DRKG - Drug Repurposing Knowledge Graph for Covid-19

Vassilis N. Ioannidis*[1,2], Xiang Song*[4], Saurav Manchanda*[1,2], Mufei Li[4], Xiaoqin Pan[3], Da Zheng[1], Xia Ning[5], Xiangxiang Zeng[3], and George Karypis*[1,2]

[1]Amazon Web Services AI
[2]University of Minnesota
[3]Hunan University
[4]Amazon Web Services Shanghai AI Lab
[5]The Ohio State University

***Abstract*—The timeline of the Covid-19 pandemic showcases the dire need for fast development of effective treatments for new diseases. Drug-repurposing (DR) is a drug discovery strategy from existing drugs that significantly shortens the time and reduces the cost compared to denovo drug discovery. This paper details an effort to create a comprehensive biological knowledge graph relating genes, compounds, diseases, biological processes, side effects and symptoms termed Drug Repurposing Knowledge Graph (DRKG). DRKG includes information from different databases such as DrugBank, String, GNBR, and data collected from recent publications particularly related to Covid19. DR can be addressed by predicting novel interactions among genes and chemical compounds, which can be formulated as a link prediction task over the DRKG. We provide methods to analyze the constructed DRKG and filter noisy links and nodes. Graph machine learning models may utilize DRKG for DR and predict whether existing drugs successfully inhibit certain pathways related to Covid-19 host proteins. Using state-of-the-art knowledge graph embedding models we learn embeddings for entities and relations in the DRKG. We also perform analysis to validate that the graph structure and the learned embedding are of high quality. Finally, using the learned KGE on the proposed DRKG we evaluate drug repurposing for Covid-19. The results corroborate that several drugs used in clinical trials were identified as possible drug candidates. Finally, by comparing the drug repurposing results of DRKG with those of the constituent databases, we confirm he merits to constructing a comprehensive DRKG. The methods presented in this paper are implemented in the efficient deep graph learning (DGL)[1] library. The DRKG, entity & relation embeddings, and the source code for the analysis presented in this paper is publicly available.[2]**

I. Introduction

The International Public Health Emergency of corona virus disease 2019 (COVID-19) exemplifies the urgency for improving the efficiency and speed of discovering new treatments. Unfortunately, de novo drug discovery is a lengthy and costly process typically requiring 10–15 years and costing over $2.6 billion for each new FDA approved drug [1], [2], [3]. An alternate approach that can dramatically reduce the time to discover new treatments is *drug repurposing* (DR) (or repositioning), which seeks to redevelop existing drugs for use in different diseases. DR leverages the fact that common molecular pathways contribute to different diseases and hence some drugs may be reused [4]. It capitalizes on the existence of detailed information on approved drugs and many abandoned compounds related to their pharmacology, formulation, dose, and potential toxicity [4].

Drug repurposing relies on identifying novel interactions among biological entities like genes and compounds. Traditional approaches for doing that rely on costly and time-consuming experimental methodologies [5]. As a result, several approaches have been developed that aim to leverage the diverse types of information that already exists about the drugs, their targets, and the diseases in order to reduce the cost and speedup drug repurposing. Among them, approaches that represent the existing information in a form of a knowledge graph and deploy graph-based machine learning techniques based on graph neural networks [6] and knowledge graph embedding models [7] have gained popularity [8], [9], [10].

A. Contributions

In this project we construct a diverse Drug Repurposing Knowledge Graph, termed DRKG, and provide a set of machine-learning tools to be used for speeding up drug repurposing. We collect interactions from the following publicly-available data sources: (i) Drugbank [11], (ii) Global Network of Biomedical Relationships (GNBR) [12], (iii) Hetionet[13], (iv) STRING [14], (v) IntAct [15], (vi) DGIdb [16], and (vii) relations from bibliographic sources [17], [18], [9]. We map the biological entities of the different databases to a common ID space, which allows us to link entities across databases, and we filter the initial data to remove noisy links and entities. In total, DRKG contains 97,055 vertices belonging to 13 types of entities, and 5,869,294 edges belonging to 107 types of relations. In addition, DRKG contains a number of Covid-19 related proteins and genes, as extracted from recent publications [17], [18], [9], [13], [19].

To analyze DRKG, we formulate and solve the link prediction task using models that compute knowledge graph embedding (KGE) [7]. We perform analysis to validate that the graph structure and the learned embeddings are of high

*equal contribution
[1]https://www.dgl.ai/
[2]https://github.com/gnn4dr/DRKG/

quality. Our analysis shows that similar biological entities and relations have similar embeddings that corroborates insights from biology and hence DRKG can be used for developing machine learning models. Finally, we used these embeddings to evaluate how well DRKG can be used to identify drugs that can be repurposed for Covid-19. Our results show that among the highest scoring drugs, there are several drugs undergoing clinical trials.

These results illustrate that using the DRKG, one can apply machine learning models to predict new links and facilitate drug repurposing for novel diseases. Finally, by comparing the drug repurposing results of DRKG with those of the constituent databases, we confirm he merits to constructing a comprehensive DRKG.

The DRKG, entity & relation embeddings, and the source code for the analysis presented in this paper is publicly made available.[3]

II. Background

A. Definitions & Notation

A graph $G = (V, E)$ is composed of two sets—the set of nodes V (also called vertices) and the set of edges E (also called arcs). Each edge connects a pair of nodes indicating that there is a relation between them. This relation can either be undirected, e.g., capturing symmetric relations between nodes, or directed, capturing asymmetric relations. Depending on the edges' directionality, a graph can be *directed* or *undirected*. Graphs can be either *homogeneous* or *heterogeneous*. In a homogeneous graph, all the nodes represent instances of the same type and all the edges represent relations of the same type. In contrast, in a heterogeneous graph, the nodes and edges can be of different types. Finally, a graph can either be a *simple graph* or a *multigraph*. In a simple graph there is only a single directed edge connecting a pair of nodes and it does not have loops. In contrast, in a multigraph there can be multiple (directed) edges between the same pair of nodes and can also contain loops. These multiple edges are typically of different types and as such most multigraphs are heterogeneous.

A *knowledge graph* (KG) is a directed heterogeneous multigraph whose node and relation types have domain-specific semantics. KGs allow us to encode the knowledge into a form that is human interpretable and amenable to automated analysis and inference. A node in a knowledge graph represents an entity and an edge represents a relation between two entities. The edges are usually in the form of triplets (h, r, t), each of which indicates that a pair of entities h (*head*) and t (*tail*) are coupled via a relation r.

Knowledge graph embeddings are low-dimensional representation of entities and relations. These embeddings carry the information of the entities and relations in the knowledge graph and are widely used in tasks, such as knowledge graph completion and recommendation. Throughput the paper, we denote the embedding vector of head entity, tail entity and relation with $\mathbf{h}$, $\mathbf{t}$ and $\mathbf{r}$, respectively; all embeddings have the same dimension size of d.

[3]https://github.com/gnn4dr/DRKG/

TABLE I: Knowledge graph models.

Models	score function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$
TransE [22]	$-\|\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|\|_{1/2}$
TransR [23]	$-\|\|M_r\mathbf{h} + \mathbf{r} - \mathbf{M_r t}\|\|_2^2$
DistMult [24]	$\mathbf{h}^\top \text{diag}(\mathbf{r})\mathbf{t}$
ComplEx [25]	$\text{Real}(\mathbf{h}^\top \text{diag}(\mathbf{r})\overline{\mathbf{t}})$
RESCAL [26]	$\mathbf{h}^\top \mathbf{M_r t}$
RotatE [27]	$-\|\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|\|^2$

M_r is a relation-specific projection matrix, $\overline{\mathbf{t}}$ is the conjugate of the complex vector $\mathbf{t}$, $\circ$ denotes Hadamard product, and TransE uses L1 or L2 norm in its score function.

