PLOS ONE

RESEARCH ARTICLE

HetIG-PreDiG: A Heterogeneous Integrated Graph Model for Predicting Human Disease Genes based on gene expression

Kathleen M. Jagodnik[1,2,3], **Yael Shvili**[4], **Alon Bartal**[1]*

1 The School of Business Administration, Bar-Ilan University, Ramat Gan, Israel, **2** Department of Psychiatry, Harvard Medical School, Boston, MA, United States of America, **3** Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States of America, **4** Department of Surgery A, Meir Medical Center, Kfar Sava, Israel

* alon.bartal@biu.ac.il

Citation: Jagodnik KM, Shvili Y, Bartal A (2023) HetIG-PreDiG: A Heterogeneous Integrated Graph Model for Predicting Human Disease Genes based on gene expression. PLoS ONE 18(2): e0280839. https://doi.org/10.1371/journal.pone.0280839

Editor: Attila Gursoy, Koc Universitesi, TURKEY

Received: June 12, 2022

Accepted: January 10, 2023

Published: February 15, 2023

Data Availability Statement: Source code and preprocessed datasets are available at: https://github.com/bartala/disease_gene.

Funding: K.M.J. was supported by a Mortimer B. Zuckerman STEM Leadership Program post-doctoral fellowship in the School of Business Administration at Bar-Ilan University and in the Departments of Psychiatry at Harvard Medical School and Massachusetts General Hospital. We thank Bar-Ilan University's Data Science Institute (DSI) for partially supporting this research.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Graph analytical approaches permit identifying novel genes involved in complex diseases, but are limited by (i) inferring structural network similarity of connected gene nodes, ignoring potentially relevant unconnected nodes; (ii) using homogeneous graphs, missing gene-disease associations' complexity; (iii) relying on disease/gene-phenotype associations' similarities, involving highly incomplete data; (iv) using binary classification, with gene-disease edges as positive training samples, and non-associated gene and disease nodes as negative samples that may include currently unknown disease genes; or (v) reporting predicted novel associations without systematically evaluating their accuracy. Addressing these limitations, we develop the Heterogeneous Integrated Graph for Predicting Disease Genes (HetIG-PreDiG) model that includes gene-gene, gene-disease, and gene-tissue associations. We predict novel disease genes using low-dimensional representation of nodes accounting for network structure, and extending beyond network structure using the developed Gene-Disease Prioritization Score (GDPS) reflecting the degree of gene-disease association via gene co-expression data. For negative training samples, we select non-associated gene and disease nodes with lower GDPS that are less likely to be affiliated. We evaluate the developed model's success in predicting novel disease genes by analyzing the prediction probabilities of gene-disease associations. HetIG-PreDiG successfully predicts (Micro-F1 = 0.95) gene-disease associations, outperforming baseline models, and is validated using published literature, thus advancing our understanding of complex genetic diseases.

1 Introduction

Understanding the complex biological phenomena involved in human diseases is essential for developing new preventive and therapeutic strategies [1]. Since authoritative sets of genetic associations for many diseases are unknown [1], and the experimentation necessary to validate

these associations is costly and time consuming, researchers have developed computational methods, including machine learning (ML) models to discover gene-disease associations [2]. Analyzing biological data using graphs can identify complex interactions among entities (e.g., genes and diseases), and it facilitates the detection of variations (e.g., genetic mutations) via structural changes in the graph [3–5].

Some disease gene prediction models [6] assume that genes associated with biologically similar diseases have similar graph structures. Those models miss potentially relevant nodes beyond the local neighborhood of a node. Network diffusion models extend beyond the local neighborhood of known disease genes by walking over the edges of a biological graph [7–11]. For example, the MEXCOwalk algorithm [11] performs an edge-weighted random walk on a graph to identify cancer gene modules, and the HotNet2 algorithm [8, 9] employs a directed network diffusion model to assess the significance of mutations in genes, and the local topology of interactions among encoded proteins, to identify mutated subnetworks in a genome-scale interaction network. However, network diffusion models that apply walks over the graph miss potentially relevant nodes that are unconnected (e.g., isolated components or nodes) due to, e.g., missing or unknown data. For example, Del Sol et al. [12] reported that complete miRNA networks accurately represent healthy tissues, whereas cancer tissues are characterized by disjointed, disconnected sub-networks. Additionally, the lack of data about gene–gene associations is often the source of the "missing heritability" problem in which known interactions can explain only a small portion of a disease [13]. Graph-based models that assume that genes with high phenotypic similarity associate with the same disease [14] also rely on highly incomplete data that can lead to poor model performance. In addition, many models for predicting gene-disease associations construct a homogeneous graph based on a single type of data such as protein-protein interactions (PPI) [15, 16]. However, using a single type of data ignores the complexity inherent in gene-disease associations [17, 18]. For example, combining PPI with tissue-specific data is important for predicting disease genes [5].

In recent years, researchers have represented nodes as numeric vectors (embeddings) in a low-dimensional space while preserving node and graph topological similarity using neural networks [19]. These embedding vectors allow ML methods to predict disease genes in graphs [20], among other tasks. Automatic feature learning from graphs [19, 21–24] has been widely studied. Graph embeddings were successfully used in capturing the biological structures of proteins [25]; reducing data noise in graphs [26] by using tasks such as node classification, link prediction, and clustering [21]; and detecting drug-drug side-effects [27].

Node embedding models for identifying disease genes are limited by relying on gene-phenotype associations, which are highly incomplete in humans and other organisms [28]. In addition, data about gene-disease associations is often limited since these complex relationships are rare, and are usually not observed in small clinical trials, preventing ML models from learning these associations [18]. Moreover, ML models for predicting disease genes are typically approached using a binary classification of gene and disease association, by selecting a sample of (i) gene-disease edges as positive training samples, and (ii) non-associated gene and disease nodes (representing non-existing edges) as negative examples that might contain unknown disease genes. Training an ML model on those negative samples may result in poor model performance.

Another limitation of most gene-disease prediction models involves assuming that accurate performance on the test set leads to accurate predictions of novel gene-disease associations. However, predicting novel disease genes requires considering *all* potential associations between candidate genes and a disease, beyond selected gene-disease associations in the test set. The accuracy of past models when considering all candidate genes is typically supported by manually examined literature without systematic evaluation. Consequently, given a specific

disease, we do not know whether the predictions of novel genes by past models are accurate enough to be validated in wet-lab experiments, even though those models perform well on a test set.

Lack of studies that consider in a single model knowledge from both local network neighbors and non-neighbors; incorporate rich knowledge from several biological domains; and address the problem of sampling negative edges for link prediction raises the need for developing new models that will enable better prediction of disease genes. We address those limitations by developing a model to improve the prediction of disease genes: we propose a Heterogeneous Integrated Graph Model for Predicting Disease Genes (HetIG-PreDiG) with gene prioritization based on gene expression. HetIG-PreDiG (pronounced "HET-ih-jee PRED-ih-jee") detects human gene-disease associations by integrating data about gene expression in different tissues, and gene-gene, gene-disease, and gene-tissue associations into a heterogeneous graph. Using the node2vec algorithm [21], HetIG-PreDiG accounts for graph structure by learning low-dimensional representation embeddings of nodes. To extend beyond graph structure, a Gene-Disease Prioritization Score (GDPS) is developed. This GDPS reflects the association degree of a gene with a disease based on co-expression similarity across multiple tissues. Node embeddings and the GDPS are input as features to a classifier that predicts gene-disease edges. To train the classifier, we randomly select gene-disease edges as positive samples, and non-associating gene and disease nodes with lower GDPS as negative training samples, thus lowering the risk of including biologically existent yet to-date unreported disease genes. The results show that a model that considers both network structure and GDPS outperforms other baseline models. Finally, we provide a method to systematically evaluate the developed model's success in predicting novel disease genes by classifying a disease based on its candidate genes' prediction probabilities into three success level groups.

We make five novel contributions to improve the identification of human gene-disease associations: 1) developing our Gene-Disease Prioritization Score (GDPS) based on data of gene co-expression similarity. Whereas most models for predicting disease genes use gene expression to identify genes having expression most strongly associated with a disease, our model uses it together with node embeddings for learning the degree of association between any gene and a disease; 2) considering network structure using graph representation learning, and extending beyond network structure by accounting for biological similarity between unconnected gene and disease nodes using GDPS; 3) offering a solution to the problem of randomly selecting as negative training samples non-associated gene and disease nodes (reflecting non-existing edges) that might actually represent yet-unknown biological associations, by our method of favoring non-associating gene and disease pairs having lower GDPS values; 4) capturing rich biological knowledge in a single heterogeneous graph-based model, the utility of which for predicting gene-disease associations is validated by producing better results compared with baseline models, and via literature analysis; and 5) providing a new method to systematically evaluate the developed model's success level in predicting novel genes for a given disease, as the accuracy of past models that consider all candidate genes is typically supported via literature analysis, but these results are often not systematically evaluated as in the current work.

1.1 Organization

Section 2 provides a detailed overview of existing methods for predicting disease genes, and the shortcomings of those methods. Section 3 describes the HetIG-PreDiG model for predicting human disease genes, and the datasets used for learning and predicting gene-disease associations. Section 4 details the analyses performed with the developed model. Section 5

discusses the results of comparing the proposed model with baseline models, empirically evaluating the proposed model, and demonstrating its usefulness via supporting literature analysis. Section 6 discusses the strengths and limitations of this work, and interprets aspects of our results. Finally, Section 7 summarizes the contributions of this study and describes potential additional applications of our model.

2 Related work

2.1 Biomedical data

Biomedical data is often high-dimensional, incomplete, and biased due to e.g., physical measurement limitations and technological constraints [14, 26, 29]. To better understand complex biomedical phenomena such as diseases, an effective model must incorporate diverse biomedical datasets from different domains [27]. ARCHS4 [30] is a web resource that provides co-expression similarity matrices of human and mouse genes, based on RNA-seq data processed from Gene Expression Omnibus (GEO) [31]. This data can be used to detect biological functions such as gene-disease associations [30]. For example, Lachmann et al. [30] found that genes with highly correlated expression tend to share biological functions. Moreover, the authors were able to predict gene function using the extensive expression data available from ARCHS4. Other biological data sources with relevance for understanding human diseases and their treatment include the DisGeNET database [32], which provides data on Mendelian, complex, and environmental human diseases. Additionally, the Human Protein Atlas [33] serves as a map of the human proteome, providing tissue-specific gene expression data that can be used to elucidate the mechanisms of disease [34].

Building models using incomplete data can cause them to perform poorly given new data. Thus, predicting gene-disease associations for a genetic disease requires analyzing genes that are associated with the disease and genes that are likely to associate with the disease, as well as their interactions in diverse biological functions [35].

2.2 Predicting gene-disease associations

The genome-wide association study (GWAS) [36] is a widely used approach that analyzes single nucleotide polymorphisms (genetic variations) among humans for predicting new disease genes. However, predicting gene-disease associations via laboratory experiments and statistical analyses is time consuming, and often results in a large number of candidate genes with multiple false positives [37]. Moreover, GWAS mainly focuses on gene-phenotype associations, excluding the functions of biological molecules that act via complex pathways [37]. To address this gap, researchers have developed computational approaches such as networks for predicting gene-disease associations. Analyzing biological data using graphs can identify complex interactions among entities (e.g., genes and diseases), and it facilitates the detection of variations (e.g., genetic mutations) via structural changes in the graph [3–5].

Different types of graphs have been exploited for predicting disease genes [17], including homogeneous [3, 16, 38], heterogeneous [39], and multiplex graphs [40]. A homogeneous graph includes nodes and edges each of a single type, a heterogeneous graph has different types of nodes and edges, and a multiplex graph is a collection of graphs with the same set of nodes and different types of edges.

