebrand, A. M. Beedle, L. Feldcamp *et al.*, "The cacna1f gene encodes an l-type calcium channel with unique biophysical properties and tissue distribution," *Journal of Neuroscience*, vol. 24, no. 7, pp. 1707–1718, 2004.

[41] T. E. Reddy, B. E. Shakhnovich, D. S. Roberts, S. J. Russek, and C. DeLisi, "Positional clustering improves computational binding site detection and identifies novel cis-regulatory sites in mammalian gaba a receptor subunit genes," *Nucleic acids research*, vol. 35, no. 3, pp. e20–e20, 2007.

[42] B. G. Katzung and A. J. Trevor, *Basic & clinical pharmacology*. McGraw-Hill New York, NY, 2015.

[43] D. G. Powe, M. J. Voss, K. S. Zänker, H. O. Habashy, A. R. Green, I. O. Ellis, and F. Entschladen, "Beta-blocker drug therapy reduces secondary cancer formation in breast cancer and improves cancer specific survival," *Oncotarget*, vol. 1, no. 7, p. 628, 2010.

[44] M. S. Glickman and C. L. Sawyers, "Converting cancer therapies into cures: lessons from infectious diseases," *Cell*, vol. 148, no. 6, pp. 1089–1098, 2012.

[45] M. Muñoz and R. Coveñas, "Neurokinin-1 receptor antagonists as antitumor drugs in gastrointestinal cancer: A new approach," *Saudi journal of gastroenterology: official journal of the Saudi Gastroenterology Association*, vol. 22, no. 4, p. 260, 2016.

[46] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review," *Pharmacology & therapeutics*, vol. 138, no. 3, pp. 333–408, 2013.

[47] P. Minervini, C. d'Amato, and N. Fanizzi, "Efficient energy-based embedding models for link prediction in knowledge graphs," *Journal of Intelligent Information Systems*, vol. 47, no. 1, pp. 91–109, 2016.

Mohan Timilsina is a Postdoctoral researcher in a Data Science Institute at National University of Ireland Galway. He received his Ph.D in Computer Sciences from Data Science Institute at National University of Ireland Galway in 2020. His research interest includes applied machine learning, bioinformatics, graph mining and information retrieval from a network data.

Declan Patrick Mc Kernan received his Ph.D degree in Biochemistry from University College Cork. He is currently a lecturer above the bar with the Department of Pharmacology and Therapeutics, at the National University of Ireland Galway. His research interest include (i) Epigenetics (ii) Neuroinflammation and (iii) Innate immunity.

Haixuan Yang received a Ph.D.degree in Mathematics from Lanzhou University in 1996, and a Ph.D. degree in Computer science and Engineering from The Chinese University of Hong Kong, in 2007. He is currently a Lecturer with the School of Mathematics, Statistics and Applied Mathematics at National University of Ireland Galway. His research interests include machine learning, bioinformatics and statistical modelling, especially for network data.

Mathieu d'Aquin is a a Professor of Informatics specialised in data analytics and semantic technologies at the Data Science Institute, Insight Centre for Data Analytics of the National University of Ireland Galway.He has worked on applying the Semantic Web/Linked Data technologies coming out of his research, in various domains including bio-medicine, education especially through learning analytics.