B. Knowledge Graph Embedding (KGE) Models

The knowledge graph embeddings are computed so that they satisfy certain properties; i.e., they follow a given *KGE model*. A KGE model defines a score function that measures the distance of two entities relative to its relation type in the low-dimensional embedding space. This score function is defined on the triplets and during training it optimizes a loss function that maximizes the scores on triplets that exist in the knowledge graph (*positive* triplets) and minimizes the scores on triplets that do not exist (*negative* triplets). In this work we use the logistic loss for KGE model training given by

$$\min \sum_{\mathbf{h},\mathbf{r},\mathbf{t} \in \mathbb{D}^+ \cup \mathbb{D}^-} \log(1 + \exp(-y \times f(\mathbf{h}, \mathbf{r}, \mathbf{t}))) \qquad (1)$$

where $\mathbb{D}^+$ and $\mathbb{D}^-$ are the positive and negative sets of triplets, and $y = 1$ if the triplet corresponds to a positive example and -1 otherwise. The negative triplets are usually constructed by replacing the entities or relations in positive triplets with entities and relations randomly sampled from the knowledge graph. For a review of negative sampling strategies see [20].

Many score functions have been developed to train knowledge graph embeddings [21] and Table I lists the score functions of some of them. TransE and TransR are representative translational models that explore the Euclidean distance-based scoring functions, while DistMult, ComplEx, and RESCAL are semantic matching models that exploit similarity-based scoring functions.

Obtaining embeddings of entities and relations in the DRKG is beneficial for a number of downstream tasks. First, the embeddings may be used to clean the DRKG from noisy triplets that may appear due to erroneous entries in the data sources. Second, the embeddings can be used to establish the similarity of different entities and relation types such as diseases, drugs, interaction types, and obtain clusters of similar entities. Third, the embeddings can be used to predict putative new links between the DRKG's entities, which can contribute to our overall knowledge about the diseasome. For example, embeddings can be used to discover new side-effects of drugs and identify alternate therapuetic uses of existing drugs (see bellow).

III. Drug Repurposing Knowledge graph

A. Data sources

TABLE II: Extracted triplets from Drugbank.

Interaction	Relation type	#triplets
Compound:Gene	target	19158
	enzyme	4923
	carrier	720
Compound:Compound	ddi-interactor-in	1379271
Compound:ATC	x-atc	15750
Compound:Disease	treats	4968

TABLE III: Threshold values for GNBR relations.

Interaction	Frequency threshold	Confidence threshold
Gene:Gene	5	0.3
Gene:Disease	3	0.4
Compound:Gene	3	0.3
Compound:Disease	3	0.5

1) Drugbank: The DrugBank database [11] is a bioinformatics and cheminformatics knowledge base with rich drug data information as well as comprehensive drug target information. We used the latest version (version 5.1.5) which contains 13563 drug entries. We used Bio2RDF[4] to extract the triplets corresponding to the relations shown in Table II from Drugbank's original XML format. Note that the *treats* relation corresponds to Drugbank's *associated conditions*. Also, since the ATC classification is hierarchical, for each triplet involving a specific ATC code, we also included triplets corresponding to the higher-level ATC codes. For example, if (DB09344, x-atc, Atc::C05BB03) is an original triplet, we added the following triplets: (DB09344, x-atc, Atc::C05BB), (DB09344, x-atc, Atc::C05B), (DB09344, x-atc, Atc::C05), and (DB09344, x-atc, Atc::C). We excluded entities that participated in less than 2 triplets and triplets connecting to these entities. In total, we extracted from Drugbank 1419822 triplets and 18729 entities whose statistics are shown in Table X.

2) GNBR: Global Network of Biomedical Relationships (GNBR) [12] uses NCBI's PubTator [28] annotations to identify instances of chemical, gene, and disease names in Medline abstracts, and applies dependency parsing to find dependency paths between pairs of entities in individual sentences. These dependency paths are grouped into semantically-related categories, to provide relations (with confidence) between among enitities that appear together in a sentence. GNBR includes gene-gene, gene-disease, drug-gene, and drug-disease interactions.

To eliminate the entity pairs co-occurring by chance, we only considered the pairs that co-occur in the same sentence more than a threshold number of times (column 'frequency threshold' in Table III). Next, to find all relations between the pair of entities, we aggregated the confidence for the pair related by a particular relation, over all the sentences in which that pair of entities co-occurs. We again selected only those relations whose confidence is more than a threshold (column 'confidence threshold' in Table III), so as to remove noisy relations occurring by chance. Further, we also remove all the relations from a gene to itself. Our final processed dataset extracted from GNBR contains 66722 gene-gene interactions, 95400 gene-disease interactions, 80803 drug-gene interactions, and 77782 drug-disease interactions. The statistics of the extracted relations are shown in Table IV

[4]https://github.com/bio2rdf/bio2rdf-scripts

TABLE IV: Extracted triplets from GNBR.

Interaction	Relation type	#triplets
Gene:Gene	(B) binding, ligand (esp. receptors)	8164
	(W) enhances response	280
	(V+) activates, stimulates	8689
	(E+) increases expression/production	10838
	(E) affects expression/production (neutral)	418
	(I) signaling pathway	5434
	(H) same protein or complex	2509
	(Rg) regulation	11018
	(Q) production by cell population	19372
Gene:Disease	(U) causal mutations	6432
	(Ud) mutations affecting disease course	407
	(D) drug targets	500
	(J) role in pathogenesis	30234
	(Te) possible therapeutic effect	2836
	(Y) polymorphisms alter risk	1948
	(G) promotes progression	2055
	(Md) biomarkers (diagnostic)	1279
	(X) overexpression in disease	1324
	(L) improper regulation linked to disease	48384
Compound:Gene	(A+) agonism, activation	1568
	(A-) antagonism, blocking	1108
	(B) binding, ligand (esp. receptors)	7170
	(E+) increases expression/production	1970
	(E-) decreases expression/production	2918
	(E) affects expression/production (neutral)	32743
	(N) inhibits	12521
	(O) transport, channels	5573
	(K) metabolism, pharmacokinetics	12411
	(Z) enzyme activity	2821
Compound:Disease	(T) treatment/therapy (including investigator)	54020
	(C) inhibits cell growth (esp. cancers)	1739
	(Sa) side effect/adverse event	16923
	(Pr) prevents, suppresses	966
	(Pa) alleviates, reduces	2619
	(J) role in disease pathogenesis	1020
	(Mp) biomarkers (of disease progression)	495

3) Hetionet: Hetionet [13] is a heterogeneous information network of biomedical knowledge assembled from 29 different databases relating genes, compounds, diseases and other [13]. We extracted 2250197 triplets from 24 relation types as shown in Table V and 45279 entities belong to 11 entities types as shown in Table X.

4) STRING: STRING [14] is a database of established and predicted protein-protein interactions [14]. The interactions include direct (physical) and indirect (functional) associations and are extracted from computational prediction, knowledge transfer between organisms, and interactions aggregated from other databases. We extracted the interactions whose score is greater than or equal to 0.6, resulting in 1496708 triplets from 7 relation types as shown in Table VI and 18316 gene entities.

5) IntAct: IntAct is an open source database that contains molecular interaction data [15]. IntAct provides gene to gene as well as gene to chemical compounds interactions. The extracted relations and entities are in Tables VII and X.

6) DGIdb: DGIdb [16] is a drug–gene interaction database that consolidates, organizes and presents drug–gene interac-

TABLE V: Extracted triplets from the Hetionet knowledge graph.