Two network-based approaches are commonly used for predicting disease genes. *Node classification* learns features of known disease genes to predict the disease labels of genes of novel disease associations. *Link prediction* learns known gene-disease associations to predict novel gene-disease links. These two network-based approaches for predicting disease genes can be implemented via three categories: 1) network diffusion, 2) supervised ML methods in which

features for diseases and genes are first extracted and then input to ML models such as Support Vector Machines (SVMs) for predicting gene-disease associations, and 3) graph representation learning.

The next subsections describe each of these methods.

2.3 Network diffusion methods

Most network-based methods assume that genes associated with biologically similar diseases have similar network structures [6, 41]. Some network methods for predicting disease genes consider only the local neighborhood of a node [3, 10, 38, 42], thus missing biological information at greater distances on the network. This limitation is partially resolved by network diffusion models that start from known disease genes and diffuse to other nodes via walks over the edges of the biological network. For example, the Random Walk with Restart (RWR) algorithm [43] performs a random walk on a graph with a restart probability r to return to any seed node at each iteration. It explores the neighborhood of seed nodes to study their functions, under the assumption that nodes related to similar functions are closer in the network. Adopting RWR, PRINCE [44] expands RWR to a weighted PPI network, and VAVIEN [7] prioritizes candidate disease genes based on the topological similarity of proteins that is calculated using RWR to perform random walks on a PPI network. RWR was widely used in PPI networks to detect novel disease genes. For example, ORIENT [45] uses RWR to detect novel disease genes in a weighted PPI network such that genes closer to known disease genes receive a higher prioritization score. The DP-LCC model [46] also detects novel disease genes using RWR on a PPI network and a phenotype similarity network.

Heterogeneous networks of gene-disease associations, disease-disease similarities, and protein–protein interactions have also been employed for predicting disease genes. For example [39], detects disease genes in heterogeneous networks using diffusion and node classification. Other examples include the RWRH model [47] that extends RWR on a heterogeneous phenotype-gene network. RWPCN [48] predicts disease genes on a heterogeneous network of phenotypes, genes, and proteins. CIPHER [49] predicts unknown disease genes in a heterogeneous network using phenotype similarity and gene proximity. The CATAPULT model [50] predicts gene-phenotype associations by vectors generated using walks on the heterogeneous network of gene-gene and gene-phenotype associations. BiRW [51] performs RWR on a heterogeneous network of phenotypes and genes. RWRMH [40] conducts RWR on a multiplex heterogeneous network of PPIs and disease associations based on phenotype similarities. Zeng et al. [52] proposed a latent factor method with heterogeneous similarity regularization to predict unknown gene-disease associations.

While network diffusion models utilize network structure to capture biological information beyond local neighborhood, they miss global information beyond network structure, involving unconnected nodes. Networks represent existing datasets, reflecting only known (often incomplete, noisy, and biased) data collected so far [4]. Hence, the information of, e.g., two unconnected proteins in a PPI network that might share the same biological pathway [15] is ignored in diffusion models. On the other hand, gene expression data can be used to calculate similarities among nodes representing biological entities [30] and is not limited by node connectivity. However, it ignores the structure of the network.

Some studies utilized gene expression data to detect disease genes, such as Hu and Agarwal [53], who begin by identifying the top genes having expression most strongly associated with each disease. Then, they perform enrichment analyses to find significant overlaps between these top genes and diseases. Another example is the DiseaseConnect web server [54] that

utilizes gene expression profiles, gene-disease associations, and GWAS data to detect novel gene-disease associations.

Recently, biological network-based models have represented nodes using feature vectors of structural network properties, such as average distance to disease genes, and structural similarity with disease genes [17, 55]. These vectors are used for training supervised ML models such as Logistic Regression (LR) and SVM to classify genes as associated with a disease or not [50, 56].

2.4 Supervised machine learning methods

Current ML methods that are applied to biological graphs typically represent genes and proteins using feature vectors of structural network properties (e.g., degree) [55]. ML models that use PPI networks to predict unknown disease genes include, e.g., the gene ranking model described in [57] that prioritizes candidate genes using network analysis of their differential expression. Relying on network structure, it assumes that candidate disease genes are neighbors of highly differentially expressed genes. More recently, BRIDGE [58] was developed to prioritize disease genes by applying Lasso Regression to a variety of biological resources, including PPI, protein sequence, gene expression, pathway, and gene ontology data. Similarly, the IMRF [59] algorithm also utilizes diverse biological data such as PPI networks to rank disease genes by improving the Markov Random Field method. Focusing on gene-disease association data, the Know-GENE algorithm [60] prioritizes candidate genes associated with a disease by calculating gene-gene similarity using gene co-occurrence. The authors recommend [60] considering gene expression data in future work to detect genes without known disease associations, as done in the current study. Representing diverse biological data as heterogeneous graphs, Metagraph+ [61] predicts disease genes by analyzing a heterogeneous graph of PPI and gene keywords. Using gene ontology similarities, the dgMDL algorithm [62] predicts gene-disease associations in a heterogeneous PPI and gene graph. The Disjunctive Graph Integration model [63] predicts novel disease genes by applying SVM to features of a heterogeneous graph of gene co-expression, pathways, functional links, phenotype similarity, and PPI.

Other data sources than PPI networks, such as disease-phenotype associations, gene ontology annotations, and tissue-specific networks, have been utilized to predict disease genes [10, 17, 42]. The use of tissue-specific gene expression data is critical, because diseases are typically associated with a specific tissue [64]. For example, NetWAS [65] analyzes a network of genes and tissue expression data to identify disease associations.

Some of the reviewed ML models in the current section require the handcrafted generation of graph features (e.g., distance between a gene and a disease) for training a model to classify genes as likely to be associated with a disease or not. Handcrafted feature generation is time consuming and requires domain knowledge. In contrast, graph representation learning methods [19, 21, 22] automatically learn graph features, as discussed next.

2.5 Graph representation learning

Automatic feature learning from graphs [17, 19, 21–24] has been widely studied using methods such as matrix factorization and graph embeddings.

Matrix factorization methods are used for predicting previously unknown gene-disease edges. For example, the PCFM algorithm [24] uncovers hidden factors for genes and diseases from a gene-disease association matrix using a probability-based collaborative filtering model to predict disease genes. Manifold learning [66] utilizes a gene-disease association matrix to learn latent factors of genes and diseases, following the assumption that disease genes are closely located on the graph. Medusa [67] analyzes 16 heterogeneous graphs as matrices to

establish connections between non-neighboring nodes in each graph. GeneHound [68] first integrates data including literature-based phenotype and gene information. Then, it performs Bayesian matrix factorization to uncover latent factors for genes and diseases to predict new gene-disease associations.

Graph embedding methods represent nodes as numerical vectors in a low-dimensional space while preserving node and graph topological similarity using neural networks [19]. The goal of graph embedding methods is to capture the topological information of nodes and edges. Graph embeddings were successfully used in capturing the biological structures of proteins [25] and reducing data noise in graphs [26] by using tasks such as node classification, link prediction, and clustering [21].

Examples of node embedding algorithms include the SkipGram algorithm [69] that constructs associations between a node and its neighbors via random walks. DeepWalk [23] expands SkipGram to perform random walks on a graph by treating nodes as words. It was used to learn node embeddings in biological graphs for tasks such as predicting drug-target associations [70] and protein function [71]. SmuDGE [28] expands SkipGram to predict novel disease genes by combining disease–phenotype and gene–phenotype associations to generate a corpus for SkipGram-based representation learning. Then, it predicts gene–disease associations using a neural network. Building upon SkipGram, HeteWalk [20] constructs a weighted heterogeneous network by joining six public data sources including PPI, miRNA similarity network, and disease phenotype similarity network, and then performs SkipGram-based network embedding. The HIN2Vec algorithm [72] generates node embeddings for heterogeneous networks based on random walks using a three-layer neural network model, but it samples only short paths, making it inefficient for large graphs [73].

The node2vec algorithm [21] finds an embedding function such that the conditional probability of observing the neighbors of a node is maximized. It extends DeepWalk, but employs more sophisticated random walks using four parameters to select the next visited nodes: 1) number of random walks from each node, 2) walk length, 3) P—the probability to return to a previously visited node, and 4) Q—the probability to explore undiscovered nodes. The node2vec algorithm is widely used for generating node embeddings, and it presents superior performance in node classification tasks on biological networks [74]. Leveraging node2vec, several biological studies combine node2vec embeddings with other features. For example, the N2VKO algorithm [75] integrates node2vec embeddings extracted from a PPI network with biological annotations for gene-disease association prediction.

In network-based models, the structure of the network must accurately represent biological knowledge such as gene-disease associations; otherwise, feature learning will be harmed. ML models for predicting unknown disease genes might be biased because of missing gene-disease edges in the graph, due to, e.g., data that has not yet been collected (unknown). Missing edges also cause those models to ignore global information from unconnected nodes since most network-based gene-disease prediction models are limited to local node-to-node propagation. In addition, to train an ML classifier for the task of link prediction, positive (existing gene-disease edges) and negative (non-existing gene-disease edges) examples are needed. Whereas sampling positive edges from a graph is straightforward, sampling negative edges involves sampling a pair of an unconnected gene and disease. Such pairs might be biologically associated but not yet known, thus falsely used as a negative example, leading to poor model performance.

Lack of studies that consider in a single model knowledge from both local network neighbors and non-neighbors; incorporate rich knowledge from several biological domains; and address the problem of sampling negative edges for link prediction raises the need for developing new models that will enable better prediction of disease genes. We address those limitations by developing the Heterogeneous Integrated Graph Model for Predicting Disease

Genes in humans (HetIG-PreDiG) model to improve the prediction of disease genes, as detailed next.

3 Materials and Methods

3.1 HetIG-PreDiG model to predict gene-disease associations

The following five steps describe the construction of the developed Heterogeneous Integrated Graph Model for Predicting Disease Genes in humans (HetIG-PreDiG) model for predicting gene-disease associations.

Step 1—**Construct a heterogeneous graph** using two types of data: (i) gene-disease associations, and (ii) gene-tissue interactions.

Using the first data type, gene-disease associations are represented as a graph, denoted by $G_{gd} = (D_{gd}, V_{gd}, E_{gd})$. Nodes represent the sets of diseases D_{gd} and genes V_{gd}. Edges E_{gd} represent gene-disease associations.

Using the second data type, gene-tissue interactions are transformed into a gene-gene graph $G_{gg} = (V_{gg}, E_{gg})$. Nodes V_{gg} represent genes that are connected by an edge $e_{ij} \in E_{gg}$ if genes $v_i, v_j \in V_{gg}$ were reported in the same tissue.

Finally, both graphs are integrated into an undirected gene-disease heterogeneous graph $G(D, V, E) = G_{gd} \cup G_{gg}$. In G, nodes represent the sets of diseases D and genes V. Edges E represent gene-disease associations and gene-gene associations. To integrate both graphs into a heterogeneous graph, gene-gene edges $e_{ij} \in E_{gg}$ were excluded if both gene v_i and gene v_j are not in G_{gd}. More formally, we define: $G(D, V, E) = G_{gd} \cup G_{gg} = \{(D_{gd}, V_{gd} \cup V_{gg}, E_{gd} \cup E_{gg}) | \forall e_{ij} = (v_i, v_j) \in E_{gg}, \exists v_i, v_j \in V_{gd}\}$. This allows us to include additional knowledge where: (i) at least one gene in V_{gg} is in V_{gd} and the second gene of the edge is not in V_{gd}, or (ii) both genes in V_{gg} are in V_{gd}.

Step 2—**Generate a labeled set**. We aim to predict missing links between unconnected gene and disease nodes. Given a network with missing links (due to yet-undiscovered knowledge), we aim to predict these missing links.

