Funded by the European Union

Work Package 4

Project Name:	Disc4All 955735		
Date:	11/11/2024	**Release:**	Final
WP Leader (WPL):	UPF		
WP Co-leader (WPCL):	PAO		
Reviewer 1	IMIM		
Reviewer 2	Jérôme Noailly		
Document Number:	D4.5		

Revision History

Date of next revision:

Revision Date	Previous Revision Date	Summary of Changes	Changes Marked
07/10/2024	NA	First draft	YES
15/10/2024	07/10/2024	V1	YES
11/11/2024	07/10/2024	Final with demonstrator video included: https://youtu.be/C3VBF2e2O6Y?si=5Anc5-647HhWpiII	YES

Approvals

This document requires the following approvals. A signed copy should be placed in the project files.

Name	YES/NO	Title	Date of Issue	Version
J. Pinero	YES	PI	15/10/2024	V1

Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version
Working Package 4 leaders, project coordinator	D4.3	07/10/2024	Draft
Consortium, Coordinator	D4.3	11/11/2024	Final

Work Package Authorisation	
Title	Project coordinator
Person Authorised[1]	Jérôme Noailly
Date[2]	11/11/2024

Approval method

Initial internal approval by the WP leaders at UPF and IMIM, followed by project-level approval by the project coordinator. Finally, it will be shared with the other consortium members via email and the project Microsoft Teams folder.

Work Package Acceptance	
Person Accepting[3]	Jérôme Noailly
Date[4]	11/11/2024

[1] The name of the WPL

[2] The date of the agreement between the Coordinator and the WPL/person authorised

[3] The Coordinator or other person accepting the work package on the Coordinator's behalf

[4] The date of acceptance

Assessment and feedback	Modification of the references and demo video.

D 4.5 Phenotype-gene-biomarker-comorbidities association tool

Lead Beneficary: Instituto de investigaciones médicas Hospital del Mar (IMIM)

Lead Author: Francesco Gualdi

Due date: November 2024

Contents

Introduction

Intervertebral disc degeneration

Intervertebral disc degeneration is a multifactorial condition that significantly impacts the health of the spine. The intervertebral disc, composed of a gel-like nucleus pulposus (NP) surrounded by a tough annulus fibrosus (AF), play a crucial role in providing flexibility, shock absorption, and stability to the spinal column. However, with aging and various contributing factors, the disc undergoes degenerative changes that compromise its biomechanical properties. These changes include alterations in the composition and structure of the extracellular matrix, such as a decrease in proteoglycan content and disorganization of collagen fibers, leading to reduced hydration, diminished disc height, and osteophytes formation. The progressive loss of disc integrity can lead to the development of pathological conditions like disc herniation, spinal stenosis, and facet joint osteoarthritis, leading to pain and impaired spinal function [1].

Several factors contribute to IDD, including genetic predisposition, biomechanical loading, lifestyle factors, and environmental influences [2]. Genetic predisposition plays a significant role in determining an individual's susceptibility to disc degeneration, with certain gene polymorphisms associated with an increased risk of developing degenerative disc disease. Additionally, repetitive mechanical loading and trauma, as well as poor posture and sedentary lifestyle habits, can accelerate disc degeneration by inducing microstructural damage and promoting inflammatory responses within the disc tissue. Furthermore, lifestyle factors such as obesity and smoking can impair disc metabolism, further incentivizing degenerative changes [3].

The clinical manifestations of intervertebral disc degeneration are very broad and can range from asymptomatic to debilitating, depending on the severity and location of the degenerative changes. Common symptoms include chronic low back pain (LBP), most of the time due to spinal stenosis. In fact, the loss of function of the disc leads to a degenerative process affecting the surrounding anatomical areas such as joints, muscles and ligaments resulting in the narrowing of the spinal canal and compression of the nerve tissue [4].

The current approaches for the treatment of IDD comprise conservative, interventional, and surgical approaches. Conservative treatment focuses on alleviating LBP and improving quality of life through methods like physical therapy, medication, and lifestyle adjustments. In cases of severe symptoms interventional treatments involve procedures such as Intra-Discal Electrothermal Therapy (IDET), radiofrequency myeloplasty, and ozone therapy to modify disc mechanics and manage pain. Surgical options, including intervertebral disc fusion, aim to provide pain relief and functional improvement by removing damaged discs, inserting support cages, and fixing vertebrae with pedicle screws. The choice of treatment depends on the severity of symptoms and individual patient factors, with the goal of addressing IDD symptoms effectively and improving patient outcomes [5].

In recent years, advancements in regenerative medicine, tissue engineering, and biological therapies have provided promising avenues for the treatment of intervertebral disc degeneration. Strategies such as mesenchymal stem cell therapy, tissue engineering and gene therapy aim to enhance matrix synthesis, treat disc injuries and inhibit inflammatory processes within the degenerated discs. Despite these innovative approaches holding great promise for revolutionizing the treatment of degenerative disc disease, treatment of IDD remains a great challenge [6].

The treatments of IDD have been supported from a deeper insight into the biological landscape of this complex condition. In fact, current technologies have revolutionized our understanding of complex diseases by allowing the study of biological molecules at large scale.

System biology for the investigation of complex diseases

Complex diseases result from a mix of genetic, environmental, and lifestyle factors, unlike single-gene disorders. Understanding their genetic basis is challenging due to various factors like multiple genes, gene-gene and gene-environment interactions, and variable expression [7].

Omics technologies (like genomics, proteomics, metabolomics, and transcriptomics) offer a comprehensive view of biological systems by studying genes, proteins, metabolites, and gene expression. Integrating data from these disciplines helps in understanding disease mechanisms, finding new drug targets, and advancing personalized medicine and other fields for better health outcomes [8].

The integration of omics data can help to understand the etiology and origins of complex conditions, offering a broader perspective on the biological pathways implicated in their progression [9]. Network-based approaches represent one strategy for integrating diverse biological data and depicting complex biological systems [10]. Here, interplay between biological entities can be represented as biological networks where different entities interact through different types of relationships. In the biomedical field many different networks can be constructed each of one having different types of nodes and edges [11].

Originally, early attempts at modeling complex interactions in biological systems used simple networks, which are essentially uni-relational graphs [12]. Despite their initial success, these networks couldn't capture the nuanced meanings of different types of connections between entities. For example, when representing protein-protein interaction networks using these basic networks, it wasn't possible to distinguish between various interaction types like inhibition, activation, phosphorylation, etc. As a result, recent research has shifted towards using heterogeneous multi-relational networks known as knowledge graphs [13].

Knowledge graphs

Knowledge Graphs (KGs) are increasingly implemented in the biomedical field due to their potential for representing and analyzing complex biomedical data. Recent research has highlighted their importance in enabling intelligent applications such as recommendation systems, semantic search, and logical reasoning. Automated schemes have been shown to significantly reduce the cost of building knowledge graphs [14]. Current research is addressing challenges such as knowledge graph completion and extraction methods for unstructured data. There is also a growing emphasis on constructing KGs from natural language text, with a focus on named entity recognition and relation extraction [15]. The field is still facing technical challenges, but the ongoing research aims to enhance the quality and reliability of knowledge graphs through novel techniques, models, and frameworks.

One commonly used approach to infer new interactions between biological entities involves expressing the entities within KGs as low-dimensional vectors using vectorial representations that preserve the graph’s local structure known as knowledge graph embeddings (KGE). This method outperforms other approaches in terms of prediction accuracy and scalability [13].

Numerous methods have been developed to generate embeddings from KGs, and they can be broadly categorized into five main families: translational models, matrix factorization, semantic matching, random walks-based models, and deep neural networks. Refer to [16], [17] for a comprehensive overview of these methods. Recently, new techniques that combine these existing methods have emerged [18]. For example, translational methods or PageRank [19] are merged with graph attention networks (GAT) to improve predictive powers of the embeddings [20]. Several studies have been conducted to explore the potential of KGE for predicting gene-disease associations (GDAs). For instance, *Nunes et al* investigated the impact of employing rich semantic representations based on more than one ontology to predict GDAs by testing different embedding creation models and machine learning algorithms [21]. Other works have focused on the heterogeneous integration of knowledge bases with the development of a single deep learning framework for predicting GDAs starting from a KG [22], [23].