Interaction	Relation type	#triplets
Anatomy:Gene	AdG (downregulation)	102240
	AeG (expression)	526407
	AuG (upregulation)	97848
Disease:Anatomy	DlA (localization)	3602
Compound:Gene	CbG (binding)	11571
	CdG (downregulation)	21102
	CuG (upregulation)	18756
Compound:Disease	CpD (palliation)	390
	CtD (treatment)	755
Compound:Compound	CrC (resemblence)	6486
Disease:Gene	DaG (association)	12623
	DdG (downregulation)	7623
	DuG (upregulation)	7731
Disease:Disease	DrD (resemblence)	543
Gene:Gene	GcG (covariation)	61690
	GiG (interaction)	147164
	Gr>G (regulation)	265672
Disease:Symptom	DpS (present)	3357
Compound:Side-effect	CcSE (causes)	138944
Gene:Biological-process	GpBP (participation)	559504
Gene:Molecular-function	GpMF participation)	97222
Gene:Cellular-component	GpCC participation)	73566
Gene:Pathway	GpPW (participation)	84372
Compound:Pharmacologic-class	PCiC (inclusion)	1029

TABLE VI: Extracted triplets from STRING.

Interaction	Relation type	#triplets
Gene:Gene	REACTION	400426
	BINDING	315875
	ACTIVATION	81355
	CATALYSIS	343533
	INHIBITION	28959
	PTMOD	15113
	EXPRESSION	757
	OTHER	310690

TABLE VII: Extracted triplets from IntAct.

Interaction	Relation type	#triplets
Compound-Gene	DIRECT INTERACTION	155
	PHYSICAL ASSOCIATION	203
	ASSOCIATION	1447
Gene-Gene	ASSOCIATION	112390
	PHYSICAL ASSOCIATION	129318
	COLOCALIZATION	3468
	DEPHOSPHORYLATION REACTION	303
	CLEAVAGE REACTION	93
	DIRECT INTERACTION	6950
	PHOSPHORYLATION REACTION	1328
	ADP RIBOSYLATION REACTION	58
	UBIQUITINATION REACTION	371
	PROTEIN CLEAVAGE	67

TABLE VIII: Extracted triplets from DGIdb.

Interaction	Relation type	#triplets
Compound:Gene	INHIBITOR	5971
	ANTAGONIST	3006
	OTHER	11070
	AGONIST	3012
	BINDER	143
	MODULATOR	243
	BLOCKER	979
	CHANNEL BLOCKER	352
	ANTIBODY	188
	POSITIVE ALLOSTERIC MODULATOR	618
	ALLOSTERIC MODULATOR	317
	ACTIVATOR	316
	PARTIAL AGONIST	75

TABLE IX: Extracted triplets from publications.

Interaction	Relation type	#triplets
Disease:Gene	Coronavirus_ass_host_gene	129
	Covid2_acc_host_gene	332
Gene:Gene	HumGenHumGen	58094
	VirGenHumGen	535
Compound:Gene	DrugVirGen	1165
	DrugHumGen	24501

tions and gene druggability information from papers, and online databases [16]. We extracted 26290 triplets from 13 relation types as shown in Table VIII and 2551 gene entities and 6348 compound entities.

7) Data related to Covid-19: To further enrich our data with information related to Covid-19, we included interaction data from three recent publications (shown in Table IX).

The work by Ge et. al. [17] developed a data-driven drug-repurposing framework that utilizes a biological network to discover the potential drug candidates against SARS-CoV-2. From that work, we extracted the biological network describing the interactions among host human proteins, virus proteins and chemical compounds. The proteins are indexed by the UniProt ID and the chemical compounds by their InChIKey.

In an effort to discover antiviral drugs for Covid-19, Gordon et. al. [18] cloned, tagged and expressed 26 of the 29 viral proteins in human cells and identified the physically associated human proteins. They identified 67 druggable human proteins targeted by 69 existing FDA-approved drugs, drugs in clinical trials and/or preclinical compounds.

A framework for drug-repurposing is introduced by Zhou et. al. [9], where the task is to repurpose drugs that are effective for certain related coronavirus strains such as IBV, HCoV-229E, HCoV-NL63, SARS, MERS and MHV. From that work, we extract relations among the aforementioned diseases and host proteins.

B. Naming conventions and normalization

From each dataset, we extracted a list of triplets in the form of (head-entity, relation-type, tail-entity). For representing entities, we use an entity type identifier followed by a unique ID of the specific entity, e.g., Gene::22947[5]. For representing

[5]https://www.ncbi.nlm.nih.gov/gene/22947

relations, we use a naming convention that combines the name of the data source, the name of the relation, and the types of head and tail entities that are involved e.g., DGIDB::INHIBITOR::Gene:Compound.

Data sources use one of several ID spaces to represent genes, compounds, diseases and others. For example, the same chemical compound may be represented in the drugbank compound ID space in DrugBank and in the chembl compound ID space in the DGIdb. To ensure that information from different sources in integrating correctly, we map biological entities to a common ID space using the following rules:

- Compound entities are mapped to the drugbank compound ID space and if not possible to the chembl compound ID space. If a compound can not be found to either of the two we use the native ID space and we include the name of the source as part of the entity's name (e.g., Compound::brenda:169533[6]).
- Gene entities are mapped to the Entrez ID space.
- Disease entities are mapped to the MESH ID space [29].
- The remaining biological entities appear only in a single data source and hence we use the data source's ID.

These rules are applied to the biological entities per database to map the entities to the common ID space. Finally, in order to avoid relations for which we do not have enough data to train good embeddings, we exclude relations types that have less than 50 edges.

C. The DRKG – Putting everything together

The extracted set of normalized triplets from the previous datasets constitute the *Drug Repurposing Knowledge Graph* (DRKG) that we created. It contains 97,055 entities belonging to 13 entity-types. The type-wise distribution of the entities is shown in Table X. DRKG contains a total of 5,869,294 triplets belonging to 107 relation-types. Table XI shows the number of triplets between different entity-type pairs for DRKG and various data sources. The per-relation-type statistics have already been discussed in Section III-A. Figure 1 depicts the possible interactions between the entity-type pairs in DRKG.

IV. Using Knowledge Graph Embeddings on DRKG for Drug Repurposing

Drug repurposing (DR) refers to using existing drugs for new therapeutic indications [4]. In this section, we formalize the DR objective as a link prediction task over the DRKG that can be solved by KGE models. In the appendix we include several data analysis techniques to verify that the constructed DRKG and the learned KG embeddings are of high quality.

A. Formulating drug repurposing as knowledge graph completion

DR refers to using existing drugs for new therapeutic indications. In the context of knowledge graphs, DR can be formulated as to predict new links between drug entities and disease entities of link type *treat*, or between drug entities and gene entities of link type *inhibit* or *bind* where the genes are related to the disease of interest (e.g., involved in related pathways). In Section IV, we validate DR on the DRKG for the Covid-19 disease, where we use the direct link formulation of DR. By using our comprehensive DRKG, researchers can address drug-repurposing for a variety of diseases such as HIV and SARS, as well as, the novel Covid-19.

[6]https://www.brenda-enzymes.org/ligand.php?brenda_ligand_id=169533

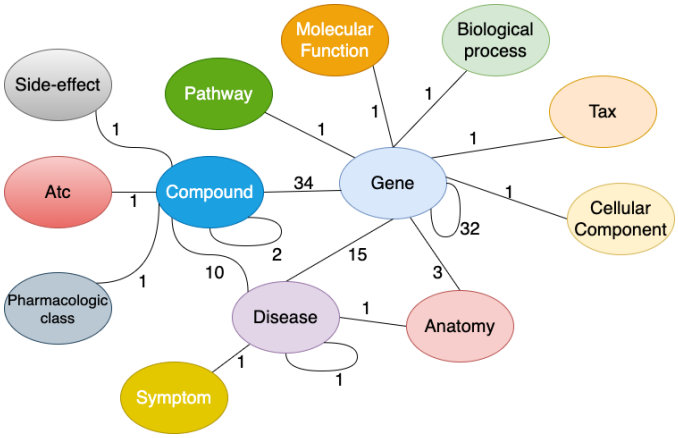

Fig. 1: Illustration of the constructed DRKG. The number next to an edge indicates the number of relation-types among the corresponding entity types in the DRKG.