Data imbalance is a known challenge when designing machine learning models, as in the case of predicting disease genes. The abundant (majority) class contains more data than the minority class. In a gene-disease network, the number of non-associated gene and disease nodes (expressed as missing edges in the majority class) far exceeds that of disease-associated genes in the minority class. The imbalanced data presents a challenge for identifying gene-disease associations. Most traditional machine learning methods are usually biased towards the majority class, and hence lead to loss of predictive performance for the minority class. Sampling methods for dealing with imbalanced datasets are frequently used [76, 77]. We applied a sampling method of positive and negative examples that is similar to other studies that used the node2vec algorithm for representing biological entities in a graph [21, 78]. The described sampling method addresses the imbalanced learning problem by sampling an equal number of negative samples and positive samples for training [78], thus ensuring that the model is not biased toward any class.

We generate the labeled dataset of edges by following three sub-steps. First, obtain positive examples by randomly selecting 20% of gene-disease edges for each disease node in G, and removing them from G, thus generating $G' = (D', V', E')$. This sub-step results in N positive examples of gene-disease edges. Second, obtain negative examples by randomly sampling an equal number (N) of unconnected node pairs composed of $N/2$ gene pairs, and $N/2$ gene and disease pairs. For each disease, we select non-associated genes with the lowest GDPS score (described in detail in Step 4). This process is iteratively repeated until the desired number

($N/2$) of non-existing gene-disease associations is achieved. Finally, the labeled dataset is split into Train (70%) and Test (30%) sets.

Step 3—**Learn node embedding vectors**. Previous studies (e.g., [21, 78]) found that deep learning techniques that use embedding vectors obtain better representation of biomedical entities (such as genes and diseases), and thus, improve prediction performance. We map nodes in G' to a low-dimensional feature space using node2vec. The vectors of each pair of nodes u, v in the Train and Test sets are aggregated into a single vector ($u + v$). We chose the node2vec algorithm since it was reported to obtain richer topological representation of the network than traditional methods [78].

Step 4—**Compute a Gene-Disease Prioritization Score (GDPS)** using a third type of data, gene-gene co-expression similarity. Compute GDPS using the developed Algorithm 1 (Fig 1) for each gene and disease pair, following [30]. GDPS uses a gene-gene co-expression similarity matrix to compute the average similarity of a gene to known disease genes as expressed by the structure of G'. The higher the GDPS value, the more likely a gene has similar functions to the disease genes, and the more likely that gene is to associate with the disease. We set GDPS for gene-gene edges to 0. When calculating the GDPS score, we use the graph G', which is the original graph G with test edges removed. This means that only gene-disease associations that are observed in G' are used to calculate GDPS. Fig 2 presents an example of Algorithm 1 (Fig 1).

Step 5—**Train ML classifier**. Concatenate the aggregated vector $u + v$ (Step 3) with the Gene-Disease Prioritization Score (Step 4) into a single feature vector for training and testing an ML model to classify the pairs of nodes $\{u, v\}$ in the Train and Test sets into one of two groups: a link will/not form. Concatenating the GDPS score to the embedding vector allows the model to learn non-linear relationships between a gene and a disease that take gene expression information into account, which is innovative [78]. To evaluate the contribution of GDPS, we compare the performance of the developed HetIG-PreDiG model with and without GDPS using 10-fold cross-validation. The best model is trained using all training examples, and then its performance is evaluated on the Test set.

Algorithm 1 : Calculate Gene-Disease Prioritization Score (GDPS)

Input: $Z_{n\times n} \leftarrow$ gene-gene co-expression similarity matrix,
$G' = (D', V', E')$: Graph G without train edges,
Disease $d \in D'$

$Z'_{n\times m} \leftarrow \{col(Z) \mid (gene, d) \in E', gene \in col(Z)\}$
$GDPS_{n\times 1} \leftarrow 0$
$i \leftarrow 0$
for $r \in rows(Z')$ **do**
    $GDPS_i \leftarrow \bar{r}$
    $i \leftarrow i + 1$
return $GDPS$

col: column; n: number of genes; m: number of genes associating with d; $\bar{r}$: row average

Fig 1. The developed algorithm to calculate Gene-Disease Prioritization Score (GDPS).

https://doi.org/10.1371/journal.pone.0280839.g001

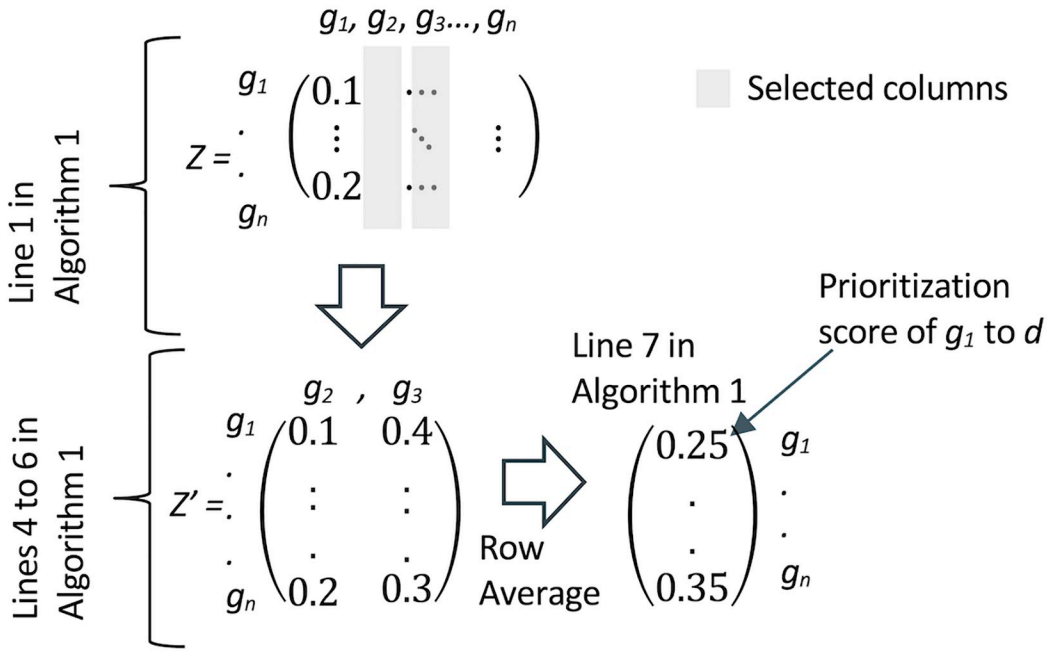

Fig 2. An illustrated example of Algorithm 1 (Fig 1). Z is a gene-gene co-expression similarity matrix. Line 1 in Algorithm 1: given disease d that is known to associate with genes g_2 and g_3, columns 2 and 3 are selected in Z to create matrix Z'. Lines 4 to 6 in Algorithm 1: each row in Z' is averaged into a Gene-Disease Prioritization Score. The prioritization score of g_1 for d is (0.1+0.4)/2 = 0.25, reflecting the average similarity of a non-associated gene g_1 to genes associated with d.

https://doi.org/10.1371/journal.pone.0280839.g002

3.2 Datasets used in model

Here we describe the datasets used for predicting gene-disease associations (Table 1).

Dataset 1 (DS1) consists of DisGeNET V7.0 [32]. DS1 contains one of the largest publicly available collections of genes associated with human diseases curated from expert repositories, GWAS catalogs, and scientific literature. DS1 contains data including `DiseaseSemanticType`—the semantic type of the disease (e.g., 'Anatomical Abnormality', 'Pathologic Function', or 'Disease or Syndrome'); `Gene symbol; Disease id; Disease name; DiseaseType` ('disease', 'phenotype', or 'group'); `Disease Specificity Index (DSI)` —ranges from 0.25 to 1 and reflects if a gene is associated with few diseases; a gene that associates with multiple diseases has a lower DSI. `Disease Pleiotropy Index (DPI)` ranges from 0 to 1 and reflects if multiple diseases that associate with a gene are similar in terms of belonging to the same Medical Subject Headings (MeSH) disease class. A gene that associates with diseases of different MeSH classes has a high DPI index; and `YearInitial` $\in$ [1940, 2020]—the year that the gene-disease association was first reported.

Table 1. Source datasets.

Dataset	DisGeNET	Human Protein Atlas	ARCHS4
Genes	21,671	15,308	35,238
Diseases	30,170	–	–
Tissues	–	63	–
Interactions	1,134,942	653,706	$(35{,}238)^2$

https://doi.org/10.1371/journal.pone.0280839.t001

Dataset 2 (DS2) consists of the Human Protein Atlas version 20.1 and Ensembl version 92.38 with information on gene-tissue interactions. DS2 is a representative tissue-specific gene expression resource with a large and comprehensive distribution of protein-coding genes in human tissues and cells [33]. DS2 contains expression profiles for proteins in human tissues with `Ensembl gene id`, `Tissue name`, `Expression level` ('High', 'Medium', 'Low', and 'Not Detected'), and the gene `Reliability` ('Approved', 'Enhanced', 'Supported', and 'Uncertain') of the expression value.

Dataset 3 (DS3), the ARCHS4 database [30], covers the majority of published RNA-seq data. It contains gene counts for humans and mice from the Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) platforms. In this study, only human gene expression data is considered. Specifically, this study uses the available pairwise Pearson correlation data of human genes across expression samples to create a gene-gene co-expression similarity matrix.

4 Analysis

This section begins by describing the implementation of the developed HetIG-PreDiG model, following the five Method steps in Section 3.1.

In Step 1, a gene-disease graph G_{gd} based on DS1 is constructed. For `DiseaseType`, only 'disease' was selected, and for `DiseaseSemanticType`, only 'Disease or Syndrome' was selected. We removed diseases with fewer than two genes, resulting in 5,417 diseases, 13,011 genes, and 179,860 gene-disease associations.

Next, we constructed a gene-gene graph based on gene-tissue data (DS2). To consider highly validated information, only gene-tissue associations that have an 'Approved' `Reliability` with 'High' `Expression level`, and genes associating with a tissue in more than a single cell type were included. This resulted in 25 tissues, 1,661 genes, and 5,191 gene-tissue interactions. The distribution of the number of genes per tissue presents an exponential decay shape with an average of 206.64 genes per tissue, a median of 195, a maximum of 463 genes associated with the *tonsil* tissue, and a minimum of 2 genes associated with *retina*. Next, we converted the selected data into a gene-gene graph G_{gg} such that genes found in the same tissue are connected by an edge. Finally, both graphs (G_{gd} and G_{gg}) are combined into a heterogeneous graph G with 5,417 diseases, 13,637 genes, 179,860 gene-disease associations, and 444,094 gene-gene associations.

In Step 2, given a graph G, we labeled a set of positive and negative examples as described in Section 3.1. This step results in the creation of graph G' by deleting positive examples (edges) from G. Next, the examples are shuffled, and split into Train (70%) and Test (30%) sets.

In Step 3, node embeddings are learned by applying node2vec to G' using the following parameters: `embedding dimensions = 64`, `walk length = 5`, `number of random walks = 10`, `p = 1`, and `q = 1`. For `p = 1`, the algorithm is less likely to revisit a node, resulting in moderate exploration and avoiding 2-hop redundancy. Setting `q = 1`, the algorithm is not biased towards visiting closer or farther nodes to the current node.

In Step 4, GDPS is computed using DS3 for each pair of gene and disease in the Train and Test sets.

To summarize, DS1 and DS2 were used to create nodes and edges in the analyzed network. DS3 was used to create the GDPS score. Specifically, we used gene expression `level` in DS2 only for selecting data for further analysis. DS3 data of gene counts (expression) is used for generating GDPS and is not used for graph structure.