In the biomedical domain, KGE have been implemented for a wide range of downstream machine learning tasks, such as drug – target prediction [24], protein - protein interaction prediction [25] and therapeutic indications [26]. Also, KGE has demonstrated the ability to achieve prediction capabilities similar to those of raw data, while also offering the advantage of reduced dimensionality compared to the original dataset. [27].While previous studies have made progress in implementing KGE methods in GDA research, we lack a proper benchmark of available methods. Existing works in this field are limited to evaluating the proposed method [22] or the comparison of different algorithms [21] without providing a deeper insight into the generated embeddings or validating a particular use case. In this work we conducted a comparison of different methods of KGE creation with unsupervised and supervised machine learning tasks. We first generated KGE from multiple ontologies and biological knowledge bases, and we implemented four state-of-the-art methods, and two novel algorithms. Subsequently, we analyzed the generated embeddings using unsupervised clustering algorithms. Furthermore, we evaluated the performance of the embeddings in a GDAs prediction task. Finally, we used the best performing model to predict potential genes associated to intervertebral disc degeneration (IDD).

Methods

Data sources

To build the KG, we mined different types of biological data from publicly available repositories: *Protein - protein interactions:* We partially integrated data from multiscale interactome (downloaded 29/06/2022) [26]. Specifically, the data were integrated from:

The biological general repository for interaction dataset (BioGRID) [28]. This is a repository of manually curated both physical and genetic interactions between proteins from 71,713 high - throughput and low - throughput publications.

The database of interacting proteins (DIP)[29] in which only physical protein - protein interactions are reported with experimental and curated evidence.

Four protein-protein interaction networks from the human reference protein interactome mapping project [30]): (HI-I-05: 2,611 interactions between 1,522 proteins; HI-II-14 13,426 interactions between 4,228 proteins, Venkatesan-09: 233 interactions between 229 proteins; Yu-11 1,126 interactions between 1,126 proteins). In addition, we integrated the last version of the Human reference interactome (HI-III-20) [30].

Physical protein-protein interaction from Menche et al. [12]). This repository integrates different resources of physical protein - protein interaction data from experimental evidence. It integrates regulatory interactions from TRANSFAC [31] database, binary interactions from yeast-two-hybrid datasets and curated interactions from IntAct [32], BioGRID and HPRD [33]. It integrates also metabolic-enzyme interactions from KEGG [34] and BIGG [35], protein complex interactions from CORUM [36], kinase-substrate interactions from PhosphositePlus [37] and signalling interactions from Vinayagam et al. [38]

Only human proteins for which direct experimental evidence of a physical interaction existed, were considered.

Ontologies: Ontologies are computational structures that aim to describe and classify the entities belonging to a certain domain in a structured and machine-readable format, in order to be implemented in a broad range of applications. The main components of the ontology are classes that represent specific entities and usually are associated with an identifier. These classes are arranged in a hierarchical way from general to more specific and are connected to each other through relations. Finally, ontologies feature metadata, formats and axioms [39] For our purpose, we integrated the following types of ontologies:

Gene Ontology (GO) [40], (downloaded 18/07/2022) is a knowledge base that aims to computationally describe biological systems ranging from molecules to organisms, as of 2023 it comprises 43,248 terms, 7,503,460 annotations across 5,267 species.

Disease Ontology (DO) [41], (downloaded 02/08/2022) is an ontological structure of standardized disease descriptors across multiple resources. The aim of the project is to provide a computable structure of integrated biomedical data in order to improve the knowledge on human diseases.

Human Phenotype Ontology (HPO) [42], (downloaded 22/08/2022) is a comprehensive logical structure that describes phenotypic abnormalities found in human diseases. This enables computational inference and interoperability in digital medicine.

We integrated HPO and DO and mapped the common codes to UMLS CUIS [43]

Gene product annotations to biological processes Proteins in the KG were mapped to their specific biological process through GO. GO annotations are statements about the function of a particular gene product, in this way, it is possible to obtain a snapshot of the current biological knowledge. We included gene annotations from the gene ontology association file (downloaded 29/06/2022).

Gene products annotations to phenotypes We integrated data of genes associated to phenotypes from 2 sources:

DisGeNET [44] is one of the largest publicly available collections of genes and variants associated with human diseases, it integrates GDAs data from curated resources with data automatically mined from the scientific literature using text-mining approaches. For our purposes we exploited DisGeNET curated (version 7.0) that integrates expert curated human gene disease associations from different data sources.To create a dataset, we used curated data from DisGeNET, comprising a total of 84,037 associations (hereafter considered as positives). We generated the same number of gene-disease non-associations (i.e. negatives) by considering that such associations were not reported in the text – mining version of DisGeNET, hence taking randomly any gene-disease pair not reported as positive.

HPO gene annotations to phenotypes: HPO (downloaded 02/08/2022) provides a file that links between genes and HPO terms. If variants in a specific gene are associated with a disease, then all the phenotypes related to that specific disease are assigned to that gene.

Phenotypes annotated to diseases We integrated annotations of phenotypes to disease from the phenotype.hpoa file from HPO ontology (downloaded 15/12/2022).

Drug-disease associations We integrated data of drug-disease pairs from the multiscale interactome [26]. This dataset is integrated by a collection of FDA approved treatments for diseases including different sources:

The drug repurposing database [45] is a database of gold-standard drug-disease pairs extracted from DrugCentral [46] and ClinicalTrials.gov

The drug repurposing hub [47] is a collection of drug-disease including 4,707 compounds. The database contains information mined from publicly and proprietary datasets that undergo manual curation.

The drug indication database [48] integrates data from 12 openly available, commercially available and proprietary information sources.

The dataset was filtered by keeping only human proteins resulting in a total number of drug - disease pairs of 5,926.

Drug - target interaction We obtained a dataset of drugs and their mode of actions on target proteins by integrating DrugBank [49] and the drug repurposing hub. Proteins that were not included in the protein – protein interaction network were removed.

KGE generation algorithms

We tested four state-of-the-art algorithms based on different principles and we implemented two novel methods to generate embeddings, referred to as BioKG2vec and Dlemb. For all experiments, the embeddings vector dimension was 100 and we set the number of epochs to 15.

RotatE RotatE [50] is a KGE generation algorithm that maps relations and entities to the complex vector space. The relations are considered as rotations from the source entity to the target entity. The principle lays on the assumption that given the triple "(h,r,t)", where h is the head, r is the relation, t is the tail e.g. "(protein1, interacts with, protein2)"the embeddings are obtained by the relation t = h ◦ r where ◦ denotes the Hadamard operation between the h and r vectors.

Relational graph convolutional networks (R-GCN) [51] R-GCN is an architecture for calculating the forward pass of relational graphs with multiple edge types. The propagation model is calculated as follows:

$$h_i^{(h+1)} = \sigma\left(\sum_{r\epsilon R}\sum_{j\epsilon N_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$

Where N_i^r is the set of neighbours of node *i* under the relation $r \in R$ and $c_{i,r}$ is a problem-specific normalization constant that is chosen beforehand. $h_j^{(l)}$ is the node vector of neighbour *j* on which applies weight matrix $W_r^{(l)}$ of relation r in the iteration *l*. $W_0^{(l)}h_i^{(l)}$ is the representation of node *i* at layer *l* i.e. a self-representation at antecedent iteration.

TransE TransE [52] is an algorithm that relies on a translational - based model. It represents relationships as translations in the embedding space. The principle lays on the assumption that given the triple *"(h,r,t)"*, where h is the head, r is the relation, t is the tail e.g. *"(protein1, interacts with, protein2)"*, the embedding of the tail should be similar to the head embedding plus the relationship embedding.

Metapath2Vec [53] is an extension of the Node2Vec model [54] well suited for heterogeneous networks. The algorithm relies on meta-path-based random walks that capture both semantic and structural correlations between different types of nodes.