B. Using knowledge graph embeddings for link prediction

Here, we analyze DRKG by learning a TransE KGE model that utilizes the ℓ_2 distance; see also Section II-B. By optimizing equation (1) we obtain vector embeddings of dimension 400×1 for all biological entities and relations participating in the DRKG. Consider the triplet (h, r, t) of the DRKG and the associated score as

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \gamma - ||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2 \qquad (2)$$

where γ is a parameter of the TransE model that is set to 12.0. For each triplet (h, r, t), the closer $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ is to γ, the more confident the model is that the head h and tail t entities are connected under relation r.

C. Drug repurposing using different relation types

In this section, we evaluate our DRKG in the drug repurposing task for the Covid-19 using the trained TransE KGE model in Section B. Here we use corona-virus diseases, including SARS, MERS and SARS-COV2, as target diseases representing Covid-19. We consider two formulations for the DR task. The first one predicts direct links between the disease entities and the drug entitites in the DRKG, while the second one predicts links among gene entities that are inhibited by drug entities where the genes are associated with the target disease.[7] We select FDA-approved drugs in Drugbank as candidates, while we exclude drugs with molecule weight less than 250 daltons, as many of certain drugs are actually

[7]The corresponding code in GitHub https://github.com/gnn4dr/COVID-19-KG/tree/master/drug_repurpose

TABLE X: Number of nodes per node type in the DRKG and the data sources.

Entity type	Drugbank	GNBR	Hetionet	STRING	IntAct	DGIdb	Bibliography	Total Entities
Anatomy	-	-	400	-	-	-	-	400
Atc	4,048	-	-	-	-	-	-	4,048
Biological Process	-	-	11,381	-	-	-	-	11,381
Cellular Component	-	-	1,391	-	-	-	-	1,391
Compound	9,708	11,961	1,538	-	153	6,348	6,250	24,313
Disease	1,182	4,746	257	-	-	-	33	5,103
Gene	4,973	27,111	19,145	18,316	16,321	2,551	3,181	39,220
Molecular Function	-	-	2,884	-	-	-	-	2,884
Pathway	-	-	1,822	-	-	-	-	1,822
Pharmacologic Class	-	-	345	-	-	-	-	345
Side Effect	-	-	5,701	-	-	-	-	5,701
Symptom	-	-	415	-	-	-	-	415
Tax	-	215	-	-	-	-	-	215
Total	19,911	44,033	45,279	18,316	16,474	8,899	9,464	97,238

TABLE XI: Number of interactions in the DRKG and the data sources.

Entity-type pair	Drugbank	GNBR	Hetionet	STRING	IntAct	DGIdb	Bibliography	Total interactions
Gene:Gene	-	667,22	474,526	1,496,708	254,346	-	58,629	2,350,931
Compound:Gene	24,801	80,803	51,429	-	1,805	26,290	25,666	210,794
Disease:Gene	-	95,399	27,977	-	-	-	461	123,837
Atc:Compound	15,750	-	-	-	-	-	-	15,750
Compound:Compound	1,379,271	-	6,486	-	-	-	-	1,385,757
Compound:Disease	4,968	77,782	1,145	-	-	-	-	83,895
Gene:Tax	-	14,663	-	-	-	-	-	14,663
Gene:Biological-process	-	-	559,504	-	-	-	-	559,504
Disease:Symptom	-	-	3,357	-	-	-	-	3,357
Anatomy:Disease	-	-	3,602	-	-	-	-	3,602
Disease:Disease	-	-	543	-	-	-	-	543
Anatomy:Gene	-	-	726,495	-	-	-	-	726,495
Gene:Molecular-function	-	-	97,222	-	-	-	-	97,222
Compound:Pharmacologic-class	-	-	1,029	-	-	-	-	1,029
Gene:Cellular-component	-	-	73,566	-	-	-	-	73,566
Gene:Pathway	-	-	84,372	-	-	-	-	84,372
Compound:Side-effect	-	-	138,944	-	-	-	-	138,944
Total	1,424,790	335,369	2,250,197	1,496,708	256,151	26,290	84,756	5,874,261

supplements and we exclude them for simplicity. This amounts to 8104 candidate drugs. We collect 32 clinical trial drugs for Covid-19 to validate our predictions and the drug names can be found in Table XII.

For predicting links among disease and drugs with the relation treatment, we identify the disease entities in our DRKG that are related with the Covid-19 target disease. These disease nodes constitute our expanded target set, since in certain cases as Covid-19, we may have multiple disease nodes representing this novel disease or being related to it such as SARS and MERS diseases. The disease nodes for Covid-19 are in Table XIII. For this experiment, we select 'GNBR::T::Compound:Disease' and 'Hetionet::CtD::Compound:Disease' as the target relations since these represent that a certain drug is used for treating a disease. Next, we recover the pretrained embeddings that are obtained using the complete DRKG and find the 100 drugs with the highest score using Equation (5). Finally, to assess whether our prediction is in par with the drugs used for treatment, we check the overlap among these 100 predicted drugs and the drugs used in clinical trials. Table XIV lists the clinical trial drugs included in the top-100 predicted drugs along with their corresponding score and ranking. Evidently, using the proposed DRKG and plain vanilla KGE model, several of the commonly used drugs in clinical trials are identified. Finally on Table XV we report the top-10 highest ranked drugs for this experiment irrespective of whether these are used for clinical trials or not.

For predicting links among gene and drugs with the inhibit relation, we identify the biological gene entities in our DRKG that are related with the Covid-19 disease. These gene nodes are involved in the related pathways of Covid-19. We obtain 442 Covid-19 related genes from the relations extracted from [18], [9].

In this experiment, we select the inhibit related relation which appears in three datasources as 'GNBR::N::Compound:Gene', 'DRUGBANK::target::Compound:Gene' and 'DGIDB::INHIBITOR::Gene:Compound'. We compare the results obtained by the DRKG with the corresponding results if we trained a KGE on a subset of the DRKG that uses only relations from the three databases GNBR, DRUGBANK and DGIDB individually.

For the results, we recover the pretrained embeddings that are obtained using the KGE model and find the 100 drugs with the highest score using Equation (5) and rank them per target gene. This way we obtain 442 ranked lists of drugs. Finally, to assess whether our prediction is in par with the drugs used for treatment, we check the overlap among these 100 predicted drugs and the drugs used in clinical trials per gene. This procedure is repeated three times per relation and we compare the results of using the DRKG against the constituent databases that include an inhibit relation.

Tables XVI- XVIII list the clinical drugs included in the top-

TABLE XII: Clinical trial drugs for Covid-19.

Drug name	Drug ID
Deferoxamine	DB00746
Piclidenoson	DB05511
Losartan	DB00678
Ibuprofen	DB01050
Favipiravir	DB12466
Ruxolitinib	DB08877
Dexamethasone	DB01234
Thalidomide	DB01041
Tranexamic acid	DB00302
Tocilizumab	DB06273
Sarilumab	DB11767
Tradipitant	DB12580
Angiotensin 1-7	DB11720
Oseltamivir	DB00198
Baricitinib	DB11817
Sargramostim	DB00020
Chloroquine	DB00608
Anakinra	DB00026
Mavrilimumab	DB12534
Azithromycin	DB00207
Tetrandrine	DB14066
Ribavirin	DB00811
Tofacitinib	DB08895
Siltuximab	DB09036
Nivolumab	DB09035
Nitric Oxide	DB00435
Colchicine	DB01394
Remdesivir	DB14761
Hydroxychloroquine	DB01611
Eculizumab	DB01257
Methylprednisolone	DB00959
Bevacizumab	

TABLE XIII: Disease nodes for Covid-19.