Finally, in Step 5, we trained a logistic regression classifier using the Train set to classify pairs of nodes into one of two classes: 1) link formation (i.e., a gene is associated with the disease), or 2) no link formation.

4.1 Comparison with baseline models

To evaluate the performance of the developed HetIG-PreDiG model (Section 3.1), the Train and Test sets were used to train and evaluate the following algorithms for predicting gene-disease associations as baseline models: RWRH, N2VKO, and HIN2Vec. These models were selected since they are frequently used and were reported to produce good performance on biological datasets [21, 47, 75]. The evaluated models were assessed using the implementations described by their authors, with their suggested parameters. In addition, we performed an ablation study by comparing the performance of HetIG-PreDiG with and without GDPS using 10-fold cross-validation.

4.2 Model evaluation: Predicting novel disease genes

Instead of predicting specific gene-disease edges in the Test set (as in Section 3.1, Step 5), we adopt a more realistic approach in which, for a given disease, all candidate (non-associating) genes are to be examined. Stated differently, given a disease and a set of candidate genes, we aim to predict gene-disease associations. We replicate a scenario at a point in time when cumulative knowledge exists and other knowledge is missing (e.g., has yet to be discovered). Cumulative knowledge is represented by $G' = (D', V', E')$, reflecting gene-disease associations in the Train set. We define missing knowledge as gene-disease associations discovered in the year 2020 (DS_{2020}) (the most recent gene-disease discovery year in DS3) that appear in the Test set. We predict those associations by considering *all* candidate genes in G' that are not associated with a disease $d \in D' \cap DS_{2020}$. More formally, we seek to predict the following gene-disease associations $\{(g, d) | g \in V' \wedge d \in D' \cap DS_{2020} \wedge (g, d) \notin E'\}$.

We assume that gene-disease associations in DS_{2020} are unknown during the prediction process only. When the prediction process is complete, we evaluate our model for predicting novel disease genes (candidate genes predicted to associate with d), using the developed Overlap measure (1):

$$Overlap(f_{\%}) = |P_{f_{\%}} \cap I_d| / |I_d| \quad (1)$$

$P_{f\%}$—The set of candidate genes predicted to associate with $d \in DS_{2020}$, located in the top $f_{\%}$ of a ranking, based on the gene's prediction probability.

I_d—The set of genes associated with $d \in DS_{2020}$.

The higher the Overlap score, the more successful the developed model at identifying genes that associate with a disease. When a disease $d \in DS_{2020}$ has a high Overlap score, the genes predicted to associate with d can be further validated in wet-lab experiments. An Overlap of 1 indicates that all genes that are known to associate with d were identified, and an Overlap of 0 indicates that no genes associating with d were identified.

Increasing $f_{\%}$ will result in the inclusion of more genes (e.g., $f_{\%} = 1$ considers all candidate genes) leading to a higher Overlap score, but also requires experimenting with more candidate genes in a wet lab. To optimize the set of candidate genes, we define the *Ratio* (2) between the number of diseases with Overlap = 1 and $f_{\%}$. We aim to maximize the number of diseases with Overlap = 1 and minimize $f_{\%}$, i.e., find the highest Ratio.

$$Ratio = |\{d \in Diseases | Overlap(d) = 1\}| / f_{\%} \quad (2)$$

We construct the set DS_{2020} as follows: first, we select all gene-disease associations in DS1 with `YearInitial` = 2020; `diseaseType` = 'disease'; and `diseaseSemanticType` = 'Disease or Syndrome'. Second, we remove diseases with only a single gene association discovered in the year 2020. Third, we keep only gene-disease edges that are not in G'. Fourth, we discard disease associations with highly connected genes (which are more trivial to predict) by selecting genes in DS1 with a `DSI` above the average `DSI` of all genes, and with `DPI` below the average `DPI` of all genes.

Next, we follow five steps to predict novel gene-disease associations. First, given $G' = (D', V', E')$ and a disease $d \in D' \cap DS_{2020}$, generate embedding vectors of each candidate gene $g \in V'$ not associated with d (i.e., $(g, d) \notin E'$), and embeddings of d. Second, sum embedding vectors of each candidate gene and d. Third, for each candidate gene, compute its GDPS score with d and concatenate it to the aggregated embedding vector in the Second step. Fourth, apply HetIG-PreDiG for each candidate gene-disease pair to predict via classification whether a link will form between them (the 'link formation' class), or not. Fifth, keep only novel genes that are predicted to associate with d (i.e., classified into the 'link formation' class) with their probability of affiliating with that class (prediction probability is used later in Section 5.3). Fig 3 summarizes the steps of model evaluation.

Next, given a disease, we estimate the success of our 'HetIG-PreDiG with GDPS' model's success in predicting gene-disease associations without the ability to calculate the Overlap score.

4.3 Model evaluation: Real-world scenario

In a real-world scenario, there is no test set. We can predict novel gene-disease associations similar to Section 4.2, but we cannot calculate the Overlap score (I_d is unknown). Hence, we cannot evaluate our predictions' correctness. Given a single disease, we aim to estimate the prediction performance of our model in identifying novel gene-disease associations. While model performance is typically evaluated based on the prediction success of the entire test set (e.g., F1 score), a single prediction might be wrong (e.g., false positive); thus, we do not know if our single prediction is successful.

To overcome this challenge, we identify via classification, diseases that are more likely to have higher Overlap values. We learn the patterns of the prediction probabilities of genes that were predicted to associate with diseases in Section 4.2. Prediction probabilities have been found informative for biological problems such as estimating the plasma effect-site equilibration rate constant [79] and anesthetic depth [80]. Under the assumption that prediction probabilities contain meaningful information about the Overlap score, we learn their patterns by: (i) calculating the mean prediction probability of gene-disease associations for each disease $d \in DS_{2020}$; (ii) analyzing the mean prediction probabilities for diseases $d \in DS_{2020}$ to automatically detect the optimal number of clusters using the Ckmeans.1d.dp R library [81] that is a variant of K-means for one-dimensional data; and (iii) classifying each $d \in DS_{2020}$ to clusters of different Overlap range using the Ckmeans.1d.dp R library.

To summarize, using the 'HetIG-PreDiG with GDPS' model, we first predict the genes associated with a disease. Then, using the developed classifier in Step *iii*, we can classify the disease based on its prediction probabilities and estimate its Overlap score based on d's cluster affiliation. For high Overlap (~ 1), experiments in a wet lab may be warranted.

We make predictions for each $d \in DS_{2020}$ using all candidate genes in the dataset. Then, we determine whether the top-ranked genes predicted by the developed model are novel by conducting an automated literature search to find papers indexed by PubMed that support the predicted gene-disease associations. Using the PubMed API, we collect papers containing the

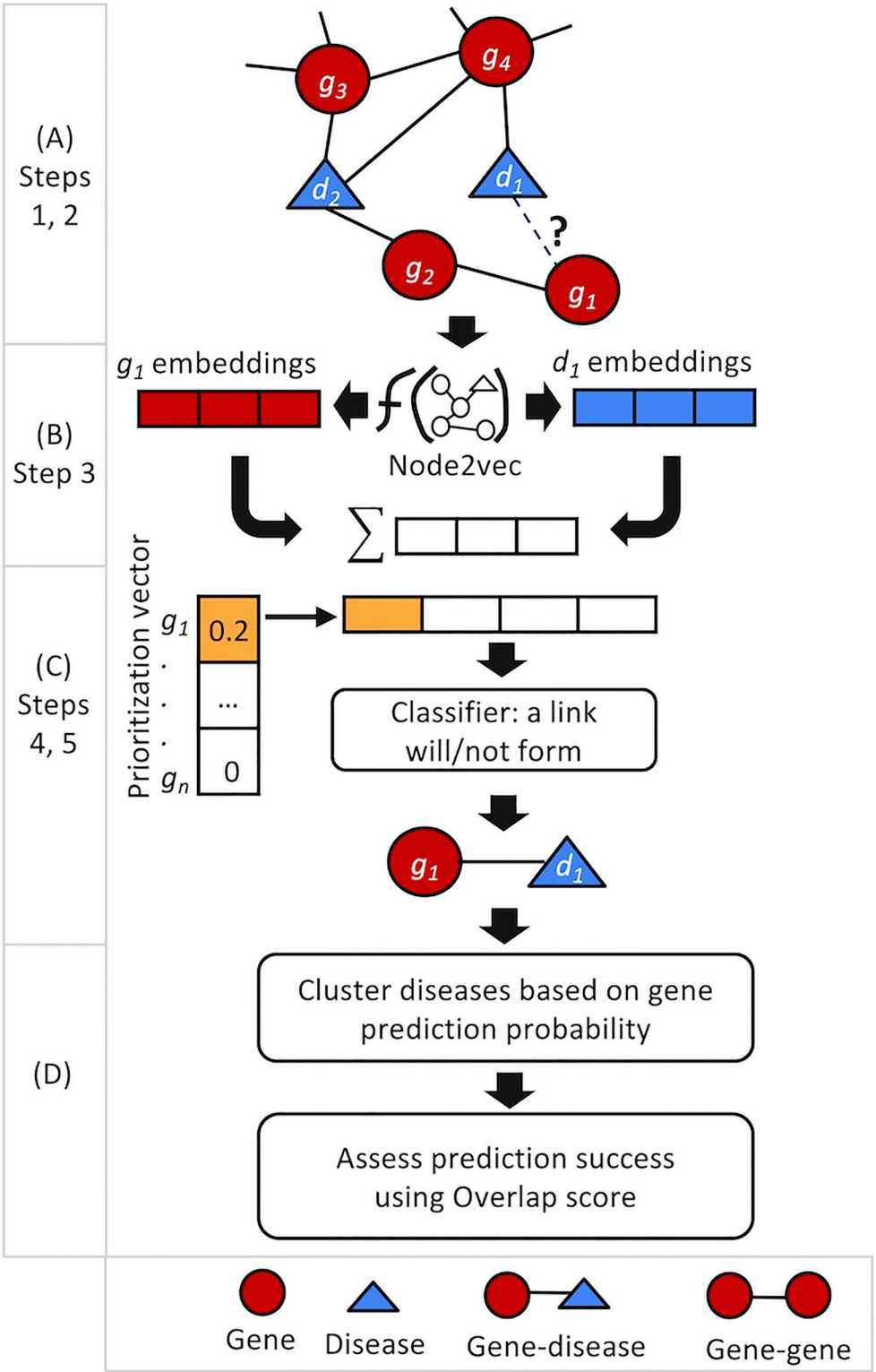

Fig 3. An illustration of the developed model using an example: Predict if gene g_1 is associated with disease d_1. (A) Steps 1,2 (Section 3.1): Heterogeneous graph (G') with genes and diseases after removal of selected Train/Test gene-disease edges. (B) Step 3: Learn node embeddings. The embeddings of d_1, g_1 are aggregated into a single vector denoted by Σ. (C) Steps 4, 5: Compute Gene-Disease Prioritization Scores. Compute GDPS(g_1, d_1) in the illustrated example. Gene expression data involving gene function is integrated into the model via the GDPS prioritization score that

provides expression similarity across multiple tissues. Next, the GDPS prioritization score is concatenated with the vector Σ and fed into a classifier that is trained to predict associations between pairs of genes and diseases. In the illustrated example, GDPS(g_1, d_1) (colored in orange) is concatenated to vector Σ. Finally, in the current example, the developed model predicts a new edge between g_1 and d_1. (D) Assess success in predicting novel genes: cluster a disease based on its gene prediction probabilities.

https://doi.org/10.1371/journal.pone.0280839.g003

disease's name and its top-ranked predicted genes together within the title and/or abstract fields. We further validate the reported papers by manually examining the complete set of results for a subset of the diseases.