BioKG2vec BioKG2vec relies on a biased random-walk approach in which the user can prioritize specific connections by assigning a weight to edges. In the KG defined in this work we used 4 different node-types: drug, protein, function and disease. Then, the probability of visiting a specific neighbour at every step is given by the equation:

$$P(n_i) = \frac{\left(n_i\left(1+\frac{w_i}{n_i}\right)\right)}{W}$$

where $P(n_i)$ is the probability for the random walker to visit a specific node type, n_i is the number of paths leading to the node (of the same type), w_i is the assigned weight (also specific for the type) and W equals to the node degree plus the sum of all weights (i.e. $\sum_i w_i$). To detect the optimal weights for the prediction of GDAs we performed a grid search assigning weights prioritizing drug -> protein -> function -> disease. Moreover, the walker stores the information of the visited edge type, and this information is used as input for Word2Vec algorithm in the embedding generation

step. Thus, the algorithm handles different edges and nodes behaving differently for each node type being visited and storing the edge type of information too.

BioKG2vec is available at https://zenodo.org/badge/latestdoi/624339823.

Dlemb Dlemb is a shallow neural network (NN) that consists of 3 layers: the input layer, embedding layer and output layer. The input layer takes as input KG entities as numbers and outputs them to the embedding layer. In the dot layer the scalar product of the vector is computed and normalized so the result is a number that ranges between -1 and 1. A false relation yields -1 while true relations produce +1. Then, the RMSE is calculated between the dot product and the expected value. Finally, the ADAM optimizer is used to adjust the embeddings layer directly since these are parameters of the neural network so that the model can be fitted to the data.

Dlemb is available at https://zenodo.org/badge/latestdoi/635382680.

Methods to combine embeddings

We used 4 strategies to combine gene and disease embeddings to obtain GDAs representations: 1) Sum, which consisted of the addition of both vectors; 2) Average, in which we averaged them; 3) concatenation, in which the result is a vector in a larger dimension, representing a pair gene-disease by concatenating both vectors; 4) Hadamard product (i.e. each element is produced by the product of the elements of the two vectors). For this work we produced embeddings of fixed dimension (i.e. 100) in the space of reals (i.e.$\mathbb{R}^{100}$).

Unsupervised analysis of the embeddings

We assessed the quality of the embeddings performing k-means unsupervised clustering. Specifically, we used function and compartment-based classification to group gene products in 16 different categories from human protein atlas (HPA) [55]. For diseases, we used annotations from UMLS to ICD-9 [56], that classify diseases into macro classes. We then used various evaluation scores for the comparison, such as the silhouette score, defined as:

$$\frac{b - a}{max(a, b)}$$

where b is the mean distance between a sample and all other points in the nearest cluster (nearest – cluster distance) and a is the mean distance between a sample and all other points in the same class (inter – cluster distance). We calculated this score for different cluster sizes ranging from 10 to 20 for genes (the gold standard number of clusters is 16) and from 10 to 20 for diseases (the gold standard number of clusters is 16).

Finally, we evaluate the homogeneity score, defined as:

$$1 - \frac{H(Y_{true}|Y_{pred})}{H(Y_{true})}$$

That is a measure that quantifies the similarity of samples in each cluster. Where the Y_{true} is the number of classes, Y_{pred} is the number of clusters and $H(Y_{true}|Y_{pred})$ represents the ratio between the number of classes Y_{true} in cluster Y_{pred} and the total number of samples in cluster Y_{pred}. When all the entities in the cluster belong to a class the homogeneity score equals 1.

Then, for visualization purposes, we performed UMAP dimensionality reduction on the embeddings and plotted the first 2 UMAP embeddings of gene and disease embeddings. Only 3 classes of genes and diseases are plotted.

Grid Search to select the best predictive model.

We performed a grid search cross-validation to find the best combination of embedding creation algorithm, GDAs representation and predictive machine learning (ML) and deep learning (DL) algorithms implemented in Scikit-learn [57] and Pytorch [58] respectively. In the grid-search experiment we created a KG in which we integrated all the biological data and 80% of curated GDAs from DisGeNET. We tested the predictions in the remaining 20% of GDAs that weren't used in the embeddings creation step. To avoid data leakage, we excluded diseases with over 20 associated genes, of which more than 90% were shared with another disease. Additionally, we made sure that in the validation dataset there were no GDAs included in the HPO data. For each algorithm, we fitted a grid of parameters (Table 1) maximizing the area under the receiver operating-characteristic curve (ROCAUC). With this, we tested a total of 120 combinations for the grid search (Supplementary Table 1). Then, the best parameter combination was evaluated on the test set by assessing additional evaluation metrics, such as:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$$FPR = \frac{FP}{FP + TN}$$

We also report the area under the precision recall curve (AUPRC).

Table 1: Search spaces of the algorithms tested during the grid search cross validation.

ALGORITHM	**PARAMETERS**	**VALUES**
LR	C	0.001, 0.01, 1, 5, 10, 25
	PENALTY	L1, L2
RANDOM FOREST	MAX DEPTH	2, 4, 6, None
	N. OF ESTIMATORS	20, 50, 100
XGBOOST	COLSAMPLE BY TREE	0.3, 0.7
	GAMMA	0, 0.5
	LEARNING RATE	0.03, 0.3
	MAX DEPTH	2, 6
	N. OF ESTIMATORS	100, 150
	SUBSAMPLE	0.4, 0.6
SVM	C	0.1, 1, 10
	GAMMA	0.001, 0.01, 0.1
	KERNEL	rbf, poly
FFN	N. OF LAYERS	2, 3
	N. OF NODES FIRST LAYER	50, 100, 150
	N. OF NODES SECOND LAYER	20, 50
	ACTIVATION FUNCTION	sigmoid, tanh, relu
	LOSS FUNCTION	Binary cross-entropy, hinge
	BATCH SIZE	30, 100
	EPOCHS	20, 60

Ontology preprocessing and heterogeneous data integration

Once we selected the model with the highest predictive power, we investigated the influence of integrating heterogeneous biological data in the KG on the GDAs predictions. For this experiment we only used ontological data. Ontologies are complex, standardized data structures composed of classes, relations, axioms and metadata all of which are included in the raw ontology. Moreover, we tested the effect of implementing a pre-processing step in the ontology in which only classes and relations were maintained as a graph structure (axioms and metadata were excluded). We studied the following combinations of data sources:

HPO + HPO annotations raw

HPO + HPO annotations pre-processed

HPO + HPO annotations + GO + GO annotations (all) pre-processed

We used a comparison based on two metrics. For this experiment, we created embeddings with the Mtapath2vec algorithm, using concatenation for GDAs representation, and SVM as classification algorithm. For the processing of the ontologies nxontology and pronto [59] python libraries were used.

Influence of GDAs in the KG for GDA-predictions.

We tested the influence of adding increasing GDAs proportions in the KG. For this experiment, we used 20% 50% 80% and 100% of DisGeNET and we included it in the KG. Then we generated embeddings from the KGs with Metapath2Vec and we trained a SVM on 80% of DisGeNET. We tested the model on the 20% of remaining associations and calculated ROCAUC and AUPRC as evaluation metrics.

Comparison with randomly generated embeddings

To show that the information is efficiently translated from the KG to the vectorial space, we compared the performance of Metapath2Vec generated embeddings and random embeddings of the same size. We aimed to assess the effectiveness of translating the information encoded in the KG into embeddings by comparing KGE with a null model. To conduct this evaluation, we created 100-dimensional random embeddings for each gene and disease, represented GDAs through concatenation, and tested their predictive capabilities. The number of associations is a latent variable that can be learned by ML to produce good predictions. This can be considered a potential bias. Therefore, we further tested the effect of removing the number of associations stratifying DisGeNET diseases by the number of associated genes. We divided the data into 23 groups in which the number of associations for every disease has a maximum difference of 20. Then we selected a disease belonging to every class, generated negative associations and performed a five-fold cross validation on the data with the best performing algorithm. We evaluated accuracy, precision, recall, f1 score and ROCAUC across every fold.