Disease node	Disease node
Disease::SARS-CoV2 E	Disease::SARS-CoV2 M
Disease::SARS-CoV2 N	Disease::SARS-CoV2 Spike
Disease::SARS-CoV2 nsp1	Disease::SARS-CoV2 nsp10
Disease::SARS-CoV2 nsp12	Disease::SARS-CoV2 nsp13
Disease::SARS-CoV2 nsp14	Disease::SARS-CoV2 nsp15
Disease::SARS-CoV2 nsp2	Disease::SARS-CoV2 nsp4
Disease::SARS-CoV2 nsp5	Disease::SARS-CoV2 nsp11
Disease::SARS-CoV2 nsp5_C145A	Disease::SARS-CoV2 nsp6
Disease::SARS-CoV2 nsp7	Disease::SARS-CoV2 nsp8
Disease::SARS-CoV2 nsp9	Disease::SARS-CoV2 orf10
Disease::SARS-CoV2 orf3a	Disease::SARS-CoV2 orf3b
Disease::SARS-CoV2 orf6	Disease::SARS-CoV2 orf7a
Disease::SARS-CoV2 orf8	Disease::SARS-CoV2 orf9b
Disease::MESH:D045169	Disease::MESH:D045473
Disease::MESH:D001351	Disease::MESH:D065207
Disease::MESH:D028941	Disease::MESH:D058957
Disease::MESH:D006517	Disease::SARS-CoV2 orf9c

100 predicted drugs across all the genes with their corresponding number of hits for DRKG and the constituent databases. The number of hits shows in how many gene ranked lists the suggested drug appeared in the top-100 ranked drugs. For example if a drug appears in the top-100 ranked drugs for all genes the number of hits is 442. This is the maximum number of possible hits. It can be observed, that several of the commonly used drugs in clinical trials appear high on the predicted list. Furthermore, the number of hits using the DRKG is significant higher comparing to the constituent databases, which corroborates the merits to constructing a comprehensive DRKG.

V. Conclusion

This paper constructed a DRKG from a collection of data sources that can be utilized for general drug repurposing tasks.

TABLE XIV: Drug treats disease scores for Covid-19.

Drug name	Score	Ranking in top-100
Ribavirin	-0.21	0
Dexamethasone	-1.00	4
Colchicine	-1.08	8
Methylprednisolone	-1.16	16
Oseltamivir	-1.39	49
Deferoxamine	-1.51	87

TABLE XV: Top-10 predicted drugs for Covid-19.

Drug name	Ranking	Drug name	Ranking
Ribavirin	1	Isotretinoin	6
Azathioprine	2	Methotrexate	7
Prednisone	3	Bleomycin	8
Streptomycin	4	Colchicine	9
Dexamethasone	5	Budesonide	10

To further facilitate efforts of researchers in repurposing drugs for Covid-19 we also include in DRKG proteins and genes related to Covid-19, as extracted from relevant papers. We train KGE models on the DRKG and obtain embeddings for entities and relation types. We also validate that the DRKG structure and the learned embeddings are of high quality. Finally, we evaluate the DRKG in the drug repurposing task for Covid-19. It is observed that several of the widely used drugs in clinical trials are identified by our method.

Our future research efforts will focus on including more biological entities in the DRKG, enhancing the entities with attributes such as chemical sequence for compounds and developing deep graph learning models that are dedicated for drug repurposing.

References

[1] A. Sertkaya, A. Birkenbach, A. Berlind, and J. Eyraud, "Examination of clinical trial costs and barriers for drug development: report to the assistant secretary of planning and evaluation (aspe)," *Washington, DC: Department of Health and Human Services*, 2014.

[2] J. Avorn *et al.*, "The $2.6 billion pill–methodologic and policy considerations," *N Engl J Med*, vol. 372, no. 20, pp. 1877–1879, 2015.

[3] J. Setoain, M. Franch, M. Martínez, D. Tabas-Madrid, C. O. Sorzano, A. Bakker, E. Gonzalez-Couto, J. Elvira, and A. Pascual-Montano, "Nffinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning," *Nucleic acids research*, vol. 43, no. W1, pp. W193–W199, 2015.

TABLE XVI: Predicted drugs for Covid-19 using the GNBR::N::Compound:Gene relation.

DRKG		GNBR	
Drug name	# hits	Drug name	# hits
Dexamethasone	401	Dexamethasone	262
Thalidomide	336	Chloroquine	233
Chloroquine	258	Thalidomide	177
Deferoxamine	111	Methylprednisolone	99
Colchicine	108	Losartan	59
Methylprednisolone	105	Azithromycin	57
Losartan	92	Tetrandrine	56
Ribavirin	92	Deferoxamine	53
Ruxolitinib	47	Tofacitinib	39
Tofacitinib	33	Colchicine	39
Hydroxychloroquine	14	Ribavirin	28
Piclidenoson	6	Hydroxychloroquine	17
Azithromycin	5	Oseltamivir	8
Oseltamivir	1	Sargramostim	5
Sargramostim	1	Baricitinib	1

TABLE XVII: Predicted drugs for Covid-19 using the DRUGBANK::target::Compound:Gene relation.

DRKG		DRUGBANK	
Drug name	# hits	Drug name	# hits
Ruxolitinib	359	Colchicine	13
Dexamethasone	239	Deferoxamine	11
Thalidomide	95	Dexamethasone	9
Colchicine	68	Thalidomide	8
Baricitinib	37	Ribavirin	7
Losartan	35	Sargramostim	5
Tofacitinib	27	Eculizumab	4
Chloroquine	19	Bevacizumab	4
Bevacizumab	11		
Tocilizumab	10		
Sarilumab	9		
Ribavirin	9		
Hydroxychloroquine	8		
Siltuximab	7		
Methylprednisolone	7		

TABLE XVIII: Predicted drugs for Covid-19 using the DGIDB::INHIBITOR::Gene:Compound relation.

DRKG		DGIDB	
Drug name	# hits	Drug name	# hits
Dexamethasone	369	Thalidomide	20
Ruxolitinib	253	Tocilizumab	14
Thalidomide	203	Ribavirin	12
Hydroxychloroquine	137	Anakinra	11
Chloroquine	92	Ruxolitinib	10
Losartan	64	Bevacizumab	10
Methylprednisolone	32	Methylprednisolone	6
Colchicine	28	Deferoxamine	5
Ribavirin	18	Siltuximab	4
Deferoxamine	11	Dexamethasone	3
Piclidenoson	11	Eculizumab	2
Azithromycin	7		
Bevacizumab	6		

TABLE XIX: Predicted drugs for Covid-19 using the Hetionet::CbG::Compound:Gene relation.

DRKG		Hetionet	
Drug name	# hits	Drug name	# hits
Ruxolitinib	426	Ruxolitinib	438
Dexamethasone	322	Ribavirin	330
Thalidomide	305	Thalidomide	270
Ribavirin	132	Dexamethasone	122
Losartan	116	Colchicine	120
Colchicine	97	Hydroxychloroquine	99
Chloroquine	82	Chloroquine	50
Hydroxychloroquine	66	Oseltamivir	36
Azithromycin	35		
Oseltamivir	22		
Deferoxamine	9		
Methylprednisolone	6		

[4] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[5] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea *et al.*, "A comprehensive map of molecular drug targets," *Nature reviews Drug discovery*, vol. 16, no. 1, p. 19, 2017.

[6] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 5165–5175.

[7] D. Zheng, X. Song, C. Ma, Z. Tan, Z. Ye, J. Dong, H. Xiong, Z. Zhang, and G. Karypis, "Dgl-ke: Training knowledge graph embeddings at scale," 2020.

[8] R. Gramatica, T. Di Matteo, S. Giorgetti, M. Barbiani, D. Bevec, and T. Aste, "Graph theory enables drug repurposing–how a mathematical model can drive the discovery of hidden mechanisms of action," *PloS one*, vol. 9, no. 1, 2014.

[9] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, "Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2," *Cell discovery*, vol. 6, no. 1, pp. 1–18, 2020.

[10] L. Udrescu, L. Sbârcea, A. Topîrceanu, A. Iovanovici, L. Kurunczi, P. Bogdan, and M. Udrescu, "Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[11] W. DS, F. YD, G. AC, L. EJ, M. A, G. JR, S. T, J. D, L. C, S. Z, A. N, I. I, L. Y, M. A, G. N, W. A, C. L, C. R, L. D, P. A, K. C, and W. M, "Drugbank 5.0: a major update to the drugbank database for 2018 nucleic acids res," *PubMed*, 2017.