4.4 Comparison of gene-disease association predictions with the literature

In this analysis, the predictions of HetIG-PreDiG are evaluated by demonstrating its capability to predict novel gene-disease associations that are not present in DS1. Studies are presented that support the highest top-ranked genes predicted to associate with a disease. We focus on the diseases in DS_{2020}. For each of the Top 10 predicted genes for each disease, we searched for supporting literature to determine whether those genes are novel using automated and manual searches.

Using an automated search, we surveyed the literature using PubMed's API as described in Section 4.3 to find existing associations between each disease and its Top 10 predicted genes. To evaluate the developed model's success in detecting novel disease genes, we calculated the *Success Rate* defined by the ratio between the number of genes with supporting literature evidence and the number of top-ranked genes (i.e., 10).

5 Results

5.1 Comparison with baseline models

The prediction performance of the HetIG-PreDiG model outperformed the prediction performances of the baseline models, as listed in Table 2, which reports the average and standard deviation of F1 score across the 10-fold cross-validation.

Regarding the ablation study, HetIG-PreDiG with GDPS outperformed HetIG-PreDiG without GDPS (Table 2). When evaluated on the Test set, HetIG-PreDiG with GDPS outperformed (Recall 0.93, Precision 0.97, and Micro-F1 score 0.95) the HetIG-PreDiG model without GDPS (Recall 0.87, Precision 0.88, and Micro-F1 score 0.88). Fig 4 shows model performances evaluated using the Receiver Operating Characteristic (ROC) curve, and the Area Under the ROC Curve (AUC). The addition of the GDPS score improves the prediction performance of model 'HetIG-PreDiG with GDPS' compared with model 'HetIG-PreDiG without

Table 2. Comparison of the developed model (HetIG-PreDiG) with the baseline models using 10-fold cross-validation for prediction of the top 30% of predicted disease genes. The source code of baseline models is listed. The Micro-F1 score column indicates average and standard deviation model performance of predicting gene-disease associations.

Model	Description and Source Code	F1 score
HetIG-PreDiG	The developed model with the developed GDPS score. https://github.com/bartala/disease_gene	0.95 ± 0.021
HIN2Vec	Generates node embeddings for heterogeneous networks based on random walks using a three-layer neural network model, but it samples only short paths, making it inefficient for large graphs. https://github.com/csiesheep/hin2vec	0.89 ± 0.036
HetIG-PreDiG no GDPS	The developed model with node embeddings only, without Gene-Disease Prioritization Score (GDPS).	0.88 ± 0.028
N2VKO	Integrates node2vec embeddings extracted from a PPI network with biological annotations for gene-disease association prediction. https://github.com/sezinata/N2VKO	0.82 ± 0.131
RWRH	Extends RWR on a heterogeneous phenotype-gene network. https://github.com/alberto-valdeolivas/RWR-MH	0.79 ± 0.058

https://doi.org/10.1371/journal.pone.0280839.t002

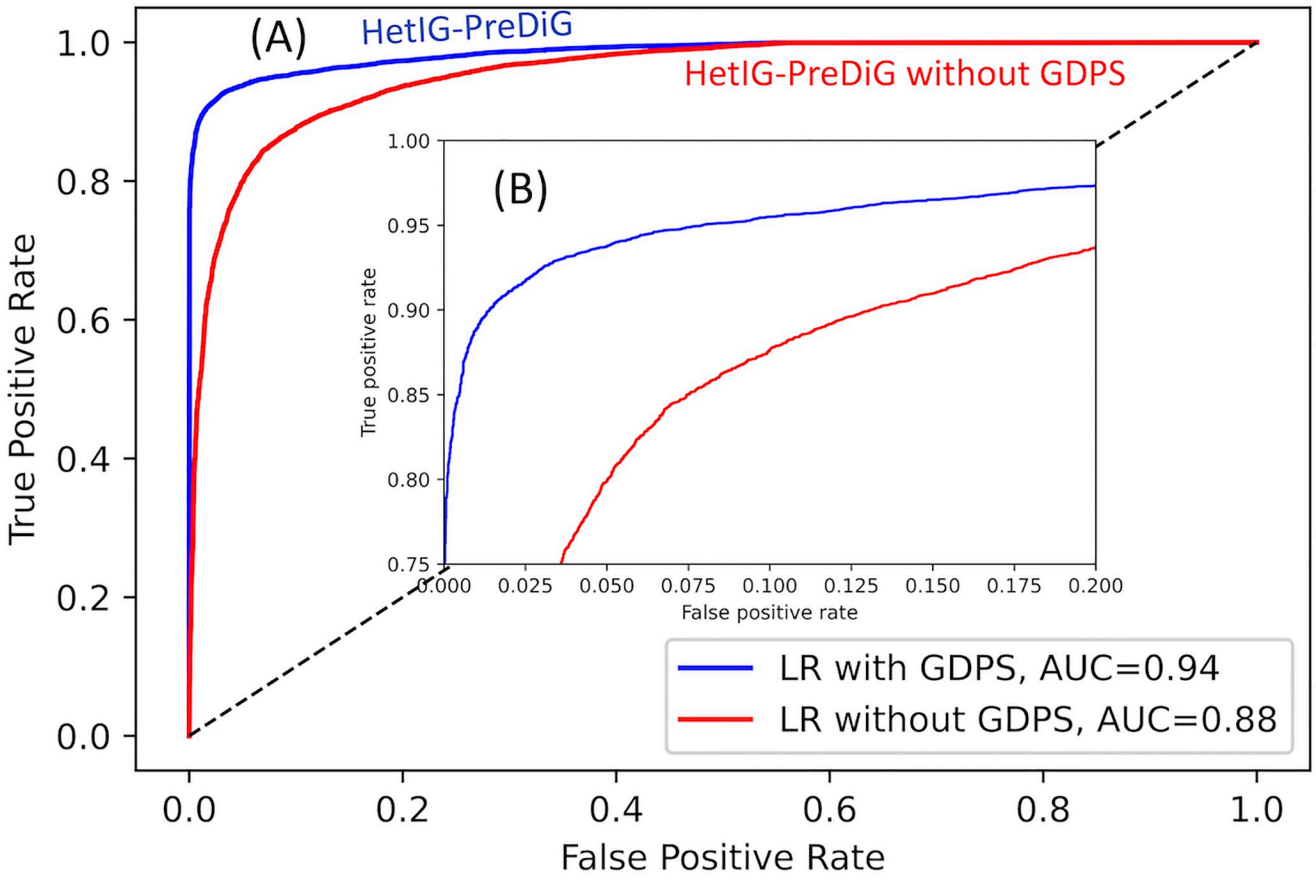

Fig 4. An ablation study. Demonstrating the effectiveness of using the developed Gene-Disease Prioritization Score (GDPS) in the developed HetIG-PreDiG model that incorporates a logistic regression (LR) classifier. (A) Receiver operating characteristic (ROC) curves of the developed model with (blue), and without (red) GDPS. The area under each ROC curve (AUC) is indicated. (B) A zoomed view of the top left corner of Fig 4A.

https://doi.org/10.1371/journal.pone.0280839.g004

GDPS' (Fig 4). The HetIG-PreDiG model developed in this study considers information not only from the structure of the network using node embeddings but also from biological information not reflected in the network, using gene co-expression similarity.

The next section demonstrates the capability of the developed HetIG-PreDiG model for predicting novel gene-disease associations.

5.2 Model evaluation: Predicting novel disease genes

To evaluate the model's success in predicting novel gene-disease associations, we followed the steps described in Secion 4.2. This resulted in 30 diseases in DS_{2020} (Table 3) associating with 11,667 genes. For each disease d, we computed the Overlap score for different $f_{\%} \in [10, 100]$ rankings, with steps of 10. We focus on $f_{\%} = 30\%$ since it presented the highest Ratio in Eq. (2) as illustrated in Fig 5. The Overlap score was 1 for 10 diseases; it was in the range (0.5, 1) for 2 diseases; and it was in the range [0, 0.5] for 18 diseases, indicating the identification of all, most, and few associating genes predicted for $f_{30\%}$, respectively. As expected, we find that the Overlap score increases as the $f_{\%}$ increases. In contrast to $f_{\%} = 30\%$, at $f_{\%} = 100\%$, the model predicts an Overlap = 1 for 22 diseases, Overlap $\in$(0.5, 1) for 4 diseases, and an Overlap $\in$[0.5, 1] for 4 diseases.

Table 3. A summary of literature evidence for the Top 10 predicted genes that are not associating with a disease for each of the selected 30 diseases. The Success Rate column is the ratio between the number of predicted genes with PubMed supporting evidence divided by 10.

Disease Name	Success Rate	Supporting PubMed Papers
Alzheimer's Disease	1.0	91
Anemia	0.9	453
Anorexia	0.9	385
Arthritis	0.9	551
Obesity	0.9	102
Parkinson Disease	0.9	76
Colitis	0.8	992
Dermatitis	0.8	243
Diabetes Mellitus	0.8	150
Myocarditis	0.8	33
Degenerative Polyarthritis	0.8	23
Rheumatoid Arthritis	0.8	177
Sepsis	0.8	204
Crohn's Disease	0.7	52
Hyperglycemia	0.7	117
Liver Fibrosis	0.7	90
Alopecia	0.6	134
Coronary Artery Disease	0.6	13
Diabetic Nephropathy	0.6	12
Hepatitis B	0.6	17
Hepatitis C	0.6	17
Retinitis Pigmentosa	0.6	28
Cerebral Infarction	0.5	30
Hypertrophic Cardiomyopathy	0.5	49
Idiopathic Pulmonary Fibrosis	0.5	35
Retinopathy of Prematurity	0.5	6
Anophthalmia and Pulmonary Hypoplasia	0.4	4
Chagas Disease	0.4	18
Septicemia	0.4	62
Amyloidosis	0.3	23

https://doi.org/10.1371/journal.pone.0280839.t003

Table 3 provides a summary of the literature evidence for the Top 10 predicted genes not associated with a disease for each of the selected 30 diseases. The Success Rate column is the ratio between the number of predicted genes with PubMed supporting evidence divided by 10. For example, for the disease Anemia, 9 of the Top 10 predicted genes for this disease had PubMed literature support, with a total of 453 PubMed entries supporting these 9 gene-disease predictions.