Generalizability of the model.

The predictive model selected was tested to predict associations for diseases not used in the training set. The rationale behind this experiment was to understand the capabilities of the model to predict gene-disease associations of new diseases, proving that the biological information encoded in the embeddings was generalizable.

To assess this, we trained the model on GDAs belonging to diseases of a specific ICD-9 disease class and then we tested the model on all other classes.

Performance of the algorithms

We compared the performance of the algorithm with the top predictive power i.e. Metapath2Vec, BioKG2vec and Dlemb. We performed n = 10 experiments by randomly selecting 1000 nodes from the knowledge graph, creating the subnetwork and producing the embeddings. We calculated the difference of the running time (in seconds) as percentage with the following formula:

$$\frac{T_1 - T_2}{T_1} \times 100$$

Being T1 the running time of Metapath2Vec and T2 the running time of either BioKG2vec or Dlemb. The experiment was conducted on an 8-core intel i7 machine.The experiment was conducted on an 8-core intel i7 machine.

Intervertebral Disc Degeneration Biomarker Prediction

We tested the model to predict genes associated with IDD. We used the model selected through grid search cross validation with concatenation of the embeddings for the GDAs representation. Lastly, we performed a function enrichment analysis using g:Profiler [60] on the set of prioritized genes with a probability greater than 0.95 to be associated to IDD

Results

Data integration and KG structure

We integrated multiple sources of data in the form of KG for a total of 95952 nodes and 2,183,603 edges. The KG contains 4 types of nodes: drugs (n = 2,991), phenotypes (n = 28,374), proteins (n =21,019) and functions (n = 43,568). These entities are connected by 81 different types of relationships represented as edges. The relationships are obtained through different data sources, 18,282 proteins interacting among each other (87, 1356 edges), 19,409 proteins annotated to 18,813 biological functions (303,404 edges) and 8, 053 proteins annotated to 13,525 phenotypes (246,006 edges). Moreover, drug information was included: 1,551 drugs annotated to 828 phenotypes for a total of 5,744 edges and 2,887 connected to 2,074 proteins they target for a total of 14,491 edges. The degree distribution of the graph follows a scale free law (Supplementary Figure 1) [61].

Unsupervised clustering of the embeddings reflects the biological classification

From the KG, we generated embeddings using six algorithms. Figure 1 shows the first 2 UMAP embeddings of genes and diseases. The embeddings tend to differentiate among gene products belonging to different groups: secreted, transcription factors, and transporters (Figure 1 A to G). Metapath2Vec, BioKG2vec, and Dlemb from a visual perspective achieve the best clustering of genes. In Figure 1, G to L only 3 categories of diseases are represented, corresponding to the ICD chapters disease of blood and blood-forming organs, diseases of the musculoskeletal system and connective tissue and mental disorders. As above, algorithms Metapath2Vec, BioKG2vec and Dlemb visually distinguished disease classes better than others.

The algorithm producing the best clustering of disease classes and gene products was Metapath2Vec, which has a higher homogeneity score for both genes and diseases. (Table 2). For the case of diseases, the silhouette score of the embeddings produced with any algorithm couldn't match the gold standard number of clusters (Supplementary Figures 2 and 3).

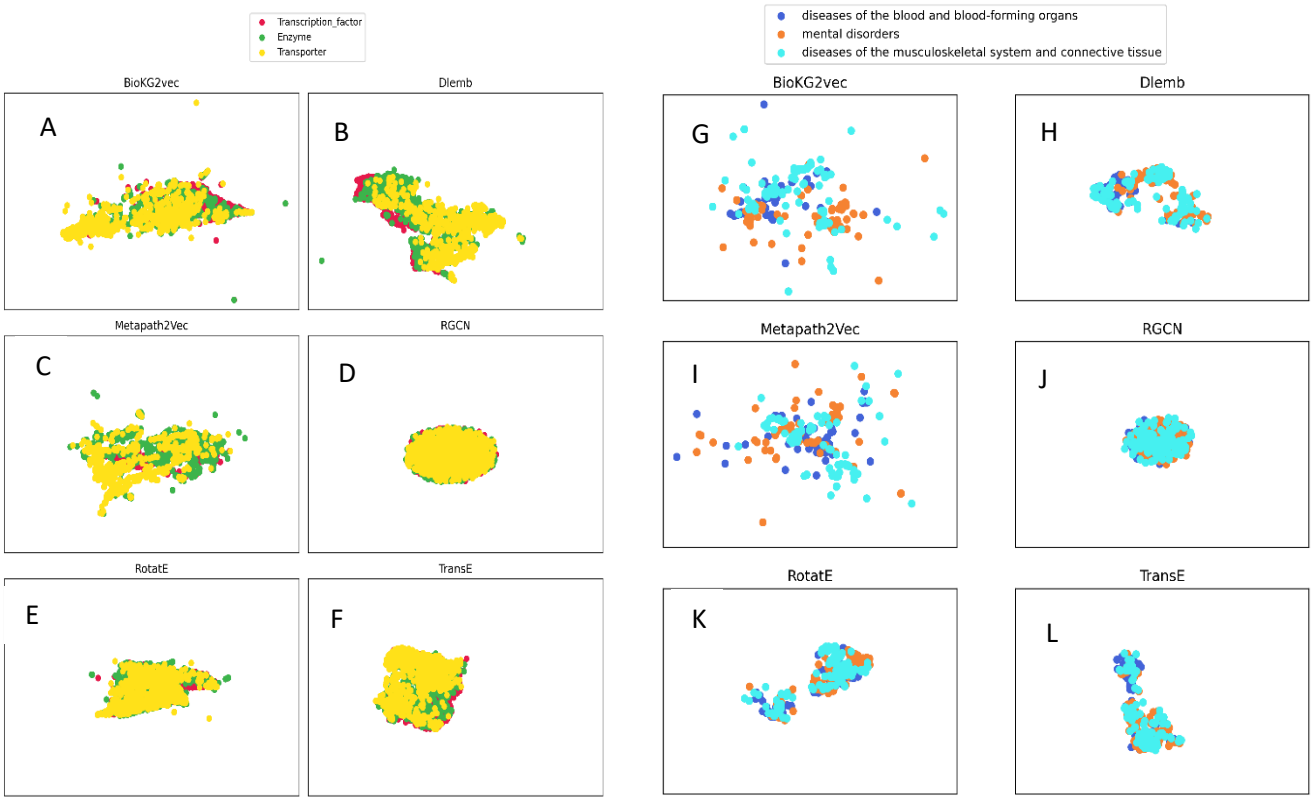

Figure 1: UMAP of gene-embeddings (panels A to F) and disease-embeddings (panel G to L) generated with BioKG2vec (A,G), Dlemb (B,H), Metapath2Vec (C,I), RGCN (D,J), RotatE (E,K) and TransE (F,L).

Table 2: Homogeneity score of K – means algorithm calculated for genes (number of clusters = 16) and diseases (number of clusters = 16). True labels are classification from ICD-9 and HPA for diseases and genes respectively

	Homogeneity score	
	Genes	**Diseases**
Metapath2Vec	**0.49**	**0.28**
Dlemb	0.35	0.17
RotatE	0.35	0.15
Trans-E	0.29	0.09
BioKG2vec	0.2	0.20
RGCN	0.008	0.02

The embeddings of gene products generated with Metapath2Vec produced more homogeneous clusters. Assigning every different gene and disease correctly to their category is a very complex task because of the high granularity of genes and disease classes (Supplementary Figures 4 and 5).

Model selection through grid search cross-validation

The best performing combination for GDA prediction was Metapath2Vec. Metapath2Vec coupled with concatenation of the gene and disease embedding as association representation and SVM with parameters C = 10 and kernel = rbf as classification algorithm. The whole output of the experiment is available in Supplementary Table 1. The following experiments were run using this combination.