[12] B. Percha and R. B. Altman, "A global network of biomedical relationships derived from text," *Bioinformatics*, vol. 34, no. 15, pp. 2614–2624, 2018.

[13] H. DS, L. A, H. C, B. L, C. SL, H. D, G. A, K. P, and B. SE, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, 2017.

[14] S. D, G. AL, L. D, J. A, W. S, H.-C. J, S. M, D. NT, M. JH, B. P, J. LJ, and von Mering C., "String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." *PubMed*, 2019.

[15] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro *et al.*, "The mintact project—intact as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, no. D1, pp. D358–D363, 2014.

[16] K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, "DGIdb 3.0: a redesign and expansion of the drug–gene interaction database," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1068–D1073, 11 2017. [Online]. Available: https://doi.org/10.1093/nar/gkx1143

[17] Y. Ge, T. Tian, S. Huang, F. Wan, J. Li, S. Li, H. Yang, L. Hong, N. Wu, E. Yuan, L. Cheng, Y. Lei, H. Shu, X. Feng, Z. Jiang, Y. Chi, X. Guo, L. Cui, L. Xiao, Z. Li, C. Yang, Z. Miao, H. Tang, L. Chen, H. Zeng, D. Zhao, F. Zhu, X. Shen, and J. Zeng, "A data-driven drug repositioning framework discovered a potential therapeutic agent targeting covid-19," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/03/12/2020.03.11.986836

[18] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, M. J. O'meara, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Huttenhain *et al.*, "A sars-cov-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing," *Nature*, 2020.

[19] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro *et al.*, "The mintact project—intact as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, no. D1, pp. D358–D363, 2014.

[20] B. Kotnis and V. Nastase, "Analysis of the impact of negative sampling on link prediction in knowledge graphs," 2017.

[21] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, Dec 2017.

[22] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems 26*, 2013.

[23] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[24] B. Yang, S. W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, May 2015.

[25] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," *CoRR*, vol. abs/1606.06357, 2016.

[26] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11, 2011.

[27] Z. Sun, Z. Deng, J. Nie, and J. Tang, "RotatE: Knowledge graph embedding by relational rotation in complex space," *CoRR*, vol. abs/1902.10197, 2019.

[28] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," *Nucleic acids research*, vol. 41, no. W1, pp. W518–W522, 2013.

[29] [Online]. Available: https://www.ncbi.nlm.nih.gov/mesh/

[30] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 6, 2013, pp. 380–384.
[31] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 19–28, 2016.
[32] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

APPENDIX

This section describes data analysis techniques we performed on the DRKG. The goal is to verify that the constructed DRKG and the learned KG embeddings are of high quality. The code is implemented in DGL-KE [7] that is an open-source package to efficiently compute knowledge graph embeddings. It contains several KGE models, including TransE, DistMult and RotatE, and introduces various optimizations that accelerate training on knowledge graphs using multi-processing, multi-GPU, and distributed parallelism. This way DGL-KE facilitates efficient training on knowledge graphs with millions of nodes and billions of edges.

A. Graph structure analysis

Here the quality of the constructed DRKG is assessed. Different data sources may describe the same triplets and hence the constructed DRKG may have some redundant edges. In this section we verify that triplets from different sources do not have significant overlapping information[8].

First, we assess what is the percentage of common triplets among each pair of edge types in the DRKG. This is important since the same relation-type among the same nodes might be described in two different datasets, and the DRKG may over-represent this relation. If the percentage of common triplets is relatively small then combining these is well justified. Towards this end, we compute the Jaccard similarity coefficient [30] among a pair of relation types $\mathcal{E}^{r_1}, \mathcal{E}^{r_2}$ that is defined as follows

$$j_{r_1,r_2} := \frac{|\mathcal{E}^{r_1} \cap \mathcal{E}^{r_2}|}{|\mathcal{E}^{r_1} \cup \mathcal{E}^{r_2}|} \tag{3}$$

where $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. The Jaccard score is a measure of similarity among the relation types, if j_{r_1,r_2} is close 1 then $\mathcal{E}^{r_1}$ and $\mathcal{E}^{r_2}$ are connecting the node pairs. Table XX reports the 10 most similar edge-pairs based on their Jaccard coefficient. As expected some relation types connect the same nodes since the information from different data sources is related. Nevertheless, the relative small values of the Jaccard coefficient for the most similar edge-type pairs indicates that there is value in including the edge types from all the sources. For example, enzyme catalysis is typically done through binding interactions with protein substrates directly or indirectly. In our DRKG, the relations "STRING::REACTION::Gene:Gene", "STRING::CATALYSIS::Gene:Gene" and "STRING::BINDING::Gene:Gene" have high Jaccard similarity values (0.608, 0.413, .307), that supports our understanding of these activities. The three relations "bioarx::DrugHumGen:Compound:Gene", "Hetionet::CbG::Compound:Gene" and "DRUGBANK::target::Compound:Gene" from different databases correspond to a same relation that compounds bind to their target genes, and in our DRKG, these relations have high pairwise Jaccard similarities. Compounds could affect gene expression after they are metabolized into active substances and moved to certain organs where the genes are usually highly expressed, or compounds and genes compete for same enzymes. In our DRKG, the high Jaccard similarity between "GNBR::E::Compound:Gene" and "GNBR::K::Compound:Gene" corresponds to such a biological process. The relations "Hetionet::AuG::Anatomy:Gene" and "Hetionet::AeG::Anatomy:Gene" both represent that the gene is over-expressed in the Anatomy, except for AuG in post-juvenile adult human samples. In our DRKG, they have high similarity under Jaccard.

The Jaccard similarity may not capture the case that an edge type is contained in another one but the two edge sets have significantly different sizes. Nevertheless, this could happen if two databases describe the same relation, but one has significantly less edges than the other. Next, we examine whether all the edges of a certain type are described by another edge type as well that is $\mathcal{E}^{r_1} \subset \mathcal{E}^{r_2}$. To accomplish this, we compute the overlap coeffcient [31] that is defined as

$$o_{r_1,r_2} := \frac{|\mathcal{E}^{r_1} \cap \mathcal{E}^{r_2}|}{\min(|\mathcal{E}^{r_1}|, |\mathcal{E}^{r_2}|)} \tag{4}$$

The overlap coefficient is close to 1 if all the edges in one edge set are also present in the other set. Table XXI reports the 10 most similar edge-type pairs based on their overlap coefficient. We observe that for certain edge-type pairs there exists significant overlap. Nevertheless, for the pair GNBR::E and GNBR::E+ the total overlap is expected since the first relation signifies that a drug affects the expression of a gene whereas the second indicates that a drug increases the expression of a gene and it is contained in the first one. The high overlap coefficient between "Hetionet::AuG::Anatomy:Gene" and "Hetionet::AeG::Anatomy:Gene" is also expected, because AuG is a special case of AeG. Similarly, "STRING::CATALYSIS::Gene:Gene" and "STRING::BINDING::Gene:Gene" can be considered as special cases of "STRING::REACTION::Gene:Gene".