5.3 Model evaluation: Real-world scenario

Using the prediction probability results of Step 5 in Section 5.2, we learn the patterns of the prediction probabilities of candidate genes predicted to associate with diseases d. Following Step i (Section 4.3), we calculate the mean prediction probability of gene-disease associations for each $d \in D' \cap DS_{2020}$. In Step ii, we analyze the mean prediction probabilities for diseases d and automatically detect the optimal number of clusters using the Ckmeans.1d.dp R library

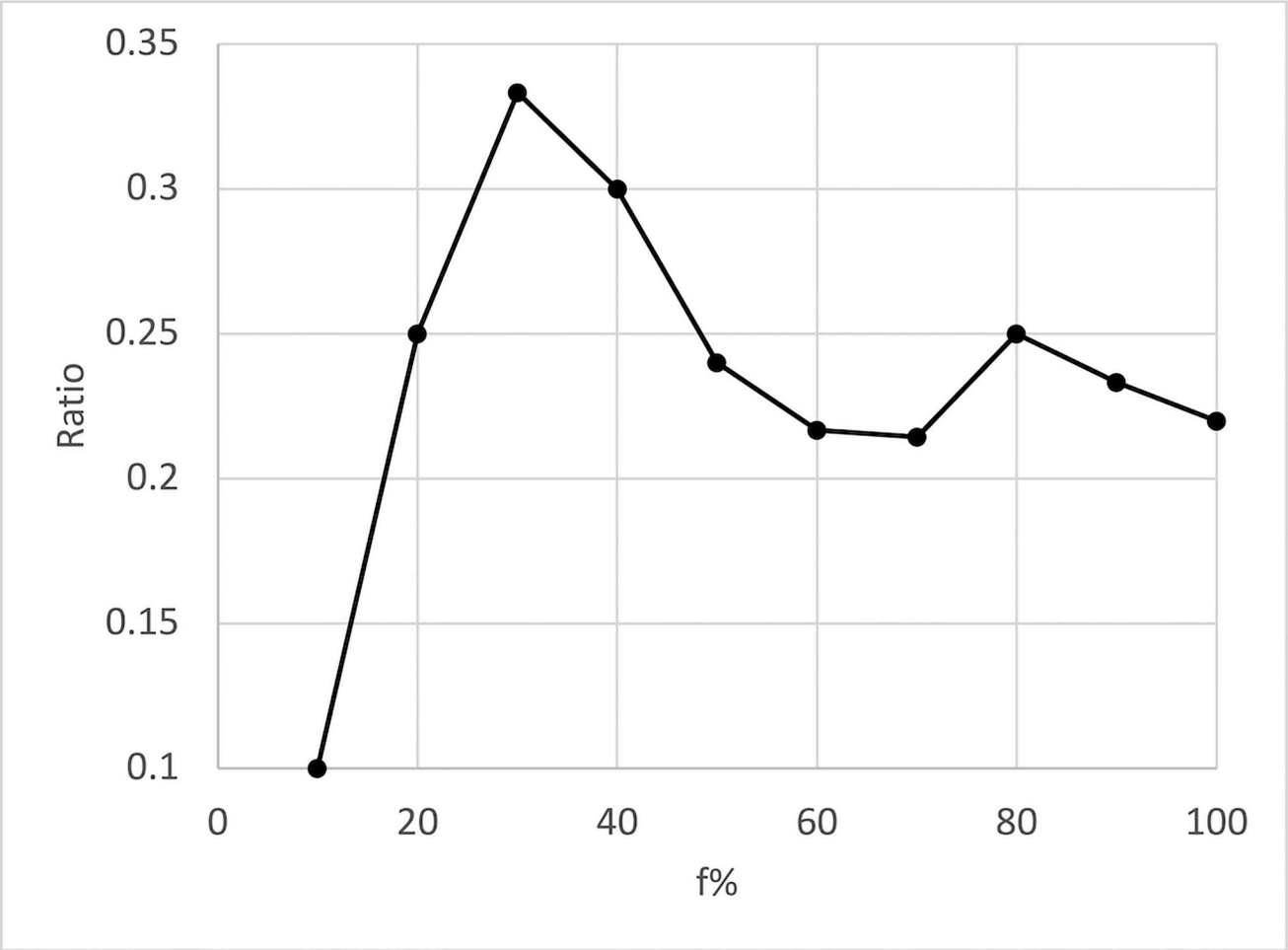

Fig 5. Ratio between the number of diseases with Overlap = 1 and $f_{\%}$. The highest ratio is achieved at $f_{\%}$ = 30%.

https://doi.org/10.1371/journal.pone.0280839.g005

[81]. The search for the optimal number of clusters was performed in the range of 2 to 9. Three clusters were identified. Lastly, in Step *iii*, we assign each *d* to a cluster based on its mean prediction probability.

Based on the Overlap scores (calculated in Section 5.2) of diseases in each cluster, the clusters were assigned a respective range of Overlap: (a) $0 \leq$ Overlap ≤ 0.5, (b) $0.5 <$ Overlap < 1, and (c) Overlap = 1 to capture (a) few, (b) most, and (c) all associating genes predicted in $P_{f\%}$, respectively.

Fig 6A presents the prediction probabilities distributions of an Overlap score for the top $f_{\%}$ = 30% in each of the three clusters. Using a Kruskal-Wallis rank (KW) test [82], we find significant differences ($P_{value} < 2.2e^{-16}$) in the density for the three classes. To detect where those differences lie, we conducted three additional KW tests between each pair of classes, and again find significant differences. Fig 6B presents a breakdown of the diseases that affiliate with each cluster in Fig 6A.

To summarize, given a disease, in a real-world scenario where the Overlap score cannot be calculated, we first predict disease genes, and then assign the disease to one of the three Overlap clusters based on the mean gene prediction probabilities. While all models make mistakes,

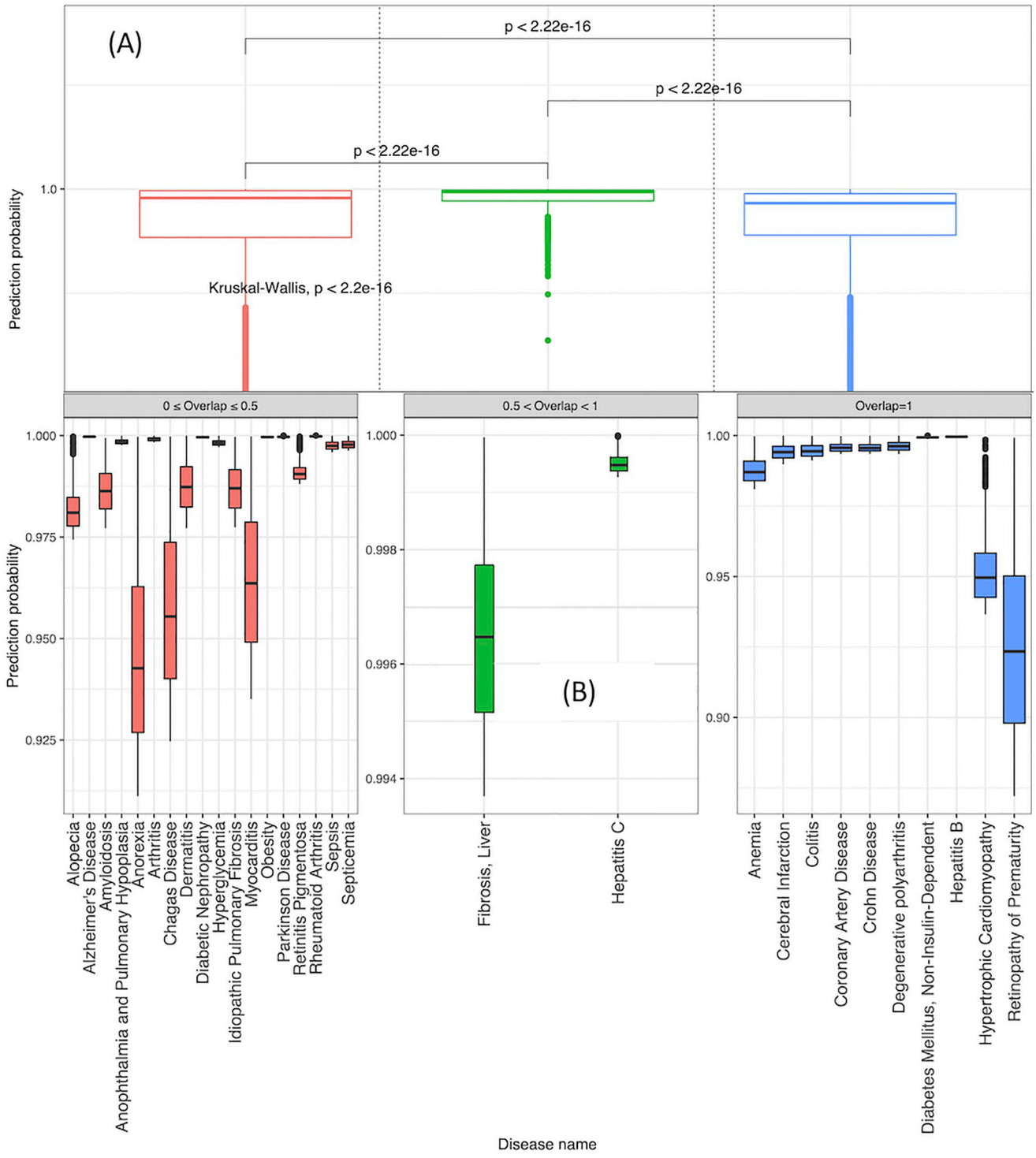

Fig 6. Box plots of the top $f_{\%}$ = 30% prediction probabilities of genes predicted to associate with a disease. (A) three levels of Overlap clusters, and (B) a breakdown showing the diseases affiliating with each cluster in (A).

https://doi.org/10.1371/journal.pone.0280839.g006

our clustering method enables us to systematically estimate the model's success in predicting novel disease genes.

While the developed model can successfully predict gene-disease associations, other predicted genes for a given disease *d* that were not yet validated might also associate with *d*, as demonstrated next.

5.4 Comparison of gene-disease association predictions with the literature

Table 3 presents a summary of the 30 diseases in DS_{2020}, their corresponding Success Rates, and the number of supporting PubMed papers reported. Using a manual search, we reviewed a subset of the 30 diseases in Table 3 for which $\leq$100 publications were reported.

Representative of the developed HetIG-PreDiG model's gene-disease association prediction quality, Table 4 provides reported literature support for the Top 10 predicted genes from the developed model, for three of the diseases listed in Table 3: Hepatitis B, Chagas Disease, and Diabetic Nephropathy. We selected those three diseases since they represent a wide range of

Table 4. Literature review for Top 10 predicted genes for 3 example diseases. Genes are listed in descending ranked order. For example, for Hepatitis B, CASP1 and BAK1 are ranked 1 (best) and 10, respectively.

Disease	Gene	Supporting PubMed Studies (PubMed IDs)
Hepatitis B	CASP1	30222506, 29803661, 24599568
	F5	NA
	PLG	33594307, 31742950, 30024701, 29384847
	MPO	11236502
	JAK2	33435880, 33249195, 32098857, 31485610, 25135878, 25931880, 23483208
	DECR1	NA
	SLC27A5	NA
	APOH	33370716
	DPP4	NA
	BAK1	32574393
Chagas Disease	VEGFA	NA
	IL4	20382097
	APOE	NA
	ICAM1	NA
	STAT3	34061845, 32892338, 31214200, 24260222, 23435997,23253440, 21984337
	IL1A	NA
	CSF2	NA
	CD14	33146244, 31379862, 27902980, 27812661, 15223603
	CXCR4	33195200, 28322302, 23597573, 16637021, 14722885
	NFKB1	NA
Diabetic Nephropathy	F2	NA
	HAMP	NA
	GPT	33232923, 16874670
	CASP1	26832955
	CSF2	31277135
	CBS	33864412
	CTNND1	21752957
	CHDH	NA
	CD14	29531593, 28058051, 25314649, 22772413, 21847775, 9498652
	MMP14	NA

https://doi.org/10.1371/journal.pone.0280839.t004

organs and disease mechanisms involved; for organs, (i) Hepatitis B involves the liver; (ii) Chagas Disease involves multiple systems including cardiovascular and digestive; and (iii) Diabetic Nephropathy involves the kidney.

Hepatitis B and Diabetic Nephropathy have literature support for 6 of the Top 10 predicted genes, and Chagas Disease is supported for 4 of its Top 10 predicted genes. All PubMed studies listed in Table 4 were manually vetted to confirm their relevance to associate each disease with its top predicted genes.

The predicted genes in Table 4 without associated literature evidence are especially interesting because these represent the HetIG-PreDig model's predicted novel gene-disease associations. We manually examined each such predicted gene-disease association, and nearly all of these predictions proved to be reasonable based on the known functions of the genes. For example, for Hepatitis B, predicted gene #2, F5; predicted gene #6, DECR1; predicted gene #7, SLC27A5, and predicted gene #9, DPP4, are all associated with liver function by the GeneCards platform [83].