Heterogeneous data integration and preprocessing

Pre-processing the ontologies leads to better ROCAUC and AUPRC compared to using embeddings generated with raw data. Nevertheless, adding heterogeneous data in the KG did not significantly affect the predictions of GDAs (Table 3). Integrating more data leads to similar performances which can be appreciated when comparing the results of generating the KG using HPO data with HPO and GO data. We must note the different results on the use of Dlemb algorithm (Supplementary Table 2). While the predictive power of Metapath2Vec is not affected by the preprocessing of the ontologies, Dlemb significantly improves the AUPRC and ROCAUC after preprocessing.

Table 3: ROCAUC and AUPRC of different experiments of GDAs predictions using Human Phenotype Ontology (HPO) + annotations (A), HPO ontology processed (B) and HPO + Gene Ontology (GO) + GO annotations. The embeddings were generated with Metapath2Vec and we used SVM as predictive algorithm, and operator concatenation for combining the embeddings.

Experiment	ROCAUC	AUPRC
HPO + HPO annotations raw	**0.95**	**0.98**
HPO + HPO annotations processed	0.93	0.97
HPO + HPO annotations + GO + GO annotations processed	0.93	0.97

The amount of training GDAs in the KG affects the prediction of GDAs

We tested the effect on the predictions caused by the increase of GDAs in the KG. We expect that increasing the amount of GDAs in the KG will increase the quality of the predictions. Figure 2 shows that the increase in the number of GDAs used for training the knowledge graph embeddings increases the values of ROCAUC and AUPRC.

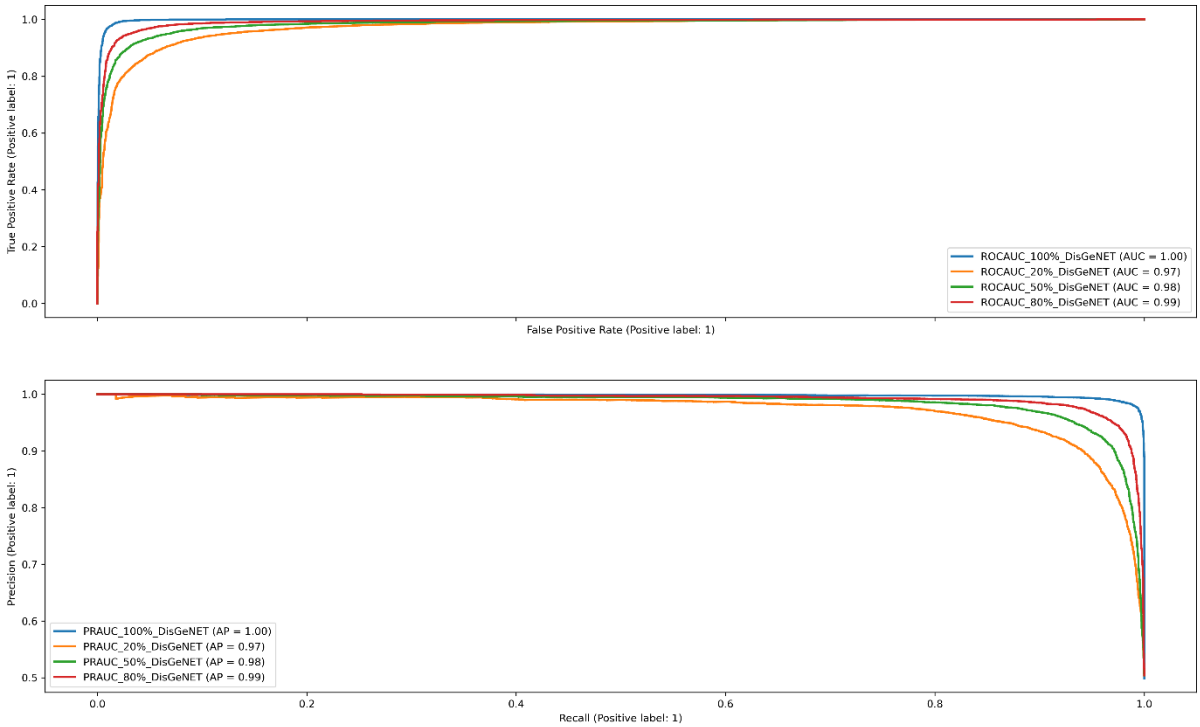

Figure 2: ROCAUC and PRAUC of the prediction of GDAs. Several KG embeddings are obtained using increasing percentages of known GDAs from 20% to 100%. Note: embeddings were generated with Metapath2Vec, using concatenation for combining embeddings and SVM for the classification/prediction algorithm.

Comparison with randomly generated embeddings

Supplementary Table 3 presents the outcomes of the experiment contrasting embeddings produced by Metapath2Vec with those generated randomly. Metapath2Vec embeddings reach an average ROCAUC of 0.93 while random generated embeddings have random metrics. These results are due to the biological information intrinsic to the embeddings since the effect of the number of GDAs was prevented by selecting associations of one disease only. In fact, the number of associations is a latent variable that is learned by the model.

Model generalization across different disease classes

Figure 3 shows the performance of the model trained on a specific ICD9 disease class and tested on all the others. Training and testing in diseases belonging to the same class leads to accurate predictions. However, embeddings generated with Metapath2Vec have poor prediction capabilities across different ICD-9 classes. Similar results were observed with randomly generated embeddings. Biological information encoded in Dlemb generated embeddings is translated across disease classes and we can see that some pairs of disease classes achieved a noteworthy prediction (e.g. the model trained for neoplasms predicts genes associated with diseases of circulatory system with an ROCAUC > 0.7) (Supplementary Figure 6). As expected, randomly generated embeddings show ROCAUC in the heatmap with random values.