B. Knowledge graph embedding analysis

We can use the embeddings of the knowledge graph to analyze how relations and entities are clustered and investigate whether certain similarities in the embedding space are consistent with their biological meanings, e.g., "GNBR::E:Compound:Gene" and "GNBR::E+:Compound:Gene" can be similar as they both represent the expression relationship between a Compound and a Gene, but "GNBR::E+" represents the positive expression. Further, we can generate the confidence of existing edges by calculating the prediction scores using the embeddings. From the perspective of the knowledge graph

[8]The corresponding notebook in GitHub https://github.com/gnn4dr/DRKG/blob/master/raw_graph_analysis

TABLE XX: Most similar edge-type pairs based on the Jaccard similarity score

Relation type 1	Relation type 1	Jaccard score
STRING::REACTION::Gene:Gene	STRING::CATALYSIS::Gene:Gene	0.608
STRING::REACTION::Gene:Gene	STRING::BINDING::Gene:Gene	0.413
STRING::CATALYSIS::Gene:Gene	STRING::BINDING::Gene:Gene	0.307
bioarx::DrugHumGen:Compound:Gene	Hetionet::CbG::Compound:Gene	0.272
bioarx::DrugHumGen:Compound:Gene	DRUGBANK::target::Compound:Gene	0.261
STRING::INHIBITION::Gene:Gene	STRING::PTMOD::Gene:Gene	0.240
GNBR::E::Compound:Gene	GNBR::K::Compound:Gene	0.215
DRUGBANK::enzyme::Compound:Gene	Hetionet::CbG::Compound:Gene	0.188
DRUGBANK::target::Compound:Gene	Hetionet::CbG::Compound:Gene	0.171
Hetionet::AuG::Anatomy:Gene	Hetionet::AeG::Anatomy:Gene	0.157

TABLE XXI: Most similar edge-type pairs based on the overlap coefficient

Relation type 1	Relation type 1	Overlap coefficient
GNBR::E::Compound:Gene	GNBR::E+::Compound:Gene	1.0
Hetionet::AuG::Anatomy:Gene	Hetionet::AeG::Anatomy:Gene	0.866
STRING::REACTION::Gene:Gene	STRING::CATALYSIS::Gene:Gene	0.818
STRING::ACTIVATION::Gene:Gene	STRING::EXPRESSION::Gene:Gene	0.680
bioarx::DrugHumGen:Compound:Gene	Hetionet::CbG::Compound:Gene	0.667
STRING::REACTION::Gene:Gene	STRING::BINDING::Gene:Gene	0.662
GNBR::E::Compound:Gene	GNBR::K::Compound:Gene	0.643
STRING::INHIBITION::Gene:Gene	STRING::PTMOD::Gene:Gene	0.565
bioarx::DrugHumGen:Compound:Gene	DRUGBANK::enzyme::Compound:Gene	0.559
INTACT::PHYSICAL ASSOCIATION::Gene:Gene	INTACT::COLOCALIZATION::Gene:Gene	0.546

embedding algorithm, e.g., TransE, if the score between two entities h and t under relation r is far away from 0, then the model cannot correctly score an edge on which it trained on. This means that the edge does not fit the underlying model and is often an indication that the corresponding edge may be incorrect.

Here, we analyze DRKG by learning a TransE KGE model that utilizes the ℓ_2 distance (Section II-B). For the analysis in this section, we split the edge triplets in training, validation and test sets as follows 90%, 5%, and 5% and train the KGE model. Finally, we obtain the entity and relation embeddings for the DRKG. Next, we apply the following methodologies to validate the quality of the learned embeddings [9].

1) Entity embedding similarity: We use t-SNE [32] to map entity embeddings to a 2D space. Figure 2 illustrates how the entity embeddings are placed in the 2D space. Different colors denote different entity types. We observe that entities from the same type are grouped together as we expected. Table XXII shows the average pair-wise cosine similarity of certain entity type within the same category and cross different categories. It can be seen that entities are more similar to each other within the same category that entities from different categories. Figure 3 shows the detailed distribution of pair-wise cosine similarity between different entities based on their embeddings. In the figure, the counts are normalized that the area under the histogram is sum to 1. It can be seen that most of the entities have low cosine similarity according to their embeddings. They are distinguishable in the current embedding space.

2) Relation type embedding similarity: We use t-SNE to map relation embeddings to a 2D space and plot it in Figure 4. It can be seen that relations are widely spread across the 2D

[9]The corresponding code in GitHub https://github.com/gnn4dr/DRKG/tree/master/embedding_analysis

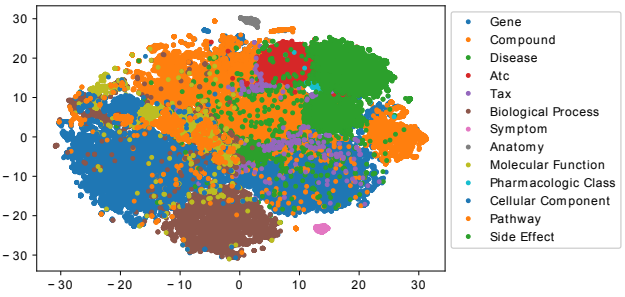

Fig. 2: Distribution of entity embeddings in 2D euclidean space

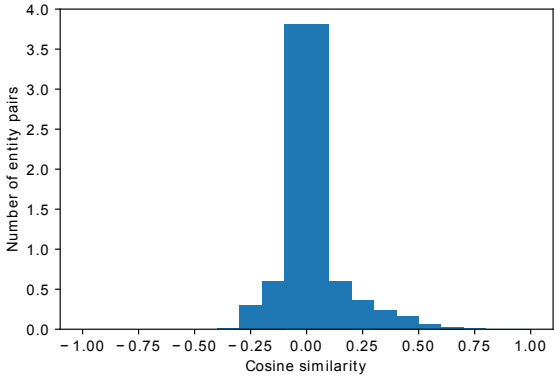

Fig. 3: Histogram of cosine similarity between entities.

TABLE XXII: Average pair-wise cosine similarity within the same category and cross different categories.

Entity Type	Avg Cosine Score	
	Within same category	Cross different categories
Gene	0.024	0.0086
Compound	0.054	0.0099
Disease	0.19	0.015
Atc	0.25	0.023
Tax	0.22	0.062
Biological Process	0.14	0.0078
Symptom	0.39	0.056
Anatomy	0.32	0.018
Molecular Function	0.21	0.016
Pharmacologic Class	0.32	0.062
Cellular Component	0.22	0.020
Pathway	0.22	0.015
Side Effect	0.20	0.016

space and relations from the same dataset do not cluster which is expected as most of relations have different meanings even from the same data source. Only a small part of relations from GNBR dataset are clustered together. There are two clustered red dots (relations from GNBR) in the Figure. Table XXIII shows all relation-type pairs with the Cosine similarity larger than 0.9. It can be seen that "GNBR::E::Compound:Gene", "GNBR::K::Compound:Gene", "GNBR::E+::Compound:Gene", "GNBR::N::Compound:Gene" and "GNBR::E-::Compound:Gene" are highly similar to each other. Some of these pairs also appear in the most similar edge-type pairs analysis on the Jaccard similarity and overlap similarity. Drugs can affect (E) gene expression (either increase (E+) or decrease (E-)) through their pharmacokinetics (K). Drugs can also inhibit genes to affect gene expressions and thus the high cosine similarity between "GNBR::N:: Compound:Gene" and the above relations are justified. "GNBR::L::Gene:Disease", "GNBR::G::Gene:Disease", "GNBR::J::Gene:Disease", "GNBR::Md::Gene:Disease", "GNBR::Te::Gene:Disease" and "GNBR::X::Gene:Disease" are highly similar to each other. The similarity between these relations could be due to that the genes are involved in the pathogenesis of the disease and therefore they can be treated as the targets or they have therapeutic effect for the disease. From the table, we can see that "bioarx::DrugHumGen:Compound:Gene" and "DRUGBANK::target::Compound:Gene" are also similar to each other, while they have similar meaning of treatment.

Figure 5 shows the detailed distribution of pair-wise cosine similarity among different relation types based on their embeddings. The counts are normalized that the area under the histogram is sum to 1. It can be seen that most of the relation embeddings have small cosine similarity. Only 0.53% of relation pairs have similarity larger than 0.9 with the maximun of 0.986 (between GNBR::E::Compound:Gene and GNBR::K::Compound:Gene).