Chagas Disease, caused by infection by the parasite *Trypanosoma cruzi*, involves immune response and inflammatory lesions, and can cause heart failure, arrhythmias, and dysfunction of the digestive system [84]. Predicted gene #1, VEGFA, is involved with vasculature, and this disease often involves impairment of cardiac and vascular function [84, 85]. Predicted gene #3, APOE, has an established role in cardiac function [86]. Predicted gene #4, ICAM1, has been reported to have both cardiac [87] and digestive [88] roles. Predicted gene #10, NFKB1, has reported functions in the cardiac [89, 90], vascular [91], immune [92], and digestive [93] systems.

For Diabetic Nephropathy, gene F2 was the top predicted gene; although PubMed does not report this association, the GeneCards platform [83] associates abnormality of the kidney with this gene. Similarly, GeneCards also reports an association of kidney function with other top predicted Diabetic Neuropathy genes having no PubMed support, including HAMP (predicted gene #2) and CHDH (predicted gene #8). Predicted gene #10, MMP14, is a member of the matrix metallopeptidase/metalloproteinase family that is well established to have a role in Diabetic Nephropathy.

6 Discussion

This study presents a model for predicting human gene-disease associations using automatic feature learning in a heterogeneous graph with Gene-Disease Prioritization Scores, based on gene co-expression data. The developed HetIG-PreDiG model outperforms baseline models, showing that a model that incorporates this study's novel Gene-Disease Prioritization Score (GDPS) achieves better prediction performance than models without GDPS. Biological data contained within the GDPS, based on gene co-expression data, allow the model to better capture the association between a gene and a disease.

Although the performance of the HetIG-PreDiG model for predicting gene-disease associations is promising, we note the following limitations: Theoretically, it is possible that two or more genes might have identical GDPS scores, which could result in no additional knowledge being added to the model. In addition, when manually analyzing the literature, we found that, with some exceptions, the reported PubMed publications were generally relevant to support the predicted gene-disease associations. However, we identified several challenges to the automation of this analysis, including the ambiguity of some gene names (e.g., some genes with symbols identical to acronyms that refer to an unrelated concept); the occasional reporting of a *lack* of association among genes and diseases; and the gene symbol and disease name appearing within the article's title and/or abstract, but not being associated to each other within the

study. Accordingly, because most of the diseases in Table 3 that are not included in Table 4 were not manually vetted, the Supporting PubMed Papers values in Table 3 should be interpreted with the caveat that the relevant papers represent a subset of the reported counts.

The top predicted genes for each disease warrant further examination, and potential experimental vetting. Of note is that the developed model's predicted gene-disease associations allow both positive and negative relationships; for example, a predicted gene might either cause or increase the probability of developing the disease, or alternatively might prevent or protect against that disease. Future analyses can separate positive from negative associations, and because these effects are often context-dependent, this will require careful analysis that may necessitate the collection of additional experimental data. Future work might also enhance existing gene-disease datasets by considering novel findings not covered within the datasets that were used in this study. Future work might also benefit from adding the co-expression data as another edge layer to the network.

Advancing beyond previous models intended to predict gene-disease associations, this study presents five main contributions. First, biomedical data is incorporated in the form of a Gene-Disease Prioritization Score (GDPS) based on gene co-expression similarity, allowing the model to better evaluate the degree of association between a gene and a disease. Second, network structure is considered by node embeddings using graph representation learning, and this analysis extends beyond network structure by accounting for biological similarity between unconnected nodes using GDPS. Third, in contrast to most existing studies that randomly select non-associating gene and disease nodes that have the potential to be gene-disease associations not yet reported in the literature, we select negative training samples by favoring non-existing gene-disease edges with lower GDPS. Fourth, we show that network data combined with gene co-expression similarity data can effectively predict gene-disease associations compared with baseline models, and demonstrate this via literature analysis. Fifth, we provide a method to evaluate the developed model's success in predicting novel disease genes.

The developed HetIG-PreDiG model can be applied to similar tasks that can be represented using networks and that benefit from incorporating gene expression data. Such tasks might include the prediction of drug-disease associations and the prediction of drug-drug interactions.

7 Conclusion

We have presented the Heterogeneous Integrated Graph for Predicting Disease Genes (HetIG-PreDiG) model that uses gene-gene, gene-disease, and gene-tissue associations data to generate accurate gene-disease association predictions, improving upon existing baseline models. Our model addresses the limitations of previous models for disease-gene prediction. This model has potential utility for other tasks that can be represented as networks and that involve gene expression data, including predicting drug-disease associations and drug-drug interactions.

Acknowledgments

We thank Alexander Lachmann, Ph.D. and Vasileios Stathias, Ph.D. for their insightful suggestions on this manuscript.

Author Contributions

Conceptualization: Kathleen M. Jagodnik, Yael Shvili, Alon Bartal.

Data curation: Kathleen M. Jagodnik, Alon Bartal.

Formal analysis: Kathleen M. Jagodnik, Yael Shvili, Alon Bartal.

Investigation: Kathleen M. Jagodnik, Alon Bartal.

Methodology: Kathleen M. Jagodnik, Alon Bartal.

Project administration: Alon Bartal.

Resources: Kathleen M. Jagodnik, Alon Bartal.

Software: Kathleen M. Jagodnik, Alon Bartal.

Supervision: Alon Bartal.

Validation: Kathleen M. Jagodnik, Yael Shvili, Alon Bartal.

Visualization: Alon Bartal.

Writing – original draft: Kathleen M. Jagodnik, Alon Bartal.

Writing – review & editing: Kathleen M. Jagodnik, Yael Shvili, Alon Bartal.

References

1. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nature reviews Drug discovery. 2018; 17(5):317. https://doi.org/10.1038/nrd.2018.14 PMID: 29472638
2. Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in bioinformatics. 2010; 11(1):96–110. https://doi.org/10.1093/bib/bbp048 PMID: 20007728
3. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature reviews genetics. 2004; 5(2):101–113. https://doi.org/10.1038/nrg1272 PMID: 14735121
4. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018; 173(7):1581–1592. https://doi.org/10.1016/j.cell.2018.05.015 PMID: 29887378
5. Yao V, Kaletsky R, Keyes W, Mor DE, Wong AK, Sohrabi S, et al. An integrative tissue-network approach to identify and test human disease genes. Nature biotechnology. 2018; 36(11):1091–1099. https://doi.org/10.1038/nbt.4246 PMID: 30346941
6. Peng J, Hui W, Shang X. Measuring phenotype-phenotype similarity through the interactome. BMC bioinformatics. 2018; 19(5):114. https://doi.org/10.1186/s12859-018-2102-9 PMID: 29671400
7. Erten S, Bebek G, Koyutürk M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. Journal of computational biology. 2011; 18 (11):1561–1574. https://doi.org/10.1089/cmb.2011.0154 PMID: 22035267
8. Killock D. HotNet2—see the wood for the trees. Nature Reviews Clinical Oncology. 2015; 12(2):66–66. https://doi.org/10.1038/nrclinonc.2014.234 PMID: 25560530
9. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics. 2015; 47(2):106–114. https://doi.org/10.1038/ng.3168 PMID: 25501392
10. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017; 18(9):551. https://doi.org/10.1038/nrg.2017.38 PMID: 28607512
11. Ahmed R, Baali I, Erten C, Hoxha E, Kazan H. MEXCOwalk: mutual exclusion and coverage based random walk to identify cancer modules. Bioinformatics. 2020; 36(3):872–879. https://doi.org/10.1093/bioinformatics/btz655 PMID: 31432076
12. Del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. Current opinion in biotechnology. 2010; 21(4):566–571. https://doi.org/10.1016/j.copbio.2010.07.010 PMID: 20709523
13. Chattopadhyay A, Lu TP. Gene-gene interaction: the curse of dimensionality. Annals of translational medicine. 2019; 7(24). https://doi.org/10.21037/atm.2019.12.87 PMID: 32042829
14. Hu Y, Shmygelska A, Tran D, Eriksson N, Tung JY, Hinds DA. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nature communications. 2016; 7(1):1–9. https://doi.org/10.1038/ncomms10448 PMID: 26835600

15. Safari-Alighiarloo N, Taghizadeh M, Rezaei-Tavirani M, Goliaei B, Peyvandi AA. Protein-protein interaction networks (PPI) and complex diseases. Gastroenterology and Hepatology from bed to bench. 2014; 7(1):17. PMID: 25436094

16. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms–disease network. Nature communications. 2014; 5(1):1–10. https://doi.org/10.1038/ncomms5212 PMID: 24967666

17. Ata SK, Wu M, Fang Y, Ou-Yang L, Kwoh CK, Li XL. Recent advances in network-based methods for disease gene prediction. Briefings in bioinformatics. 2021; 22(4):bbaa303. https://doi.org/10.1093/bib/bbaa303 PMID: 33276376

18. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019; 50:71–91. https://doi.org/10.1016/j.inffus.2018.09.012 PMID: 30467459

19. Cai H, Zheng VW, Chang KCC. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering. 2018; 30(9):1616–1637. https://doi.org/10.1109/TKDE.2018.2807452

20. Xiong Y, Guo M, Ruan L, Kong X, Tang C, Zhu Y, et al. Heterogeneous network embedding enabling accurate disease association predictions. BMC medical genomics. 2019; 12(10):1–17. https://doi.org/10.1186/s12920-019-0623-3 PMID: 31865913

21. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining; 2016. p. 855–864.

22. Cui P, Wang X, Pei J, Zhu W. A survey on network embedding. IEEE Transactions on Knowledge and Data Engineering. 2018; 31(5):833–852. https://doi.org/10.1109/TKDE.2018.2849727

23. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014. p. 701–710.

24. Zeng X, Ding N, Rodríguez-Patón A, Zou Q. Probability-based collaborative filtering model for predicting gene–disease associations. BMC medical genomics. 2017; 10(5):45–53. https://doi.org/10.1186/s12920-017-0313-y PMID: 29297351

25. Alanis-Lobato G, Mier P, Andrade-Navarro M. The latent geometry of the human protein interaction network. Bioinformatics. 2018; 34(16):2826–2834. https://doi.org/10.1093/bioinformatics/bty206 PMID: 29635317

26. Wang B, Pourshafeie A, Zitnik M, Zhu J, Bustamante CD, Batzoglou S, et al. Network enhancement as a general method to denoise weighted biological networks. Nature communications. 2018; 9(1):1–8. https://doi.org/10.1038/s41467-018-05469-x PMID: 30082777

27. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018; 34(13):i457–i466. https://doi.org/10.1093/bioinformatics/bty294 PMID: 29949996

28. Alshahrani M, Hoehndorf R. Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. Bioinformatics. 2018; 34(17):i901–i907. https://doi.org/10.1093/bioinformatics/bty559 PMID: 30423077

29. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015; 31(13):2123–2130. https://doi.org/10.1093/bioinformatics/btv118 PMID: 25717192

30. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nature communications. 2018; 9(1):1–10. https://doi.org/10.1038/s41467-018-03751-6 PMID: 29636450

31. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002; 30(1):207–210. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

32. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research. 2016; p. gkw943. https://doi.org/10.1093/nar/gkw943 PMID: 27924018

33. Thul PJ, Lindskog C. The human protein atlas: a spatial map of the human proteome. Protein Science. 2018; 27(1):233–244. https://doi.org/10.1002/pro.3307 PMID: 28940711

34. Digre A, Lindskog C. The human protein atlas—spatial localization of the human proteome in health and disease. Protein Science. 2021; 30(1):218–233. https://doi.org/10.1002/pro.3987 PMID: 33146890

35. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PloS one. 2011; 6(6). https://doi.org/10.1371/journal.pone.0020284 PMID: 21695124

36. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research. 2017; 45(D1):D896–D901. https://doi.org/10.1093/nar/gkw1133 PMID: 27899670

37. Yoon S, Nguyen HCT, Yoo YJ, Kim J, Baik B, Kim S, et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic acids research. 2018; 46(10):e60–e60. https://doi.org/10.1093/nar/gky175 PMID: 29562348

38. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010; 26(8):1057–1063. https://doi.org/10.1093/bioinformatics/btq076 PMID: 20185403

39. Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, et al. Predicting disease-related genes using integrated biomedical networks. BMC genomics. 2017; 18(1):1–11. https://doi.org/10.1186/s12864-016-3263-4 PMID: 28198675

40. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. 2019; 35(3):497–505. https://doi.org/10.1093/bioinformatics/bty637 PMID: 30020411

41. Cheng L, Wang P, Tian R, Wang S, Guo Q, Luo M, et al. LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic acids research. 2019; 47(D1):D140–D144. https://doi.org/10.1093/nar/gky1051 PMID: 30380072

42. Shim JE, Hwang S, Lee I. Pathway-dependent effectiveness of network algorithms for gene prioritization. PLoS One. 2015; 10(6):e0130589. https://doi.org/10.1371/journal.pone.0130589 PMID: 26091506

43. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics. 2008; 82(4):949–958. https://doi.org/10.1016/j.ajhg.2008.02.013 PMID: 18371930

44. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010; 6(1):e1000641. https://doi.org/10.1371/journal.pcbi.1000641 PMID: 20090828

45. Le DH, Kwon YK. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. Computational biology and chemistry. 2013; 44:1–8. https://doi.org/10.1016/j.compbiolchem.2013.01.001 PMID: 23434623

46. Zhu J, Qin Y, Liu T, Wang J, Zheng X. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. In: BMC bioinformatics. vol. 14. Springer; 2013. p. 1–11.

47. Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010; 26(9):1219–1224. https://doi.org/10.1093/bioinformatics/btq108 PMID: 20215462

48. Yang P, Li X, Wu M, Kwoh CK, Ng SK. Inferring gene-phenotype associations via global protein complex network propagation. PloS one. 2011; 6(7):e21502. https://doi.org/10.1371/journal.pone.0021502 PMID: 21799737

49. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Molecular systems biology. 2008; 4(1):189. https://doi.org/10.1038/msb.2008.27 PMID: 18463613

50. Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS, Marcotte EM. Prediction and validation of gene-disease associations using methods inspired by social network analyses. PloS one. 2013; 8(5):e58977. https://doi.org/10.1371/journal.pone.0058977 PMID: 23650495

51. Xie M, Xu Y, Zhang Y, Hwang T, Kuang R. Network-based phenome-genome association prediction by bi-random walk. PloS one. 2015; 10(5):e0125138. https://doi.org/10.1371/journal.pone.0125138 PMID: 25933025

52. Zeng X, Ding N, Zou Q. Latent factor model with heterogeneous similarity regularization for predicting gene-disease associations. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2016. p. 682–687.

53. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PloS one. 2009; 4(8):e6536. https://doi.org/10.1371/journal.pone.0006536 PMID: 19657382

54. Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic acids research. 2014; 42(W1):W137–W146. https://doi.org/10.1093/nar/gku412 PMID: 24895436

55. Ideker T, Sharan R. Protein networks in disease. Genome research. 2008; 18(4):644–652. https://doi.org/10.1101/gr.071852.107 PMID: 18381899

56. Mordelet F, Vert JP. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC bioinformatics. 2011; 12(1):1–15. https://doi.org/10.1186/1471-2105-12-389 PMID: 21977986

57. Nitsch D, Gonçalves JP, Ojeda F, De Moor B, Moreau Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC bioinformatics. 2010; 11(1):1–16. https://doi.org/10.1186/1471-2105-11-460 PMID: 20840752

58. Chen Y, Wu X, Jiang R. Integrating human omics data to prioritize candidate genes. BMC medical genomics. 2013; 6(1):1–12.

59. Chen B, Wang J, Li M, Wu FX. Identifying disease genes by integrating multiple data sources. BMC medical genomics. 2014; 7(2):1–12. https://doi.org/10.1186/1755-8794-7-S2-S2 PMID: 25350511

60. Zhou H, Skolnick J. A knowledge-based approach for predicting gene–disease associations. Bioinformatics. 2016; 32(18):2831–2838. https://doi.org/10.1093/bioinformatics/btw358 PMID: 27283949

61. Ata SK, Fang Y, Wu M, Li XL, Xiao X. Disease gene classification with metagraph representations. Methods. 2017; 131:83–92. https://doi.org/10.1016/j.ymeth.2017.06.036

62. Luo P, Li Y, Tian LP, Wu FX. Enhancing the prediction of disease–gene associations with multimodal deep learning. Bioinformatics. 2019; 35(19):3735–3742. https://doi.org/10.1093/bioinformatics/btz155 PMID: 30825303

63. Tran VD, Sperduti A, Backofen R, Costa F. Heterogeneous networks integration for disease–gene prioritization with node kernels. Bioinformatics. 2020; 36(9):2649–2656. https://doi.org/10.1093/bioinformatics/btaa008 PMID: 31990289

64. Kim P, Park A, Han G, Sun H, Jia P, Zhao Z. TissGDB: tissue-specific gene database in cancer. Nucleic acids research. 2018; 46(D1):D1031–D1038. https://doi.org/10.1093/nar/gkx850 PMID: 29036590

65. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nature genetics. 2015; 47(6):569. https://doi.org/10.1038/ng.3259 PMID: 25915600

66. Luo P, Tian LP, Chen B, Xiao Q, Wu FX. Predicting gene-disease associations with manifold learning. In: International Symposium on Bioinformatics Research and Applications. Springer; 2018. p. 265–271.

67. Zitnik M, Zupan B. Jumping across biomedical contexts using compressive data fusion. Bioinformatics. 2016; 32(12):i90–i100. https://doi.org/10.1093/bioinformatics/btw247 PMID: 27307649

68. Zakeri P, Simm J, Arany A, ElShal S, Moreau Y. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics. 2018; 34(13):i447–i456. https://doi.org/10.1093/bioinformatics/bty289 PMID: 29949967

69. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;.

70. Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics. 2017; 33(15):2337–2344. https://doi.org/10.1093/bioinformatics/btx160 PMID: 28430977

71. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018; 34(4):660–668. https://doi.org/10.1093/bioinformatics/btx624 PMID: 29028931

72. Fu Ty, Lee WC, Lei Z. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017. p. 1797–1806.

73. Hu B, Fang Y, Shi C. Adversarial learning on heterogeneous information networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019. p. 120–129.

74. Liu R, Mancuso CA, Yannakopoulos A, Johnson KA, Krishnan A. Supervised learning is an accurate method for network-based gene classification. Bioinformatics. 2020; 36(11):3457–3465. https://doi.org/10.1093/bioinformatics/btaa150 PMID: 32129827

75. Ata SK, Ou-Yang L, Fang Y, Kwoh CK, Wu M, Li XL. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. BMC systems biology. 2018; 12(9):31–44. https://doi.org/10.1186/s12918-018-0662-y PMID: 30598097

76. Zeng M, Zou B, Wei F, Liu X, Wang L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In: 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS). IEEE; 2016. p. 225–228.

77. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16:321–357. https://doi.org/10.1613/jair.953

78. Zeng M, Li M, Wu FX, Li Y, Pan Y. DeepEP: a deep learning framework for identifying essential proteins. BMC bioinformatics. 2019; 20(16):1–10. https://doi.org/10.1186/s12859-019-3076-y PMID: 31787076

79. Ellerkmann RK, Bruhn J, Soehle M, Kehrer M, Hoeft A, Kreuer S. Maximizing prediction probability PK as an alternative semiparametric approach to estimate the plasma effect-site equilibration rate constant

ke0. Anesthesia & Analgesia. 2009; 109(5):1470–1478. https://doi.org/10.1213/ANE.0b013e3181b61efd PMID: 19713250

80. Jordan D, Steiner M, Kochs EF, Schneider G. A program for computing the prediction probability and the related receiver operating characteristic graph. Anesthesia & Analgesia. 2010; 111(6):1416–1421. https://doi.org/10.1213/ANE.0b013e3181fb919e PMID: 21059744

81. Wang H, Song M. Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. The R journal. 2011; 3(2):29. https://doi.org/10.32614/RJ-2011-015 PMID: 27942416

82. Vargha A, Delaney HD. The Kruskal-Wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics. 1998; 23(2):170–192. https://doi.org/10.3102/10769986023002170

83. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Current protocols in bioinformatics. 2016; 54(1):1–30. https://doi.org/10.1002/cpbi.5 PMID: 27322403

84. Coura JR, Borges-Pereira J. Chagas disease: 100 years after its discovery. A systemic review. Acta tropica. 2010; 115(1-2):5–13. https://doi.org/10.1016/j.actatropica.2010.03.008 PMID: 20382097

85. Monterroso J. Chagas disease: A review. Journal of Alternative Medicine Research. 2021; 13(2):117–125.

86. Ellulu MS, Patimah I, Khaza'ai H, Rahmat A, Abed Y, Ali F. Atherosclerotic cardiovascular disease: a review of initiators and protective factors. Inflammopharmacology. 2016; 24(1):1–10. https://doi.org/10.1007/s10787-015-0255-y PMID: 26750181

87. Lin QY, Lang PP, Zhang YL, Yang XL, Xia YL, Bai J, et al. Pharmacological blockage of ICAM-1 improves angiotensin II-induced cardiac remodeling by inhibiting adhesion of LFA-1+ monocytes. American Journal of Physiology-Heart and Circulatory Physiology. 2019; 317(6):H1301–H1311. https://doi.org/10.1152/ajpheart.00566.2019 PMID: 31729904

88. Sumagin R, Brazil J, Nava P, Nishio H, Alam A, Luissint A, et al. Neutrophil interactions with epithelial-expressed ICAM-1 enhances intestinal mucosal wound healing. Mucosal immunology. 2016; 9(5):1151–1162. https://doi.org/10.1038/mi.2015.135 PMID: 26732677

89. Coto E, Reguero JR, Avanzas P, Pascual I, Martín M, Hevia S, et al. Gene variants in the NF-KB pathway (NFKB1, NFKBIA, NFKBIZ) and risk for early-onset coronary artery disease. Immunology letters. 2019; 208:39–43. https://doi.org/10.1016/j.imlet.2019.02.007 PMID: 30902734

90. Jin SY, Luo JY, Li XM, Liu F, Ma YT, Gao XM, et al. NFKB1 gene rs28362491 polymorphism is associated with the susceptibility of acute coronary syndrome. Bioscience reports. 2019; 39(4). https://doi.org/10.1042/BSR20182292 PMID: 30910844

91. Yenmis G, Oner T, Cam C, Koc A, Kucuk O, Yakicier M, et al. Association of NFKB 1 and NFKBIA Polymorphisms in Relation to Susceptibility of B ehçet's Disease. Scandinavian journal of immunology. 2015; 81(1):81–86. https://doi.org/10.1111/sji.12251 PMID: 25367031

92. Kaustio M, Haapaniemi E, Göös H, Hautala T, Park G, Syrjänen J, et al. Damaging heterozygous mutations in NFKB1 lead to diverse immunologic phenotypes. Journal of Allergy and Clinical Immunology. 2017; 140(3):782–796. https://doi.org/10.1016/j.jaci.2016.10.054 PMID: 28115215

93. Borm M, Van Bodegraven A, Mulder C, Kraal G, Bouma G. A NFKB1 promoter polymorphism is involved in susceptibility to ulcerative colitis. International journal of immunogenetics. 2005; 32(6):401–405. https://doi.org/10.1111/j.1744-313X.2005.00546.x PMID: 16313306