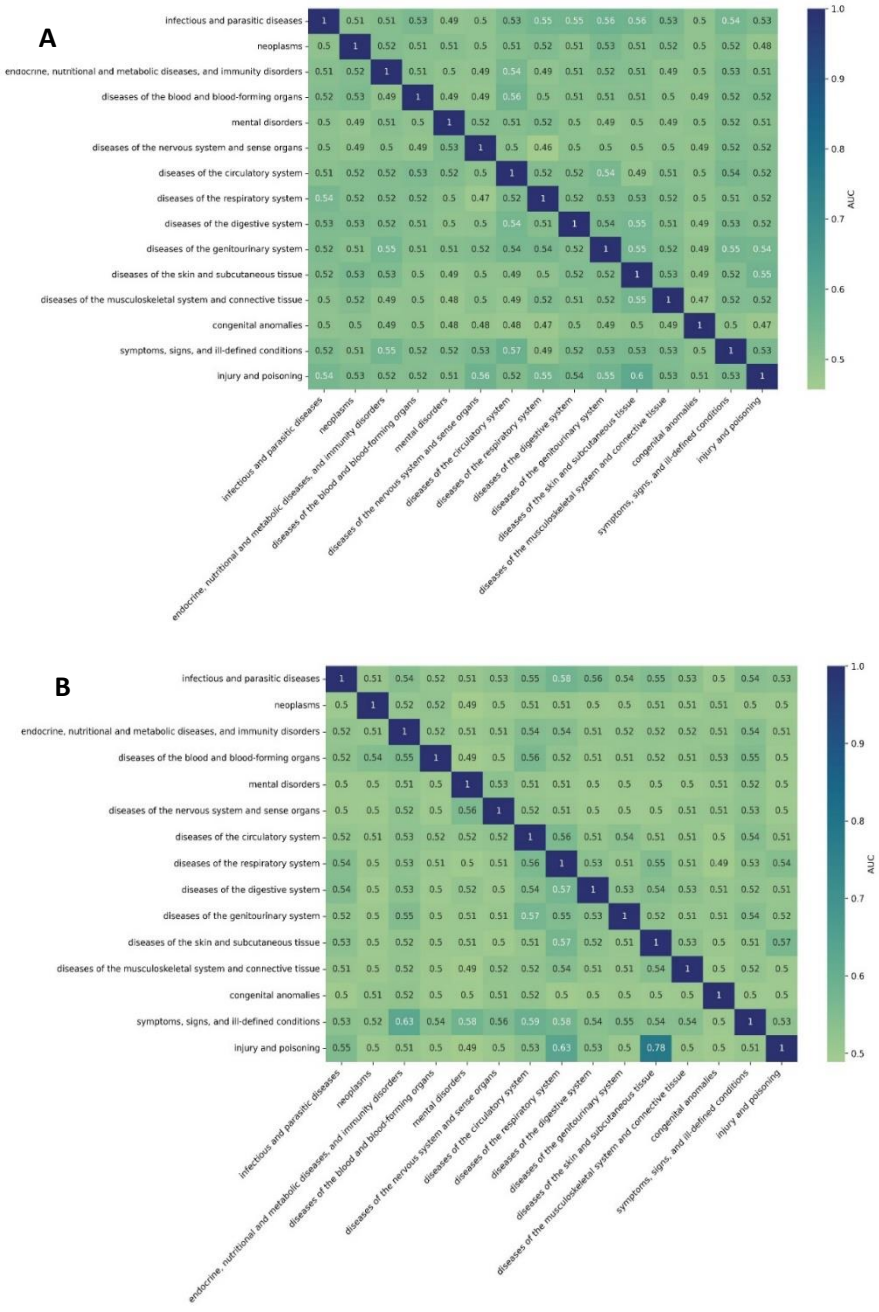

Figure 3: ROCAUC of the best performing combination for the prediction on different disease classes. A) Results for randomly generated embeddings, the ROCAUC shows random values. B) The embeddings were generated with Metapath2Vec, the GDAs representation was concatenation, and the algorithm was SVM with parameters C = 10 and kernel = rbf. We show the results of training a model on a specific ICD-9 disease class (rows), and then testing on the others (columns).

Computational Performance of the best algorithms for KGE

We compared the performance of the three algorithms that reached the highest ROCAUC during the grid search cross validation in terms of running time. In the supplementary figure 7 are reported the running times of 10 experiments. BioKG2vec and Dlemb are respectively ~100% and 360% faster than Metapath2Vec.

KGE successfully predicts genes associated to IDD

Finally, we used the selected prediction model with the best parametrization to predict GDAs for IDD. IDD is one of the main causes of low back pain, the largest cause of morbidity worldwide affecting 80% of people from Western countries during their lifetime [62]. IDD consists of the gradual deterioration of the intervertebral disc (IVD) in which the content of collagen and glycosaminoglycan decreases, and it becomes more dehydrated and fibrotic. Due to this, its anatomical areas nucleus pulposus (NP) and anulus fibrosus (AF) becomes less distinguishable [63]. Also, during IDD there is a catabolic shift in the biochemical processes of the disc environment with an increased expression of matrix degrading enzymes promoted by catabolic cytokines and vascularization of the tissues [2]. According to DisGeNET (curated sources), IDD is associated to TGFβ-1, HTRA1 and SPARC. We ran predictions for 20,951 genes, of those 445 were predicted to be associated to the disease and 93 with a probability > 0.95. The results of the top 10 prioritized genes are shown in Table 4.

The predictive analysis identifies the TGFβ-1 gene as the most promising candidate associated with Intervertebral Disc Degeneration (IDD), with isoforms TGFβ-2 and TGFβ-3 also receiving prioritization. Notably, TGFβ-1 emerges as the highest-scoring gene in DisGeNET's curated dataset related to disc degeneration. TGFβ plays a multifaceted role in various pathways associated with the homeostasis and turnover of the extracellular matrix in IDD [64]. Additionally, SMAD3 and SMAD2, integral genes in disc homeostasis, participate in the TGF-β pathway [65][66]. Matrix metalloproteinase 9 (MMP9) and matrix metalloproteinase 2 (MMP2) enzymes contribute significantly to IDD by participating in matrix degradation, targeting proteins expressed in the intervertebral disc like collagens and aggrecan. [67]. Moreover, LOX, crucial for cartilage homeostasis, presents a potential strategy for cartilage regeneration [68] , with studies indicating its anti-apoptotic effects in TNF-α treated rat NP-cells [69]. These genes were shown to have a role in IDD and could be further investigated to elucidate the mechanisms that lead to the degeneration of the disc.

To further explore the biological functions of these candidate genes, we performed a function enrichment analysis (Figure 4). The top prioritized genes are enriched in processes related to the extracellular matrix organization, pathways related to collagen formation, and extracellular matrix degradation, all of them related to IDD.

Table 4: Top 10 genes prioritized from the model with highest predictive capabilities

Gene ID	Gene Symbol	Probability
7040	TGFB1	1
4088	SMAD3	1
4318	MMP9	1
4015	LOX	1
7043	TGFB3	1
7046	TGFBR1	1
7042	TGFB2	1
1277	COL1A1	1
4313	MMP2	1
4087	SMAD2	1

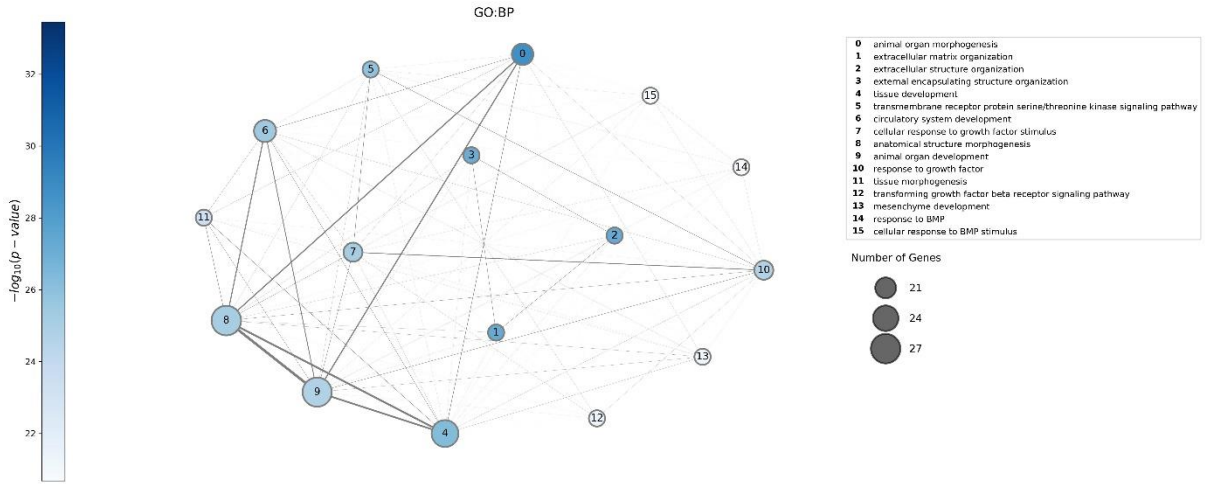

Figure 4: Gene ontology biological processes (GO:BP) function enrichment analysis on the genes with probability higher than 0.95 to be associated to C0158266 (n=93). To run the functional enrichment, we used g:Profiler. The nodes correspond to the pathway enriched in the gene set, their size is proportional to the number of genes belonging to that specific pathway and the colour is related to the significance of the enrichment in the gene set (calculated through hypergeometric distribution). An edge exists between 2 nodes if there are genes shared between the two pathways and the width of the edge is proportional to the number of the genes shared.

Discussion

In this work we investigated how KGE perform to predict gene-disease associations. First, we generated a KG by implementing heterogeneous biological information such as protein-protein interactions, gene-disease associations, drug-disease associations and drug-protein interactions, and ontologies The integration of multiple knowledge-based datasets prevented us from using syntactic-based approaches for embedding-creation such as OPA2VEC [70]. Syntactic approaches rely in the set of axioms only for obtaining the embeddings without the intermediate graph-based representation [57], so the input of the algorithm must be in Web-Ontology Language (OWL) format. Moreover, the integration of different ontologies is a challenging task and an active research topic [72].

In this study, we systematically assessed diverse methodologies for KGE construction and introduced two novel algorithms, namely BioKG2vec and Dlemb. Our comprehensive evaluation reveals that these algorithms exhibit superior performance compared to most existing methods. Notably, the parallelized implementation of both BioKG2vec and Dlemb results in substantially reduced running times in comparison to Metapath2Vec. This enhanced scalability facilitates the effective utilization of computational resources.