3) Edge prediction analysis: Towards validating each triplet of the DRKG we evaluate how well it fits the scoring function of KGE model. In order to avoid the possible bias of over-fitting the triplets in the training set, we split the whole DRKG into 10 equal folds and train 10 KGE models by picking each fold as the test set and the rest other nine folds are the training set. Following this, the score for each triplet is calculated while

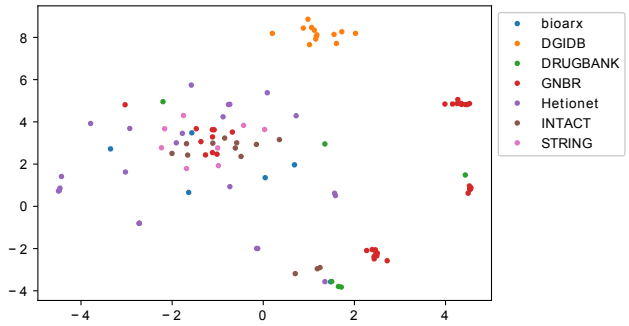

Fig. 4: Distribution of relation embeddings in 2D euclidean space

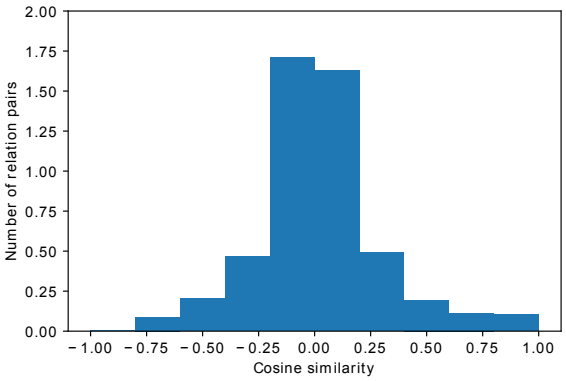

Fig. 5: Histogram of cosine similarity between relations

TABLE XXIII: Relation-type pairs with the Cosine similarity larger than 0.9 based on their embeddings.

Relation type 1	Relation type 2	Cosine score
GNBR::E::Compound:Gene	GNBR::K::Compound:Gene	0.986
GNBR::E::Compound:Gene	GNBR::E+::Compound:Gene	0.983
GNBR::N::Compound:Gene	GNBR::E-::Compound:Gene	0.970
GNBR::E::Compound:Gene	GNBR::E-::Compound:Gene	0.965
GNBR::K::Compound:Gene	GNBR::E+::Compound:Gene	0.956
GNBR::E+::Compound:Gene	GNBR::E-::Compound:Gene	0.950
GNBR::L::Gene:Disease	GNBR::G::Gene:Disease	0.942
GNBR::K::Compound:Gene	GNBR::E-::Compound:Gene	0.941
GNBR::J::Gene:Disease	GNBR::Md::Gene:Disease	0.932
GNBR::J::Gene:Disease	GNBR::Te::Gene:Disease	0.932
GNBR::L::Gene:Disease	GNBR::X::Gene:Disease	0.932
GNBR::E::Compound:Gene	GNBR::N::Compound:Gene	0.926
GNBR::G::Gene:Disease	GNBR::X::Gene:Disease	0.924
GNBR::N::Compound:Gene	GNBR::A-::Compound:Gene	0.916
GNBR::Te::Gene:Disease	GNBR::Md::Gene:Disease	0.915
GNBR::N::Compound:Gene	GNBR::E+::Compound:Gene	0.914
GNBR::B::Compound:Gene	GNBR::O::Compound:Gene	0.913
bioarx::DrugHumGen:Compound:Gene	DRUGBANK::target::Compound:Gene	0.911
GNBR::E::Compound:Gene	GNBR::A+::Compound:Gene	0.910
GNBR::E::Compound:Gene	GNBR::O::Compound:Gene	0.909
GNBR::Md::Gene:Disease	GNBR::X::Gene:Disease	0.909
GNBR::J::Gene:Disease	GNBR::X::Gene:Disease	0.908
GNBR::Md::Gene:Disease	GNBR::G::Gene:Disease	0.907
GNBR::E::Compound:Gene	GNBR::B::Compound:Gene	0.906
GNBR::A+::Compound:Gene	GNBR::E+::Compound:Gene	0.905
GNBR::L::Gene:Disease	GNBR::Ud::Gene:Disease	0.904
GNBR::N::Compound:Gene	GNBR::K::Compound:Gene	0.903
Hetionet::AdG::Anatomy:Gene	Hetionet::AuG::Anatomy:Gene	0.902
GNBR::E::Compound:Gene	GNBR::Z::Compound:Gene	0.901
GNBR::J::Gene:Disease	GNBR::G::Gene:Disease	0.901

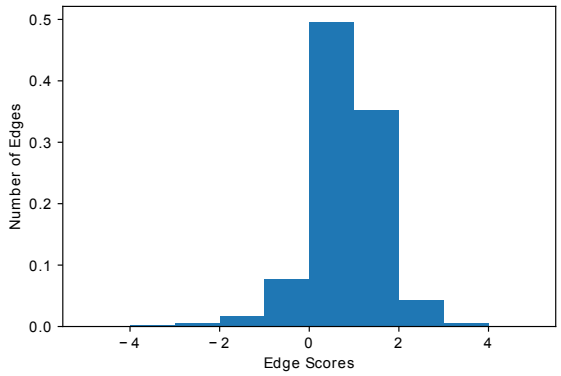

Fig. 6: Distribution of edge scores

this triplet was in the test set. Consider the triplet (h, r, t) of the DRKG and the associated score as

$$\text{score} = \gamma - ||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2 \tag{5}$$

where γ is a constant we used in training of the TransE model that is set to 12.0. For each triplet (h, r, t), the closer its $score$ is to 0, the more confident it was that the head h and tail t entities are connected under relation r. The distribution of the edge scores is depicted in Figure 6. The counts are normalized that the area under the histogram is sum to 1. From the figure, we can see that around 57.28% of edges are scored as $|score| < 1$ and 94.24% of edges are scored as $|score| < 2$. The average of $|score|$ is 0.987. We also randomly shuffle t in (h, r, t) for all triplets to construct the negative triplets (h, r, t') and calculate the $|score'|$ using the same formula. The average $|score'|$ is 2.545.

4) Link type recommendation similarity: Finally, we also evaluate how similar are the predicted links as given by the KGE model among different relation types. This task examines the similarity across relation types for the link prediction task. For a set of nodes, we measure the overlap of predicted neighbors under different relation types as given by the TransE model.

For seed node we find the top 10 neighbors under relation r_j with the highest link prediction score. Next, we repeat the same prediction for relation $r_{j'}$ and calculate the Jaccard similarity coefficient among the predicted sets of top 10 neighbors for r_j and $r_{j'}$. We repeat this process for 100 random selected seed nodes and report the average similarity score for all edge-type pairs. The most similar edge types in this context are the ones relating Genes. This may be attributed to the fact that the genes are the most represented entitites in the DRKG.

TABLE XXIV: Top 10 most similar edge-type pairs based on their link recommendation similarity.

Relation type 1	Relation type 1	Jaccard score
STRING::CATALYSIS::Gene:Gene	GNBR::Q::Gene:Gene	0.864
STRING::OTHER::Gene:Gene	STRING::ACTIVATION::Gene:Gene	0.892
STRING::OTHER::Gene:Gene	GNBR::Rg::Gene:Gene	0.888
STRING::REACTION::Gene:Gene	STRING::OTHER::Gene:Gene	0.882
STRING::OTHER::Gene:Gene	STRING::BINDING::Gene:Gene	0.880
STRING::OTHER::Gene:Gene	GNBR::Q::Gene:Gene	0.877
STRING::OTHER::Gene:Gene	bioarx::HumGenHumGen:Gene:Gene	0.875
GNBR::Q::Gene:Gene	GNBR::Rg::Gene:Gene	0.874
bioarx::HumGenHumGen:Gene:Gene	STRING::INHIBITION::Gene:Gene	0.874
STRING::ACTIVATION::Gene:Gene	STRING::BINDING::Gene:Gene	0.872