We conducted an extensive analysis of embeddings utilizing unsupervised machine learning techniques. Our investigation encompassed the integration of diverse data types and the comparison of GDA predictions using random features. Our findings revealed that augmenting the proportion of GDA within the KG enhances model performance. This observation suggests that task-specific embeddings implementation could enhance predictions, potentially leveraging the learning of pertinent features, as indicated elsewhere [27]. Furthermore, we applied KGE to prioritize new genes associated with IDD, illustrating their utility in inferring disease biomarkers even in scenarios with limited genetic data. Notably, our model, trained on a DisGeNET curated dataset containing merely 3 associations, prioritized 445 genes, which effectively reflected the underlying biology of IDD. In fact, the polygenic nature and epistatic interactions characteristic of non-communicable diseases pose challenges to comprehending the intricate biology underlying the development of complex conditions [73].

Finally, we emphasize the significance of scrutinizing the data quality employed in embedding creation, as predictive models can glean numerous latent features, potentially introducing bias to the outcomes.

Conclusions

In this work we carried out an extensive investigation on KGE from the generation and evaluation of the produced embeddings to the development of two new models for KGE generation and the utilization of the created embedding in a GDA prediction task. We showed that embeddings can effectively be implemented in the biomedical field to infer new knowledge over a certain domain. Nevertheless, many challenges remain open that require interdisciplinary collaboration to reach better outcomes in the healthcare sector.

Tool availability

Tool with embeddings generated with the top 3 best performing algorithms (Metapath2Vec, Dlemb and BioKG2vec) for GDAs association are available at:

https://github.com/freh-g/KGE

[1] Z. Sun, B. Liu, and Z.-J. Luo, "The Immune Privilege of the Intervertebral Disc: Implications for Intervertebral Disc Degeneration Treatment," *Int J Med Sci*, vol. 17, no. 5, pp. 685–692, 2020, doi: 10.7150/ijms.42238.

[2] T. Kadow, G. Sowa, N. Vo, and J. D. Kang, "Molecular Basis of Intervertebral Disc Degeneration and Herniations: What Are the Important Translational Questions?," *Clin Orthop Relat Res*, vol. 473, no. 6, pp. 1903–1912, Jun. 2015, doi: 10.1007/s11999-014-3774-8.

[3] J. P. G. Urban and S. Roberts, "Degeneration of the intervertebral disc.," *Arthritis Res Ther*, vol. 5, no. 3, pp. 120–30, 2003, doi: 10.1186/ar629.

[4] N. Kos, L. Gradisnik, and T. Velnar, "A Brief Review of the Degenerative Intervertebral Disc Disease," Dec. 01, 2019, *NLM (Medline)*. doi: 10.5455/medarh.2019.73.421-424.

[5] J. Xin, Y. Wang, Z. Zheng, S. Wang, S. Na, and S. Zhang, "Treatment of Intervertebral Disc Degeneration," Jul. 01, 2022, *Sociedade Brasileira de Matematica Aplicada e Computacional*. doi: 10.1111/os.13254.

[6] S. Kirnaz *et al.*, "Innovative Biological Treatment Methods for Degenerative Disc Disease," *World Neurosurg*, vol. 157, pp. 282–299, Jan. 2022, doi: 10.1016/j.wneu.2021.09.068.

[7] B. Jordan, "Genes and Non-Mendelian Diseases: Dealing with Complexity," *Perspect Biol Med*, vol. 57, no. 1, pp. 118–131, 2014, doi: 10.1353/pbm.2014.0002.

[8] J. M. Walker, "M e t h o d s i n M o l e c u l a r B i o l o g y ™ Series Editor." [Online]. Available: www.springer.com/series/7651

[9] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," May 05, 2017, *BioMed Central Ltd.* doi: 10.1186/s13059-017-1215-1.

[10] F. E. Agamah *et al.*, "Computational approaches for network-based integrative multi-omics analysis," Nov. 14, 2022, *Frontiers Media S.A.* doi: 10.3389/fmolb.2022.967205.

[11] B. Lee, S. Zhang, A. Poleksic, and L. Xie, "Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis," *Front Genet*, vol. 10, Jan. 2020, doi: 10.3389/fgene.2019.01381.

[12] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science (1979)*, vol. 347, no. 6224, p. 841, Feb. 2015, doi: 10.1126/science.1257601.

[13] S. K. Mohamed, A. Nounu, and V. Nováček, "Biological applications of knowledge graph embedding models," *Brief Bioinform*, vol. 22, no. 2, pp. 1679–1693, Mar. 2021, doi: 10.1093/bib/bbaa012.

[14] A. Hur, N. Janjua, and M. Ahmed, "A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead," in *Proceedings - 2021 IEEE 4th International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 99–103. doi: 10.1109/AIKE52691.2021.00021.

[15] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge Graphs: Opportunities and Challenges," *Artif Intell Rev*, vol. 56, no. 11, pp. 13071–13102, Nov. 2023, doi: 10.1007/s10462-023-10465-9.

[16] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, "A survey on knowledge graph embedding: Approaches, applications and benchmarks," *Electronics (Switzerland)*, vol. 9, no. 5, May 2020, doi: 10.3390/electronics9050750.

[17] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," *IEEE Trans Knowl Data Eng*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

[18] G. Wang, Y. Ding, Z. Xie, Y. Ma, Z. Zhou, and W. Qian, "RotatGAT: Learning Knowledge Graph Embedding with Translation Assumptions and Graph Attention Networks," in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/IJCNN55064.2022.9892206.

[19] S. Brin and L. Page, "Computer Networks and ISDN Systems," 1998. [Online]. Available: http://www.yahoo.com

[20] L. Wei, "Combining Graph Attention Mechanism and PageRank to Learn Graph-level Representations," 2022.

[21] S. Nunes, R. T. Sousa, and C. Pesquita, "Predicting Gene-Disease Associations with Knowledge Graph Embeddings over Multiple Ontologies," May 2021, [Online]. Available: http://arxiv.org/abs/2105.04944

[22] Z. Gao, Y. Pan, P. Ding, and R. Xu, "A knowledge graph-based disease-gene prediction system using multi-relational graph convolution networks.," *AMIA Annu Symp Proc*, vol. 2022, pp. 468–476, 2022, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/37128437

[23] W. Choi and H. Lee, "Identifying disease-gene associations using a convolutional neural network-based model by embedding a biological knowledge graph with entity descriptions," *PLoS One*, vol. 16, no. 10 October, Oct. 2021, doi: 10.1371/journal.pone.0258626.

[24] S. K. Mohamed, V. Nováček, and A. Nounu, "Discovering protein drug targets using knowledge graph embeddings," *Bioinformatics*, vol. 36, no. 2, pp. 603–610, Jan. 2020, doi: 10.1093/bioinformatics/btz600.

[25] X. Zhong and J. C. Rajapakse, "Graph embeddings on gene ontology annotations for protein–protein interaction prediction," *BMC Bioinformatics*, vol. 21, Dec. 2020, doi: 10.1186/s12859-020-03816-8.

[26] C. Ruiz, M. Zitnik, and J. Leskovec, "Identification of disease treatment mechanisms through the multiscale interactome," *Nat Commun*, vol. 12, no. 1, p. 1796, Mar. 2021, doi: 10.1038/s41467-021-21770-8.

[27] A. Fernández-Torras, M. Duran-Frigola, M. Bertoni, M. Locatelli, and P. Aloy, "Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque," *Nat Commun*, vol. 13, no. 1, p. 5304, Sep. 2022, doi: 10.1038/s41467-022-33026-0.

[28] R. Oughtred *et al.*, "The BioGRID interaction database: 2019 update," *Nucleic Acids Res*, vol. 47, no. D1, pp. D529–D541, Jan. 2019, doi: 10.1093/nar/gky1079.

[29] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res*, vol. 32, no. DATABASE ISS., Jan. 2004, doi: 10.1093/nar/gkh086.

[30] K. Luck *et al.*, "A reference map of the human binary protein interactome.," *Nature*, vol. 580, no. 7803, pp. 402–408, Apr. 2020, doi: 10.1038/s41586-020-2188-x.

[31] E. Wingender *et al.*, "The TRANSFAC system on gene expression regulation," 2001. [Online]. Available: www.gene-regulation.de/

[32] H. Hermjakob *et al.*, "IntAct: An open source molecular interaction database," *Nucleic Acids Res*, vol. 32, no. DATABASE ISS., Jan. 2004, doi: 10.1093/nar/gkh052.

[33] R. Goel, H. C. Harsha, A. Pandey, and T. S. K. Prasad, "Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis," 2012, *Royal Society of Chemistry*. doi: 10.1039/c1mb05340j.

[34] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," 2000. [Online]. Available: http://www.genome.ad.jp/kegg/

[35] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, "BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions," 2010. [Online]. Available: http://www.biomedcentral.com/1471-2105/11/213

[36] A. Ruepp *et al.*, "CORUM: The comprehensive resource of mammalian protein complexes," *Nucleic Acids Res*, vol. 36, no. SUPPL. 1, Jan. 2008, doi: 10.1093/nar/gkm936.

[37] P. V. Hornbeck *et al.*, "PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Res*, vol. 40, no. D1, Jan. 2012, doi: 10.1093/nar/gkr1122.

[38] A. Vinayagam *et al.*, "A directed protein interaction network for investigating intracellular signal transduction," *Sci Signal*, vol. 4, no. 189, Sep. 2011, doi: 10.1126/scisignal.2001699.

[39] J. Hastings, "Primer on Ontologies," in *Methods in Molecular Biology*, vol. 1446, Humana Press Inc., 2017, pp. 3–13. doi: 10.1007/978-1-4939-3743-1_1.

[40] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[41] L. M. Schriml *et al.*, "Human Disease Ontology 2018 update: classification, content and workflow expansion.," *Nucleic Acids Res*, vol. 47, no. D1, pp. D955–D962, Jan. 2019, doi: 10.1093/nar/gky1032.

[42] T. Groza *et al.*, "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease," *The American Journal of Human Genetics*, vol. 97, no. 1, pp. 111–124, Jul. 2015, doi: 10.1016/j.ajhg.2015.05.020.

[43] B. T. McInnes, T. Pedersen, and J. Carlis, "Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain.," *AMIA Annu Symp Proc*, vol. 2007, pp. 533–7, Oct. 2007, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18693893

[44] J. Piñero *et al.*, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Res*, vol. 48, no. D1, pp. D845–D855, Nov. 2019, doi: 10.1093/nar/gkz1021.

[45] A. S. Brown and C. J. Patel, "A standard database for drug repositioning," *Sci Data*, vol. 4, no. 1, p. 170029, Mar. 2017, doi: 10.1038/sdata.2017.29.

[46] O. Ursu *et al.*, "DrugCentral: Online drug compendium," *Nucleic Acids Res*, vol. 45, no. D1, pp. D932–D939, Jan. 2017, doi: 10.1093/nar/gkw993.

[47] S. M. Corsello *et al.*, "The Drug Repurposing Hub: a next-generation drug library and information resource," *Nat Med*, vol. 23, no. 4, pp. 405–408, Apr. 2017, doi: 10.1038/nm.4306.

[48] M. E. Sharp, "Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources," *J Biomed Semantics*, vol. 8, no. 1, p. 2, Dec. 2017, doi: 10.1186/s13326-016-0110-0.

[49] D. S. Wishart, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res*, vol. 34, no. 90001, pp. D668–D672, Jan. 2006, doi: 10.1093/nar/gkj067.

[50] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.10197

[51] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.06103

[52] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data."

[53] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2017, pp. 135–144. doi: 10.1145/3097983.3098036.

[54] A. Grover and J. Leskovec, "node2vec," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 855–864. doi: 10.1145/2939672.2939754.

[55] P. J. Thul and C. Lindskog, "The human protein atlas: A spatial map of the human proteome," *Protein Science*, vol. 27, no. 1, pp. 233–244, Jan. 2018, doi: 10.1002/pro.3307.

[56] W. H. Organization, "International classification of diseases : [9th] ninth revision, basic tabulation list with alphabetic index," 1978, *World Health Organization*.

[57] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jan. 2012, [Online]. Available: http://arxiv.org/abs/1201.0490

[58] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.01703

[59] M. Larralde *et al.*, “althonos/pronto: v2.5.4,” Apr. 2023, *Zenodo*. doi: 10.5281/zenodo.7814219.

[60] U. Raudvere *et al.*, “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Res*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.

[61] R. Albert, “Scale-free networks in cell biology,” Nov. 01, 2005. doi: 10.1242/jcs.02714.

[62] P. Bermudez-Lekerika *et al.*, “Immuno-Modulatory Effects of Intervertebral Disc Cells,” *Front Cell Dev Biol*, vol. 10, Jun. 2022, doi: 10.3389/fcell.2022.924692.

[63] J. A. Buckwalter, “Aging and Degeneration of the Human Intervertebral Disc,” *Spine (Phila Pa 1976)*, vol. 20, no. 11, pp. 1307–1314, Jun. 1995, doi: 10.1097/00007632-199506000-00022.

[64] S. Chen, S. Liu, K. Ma, L. Zhao, H. Lin, and Z. Shao, “TGF-β signaling in intervertebral disc health and disease,” Aug. 01, 2019, *W.B. Saunders Ltd*. doi: 10.1016/j.joca.2019.05.005.

[65] S. Tseranidou, M. Segarra-Queralt, F. K. Chemorion, C. Le 3 Maitre, J. Piñero, and J. Noailly, “Nucleus Pulposus Cell Network Modelling in the Intervertebral Disc”, doi: 10.1101/2024.09.18.613636.

[66] C.-G. Li *et al.*, “A Continuous Observation of the Degenerative Process in the Intervertebral Disc of Smad3 Gene Knock-Out Mice,” *Spine (Phila Pa 1976)*, vol. 34, no. 13, pp. 1363–1369, Jun. 2009, doi: 10.1097/BRS.0b013e3181a3c7c7.

[67] N. V. Vo, R. A. Hartman, T. Yurube, L. J. Jacobs, G. A. Sowa, and J. D. Kang, “Expression and regulation of metalloproteinases and their inhibitors in intervertebral disc aging and degeneration,” *The Spine Journal*, vol. 13, no. 3, pp. 331–341, Mar. 2013, doi: 10.1016/j.spinee.2012.02.027.

[68] W. Lin, L. Xu, and G. Li, “Molecular Insights Into Lysyl Oxidases in Cartilage Regeneration and Rejuvenation,” Apr. 30, 2020, *Frontiers Media S.A.* doi: 10.3389/fbioe.2020.00359.

[69] R. Zhao *et al.*, “Lysyl oxidase inhibits TNF-α induced rat nucleus pulposus cell apoptosis via regulating Fas/FasL pathway and the p53 pathways,” *Life Sci*, vol. 260, Nov. 2020, doi: 10.1016/j.lfs.2020.118483.

[70] Fatima Zohra Smaili, “OPA2Vec: combining formal and informal contentof biomedical ontologies to improvesimilarity-based prediction,” *Bioinformatics*, vol. 33, no. 16, pp. 1–7, Aug. 2017, doi: 10.1093/bioinformatics/xxxxxx.

[71] M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, “Semantic similarity and machine learning with ontologies,” Jul. 01, 2021, *Oxford University Press*. doi: 10.1093/bib/bbaa199.

[72] I. Osman, S. Ben Yahia, and G. Diallo, “Ontology Integration: Approaches and Challenging Issues,” *Information Fusion*, vol. 71, pp. 38–63, Jul. 2021, doi: 10.1016/j.inffus.2021.01.007.

[73] P. M. Visscher, L. Yengo, N. J. Cox, and N. R. Wray, “Discovery and implications of polygenicity of common diseases,” *Science (1979)*, vol. 373, no. 6562, pp. 1468–1473, Sep. 2021, doi: 10.1126/science.abi8206.