BioKG: A Knowledge Graph for Relational Learning On Biological Data

Brian Walsh
Data Science Institute, NUI Galway
Insight Centre for Data analytics
Galway, Ireland
brian.walsh@insight-centre.org

Sameh K. Mohamed
Data Science Institute, NUI Galway
Insight Centre for Data analytics
Galway, Ireland
sameh.kamal@insight-centre.org

Vít Nováček
Data Science Institute, NUI Galway
Insight Centre for Data analytics
Galway, Ireland
vit.novacek@insight-centre.org

ABSTRACT

Knowledge graphs became a popular means for modelling complex biological systems where they model the interactions between biological entities and their effects on the biological system. They also provide support for relational learning models which are known to provide highly scalable and accurate predictions of associations between biological entities. Despite the success of the combination of biological knowledge graph and relation learning models in biological predictive tasks, there is a lack of unified biological knowledge graph resources. This forced all current efforts and studies for applying a relational learning model on biological data to compile and build biological knowledge graphs from open biological databases. This process is often performed inconsistently across such efforts, especially in terms of choosing the original resources, aligning identifiers of the different databases and assessing the quality of included data. To make relational learning on biomedical data more standardised and reproducible, we propose a new biological knowledge graph which provides a compilation of curated relational data from open biological databases in a unified format with common, interlinked identifiers. We also provide a new module for mapping identifiers and labels from different databases which can be used to align our knowledge graph with biological data from other heterogeneous sources. Finally, to illustrate practical relevance of our work, we provide a set of benchmarks based on the presented data that can be used to train and assess the relational learning models in various tasks related to pathway and drug discovery.

CCS CONCEPTS

• **Applied computing → Biological networks**; **Bioinformatics**; • **Information systems → Extraction, transformation and loading**.

ACM Reference Format:
Brian Walsh, Sameh K. Mohamed, and Vít Nováček. 2020. BioKG: A Knowledge Graph for Relational Learning On Biological Data. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412776

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412776

1 INTRODUCTION

Knowledge graphs (KGs) are a popular means for modelling relational data in many systems and applications. They have currently become the backbone of many semantic web search engines and question answering systems in both academic and industrial settings [27]. This encouraged the development of many public knowledge graphs which model information from different domains such as general human knowledge [18], lexical information [20] and other domains. These knowledge graphs provide support for predictive models in different tasks and facilitate information retrieval on the original linked data.

In recent years, knowledge graphs have also become a favourable choice for modelling complex biological systems where they were used in different predictive tasks such as predicting drug protein targets [26], predicting polypharmacy side-effects [40] and the prediction of cellular functions of proteins at the tissue level [22]. In each of these tasks, KGs were used to model biological networks, and then relational learning models were used to provide new predictions. Despite the effectiveness of such approaches [24], there is a lack of open biological knowledge graphs to support them. Furthermore, current approaches rely on building customized knowledge graph by parsing and transforming open biological databases [24, 26, 29].

The effectiveness of knowledge graphs and the popularity of the RDF framework form modelling linked data have encouraged many open biological database to provide their contents in RDF format. For example, the UNIPROT [6, 7], Reactome [8] and Gene Ontology [5] databases provide an RDF version of their content which preserves both the interlinks and metadata of their contained biological entities. However, these RDF graphs only focus on a limited set of biological entity types covered by the corresponding original database. Moreover, they do not share any common entity coding system, which makes it hard to use them in concert. There is also a large body of biological data that has no RDF counterpart at all. This encouraged efforts such as the Bio2RDF project [3] to build and provide a network of linked biological data by transforming open biological databases to RDF graphs.

The Bio2RDF project consists of a set of web parsers for open biological data which consume, process and convert these database to RDF graphs. Despite the high coverage of its generated RDF graphs, they are not commonly used in the different predictive biological task by relational learning models [24]. One of the main reasons is the large volume of metadata information stored in these graphs which often decreases the predictive accuracy of relational models.

This is due to the models' tendency to over-represent the clearly-interpretable and uniform metadata links and under-represent the more subtle actual biological relationships.

The current studies of the applications of relational leaning models in biological settings commonly involve building customized biological graphs from open biological databases [24, 26, 29]. This process involves repeated procedures such as parsing the different database sources into intermediate formats then merging these format into knowledge graphs. It also involves mapping entity identifiers to a unified ID system as biological databases commonly employ different identifier systems. Such steps are frequently associated with many rather arbitrary decisions that complicate reproducibility and meaningful comparisons between the corresponding models. To address this problem, we provide a new open biological knowledge graph, BioKG, and tools for its transparent creation, updates and extensions. Contrary to existing resources like Bio2RDF, BioKG combines information from different open biological databases in a simple graph format which focuses on biological relationships while preserving basic important ontological information, and thus it allows for straightforward development and comparative evaluation of relation learning models.

We discuss related works in Section 2 and we discuss our main contributions in Sections 3,4 and 5 as follows:

(1) In Section 3, we propose a biological knowledge graph (BioKG) compiled from open source databases to support relational learning models in predictive tasks on biological data.
(2) In Section 4, we propose a software module (BioDBLinker) which provides name-id lookup and mapping of different id systems for biological entities.
(3) In Section 5, we propose a set of five benchmarking datasets for assessing the predictive accuracy of relational learning models in different tasks related to drug–protein, drug–drug and protein–protein interactions.

In Section 6, we discuss potential applications and possible issues related to the development and use of BioKG knowledge graph, and our intentions for future extensions of this work. Finally, in Section 7 we discuss our conclusions.

2 RELATED WORKS

In this section, we discuss studies and resources related to our newly proposed knowledge graphs.

2.1 Open Biological Databases

Open biological databases support research in both clinical and computational biology. They contain different types of structured and unstructured data related to different biological phenomenons. In this work, we focus on databases that provide biological data which is related to molecular and pharmacological activities, *e.g.* protein interactions, drug protein targets, etc.

In Table 1 we provide detailed statistics on a selected set of popular biological databases which are commonly used to train relational learning models in bioinformatics settings. We provide a comparison between these databases in terms of their data formats, specialities and the covered biological entities. In terms of the data format, the table shows that almost all the databases contain structured data while a subset of these databases contains both structured and unstructured data. These unstructured data are usually comments and annotations of describing pieces of the structured data as in the protein–protein interactions related comments in the UniProt database [6].

Table 1: A comparison between popular biological databases in terms of the coverage of different types of biological entities. The abbreviation S represent structured data, U represents unstructured data, DR represents drugs, PR represents proteins, GO represents gene ontology,PA represents pathways, GD represents genetic disorders, CL represents cell-lines and *CH* denotes chemicals.

	Properties		Entity coverage							
Database Name	Format	Speciality	Proteins	Drugs	Indications	Diseases	Gene Ontology	Expressions	Pathways	In BioKG ?
UniProt [7]	S/U	PR	✓	✓	✗	✓	✓	✓	✓	✓
REACTOME [8]	S	PA	✓	✗	✗	✗	✓	✗	✓	✓
KEGG [14]	S	PA	✓	✓	✗	✓	✗	✗	✓	✓
DrugBank [15]	S/U	DR	✓	✓	✗	✗	✗	✗	✓	✓
GO [5]	S	GO	✓	✗	✗	✗	✓	✗	✓	✓
CTD [19]	S/U	CH	✓	✓	✗	✗	✓	✗	✓	✓
SIDER [16]	S	DR	✗	✓	✓	✗	✗	✗	✗	✓
HPA [36]	S/U	PR	✓	✗	✗	✗	✓	✓	✗	✓
STRING [33]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
BIOGRID [32]	S	PR	✓	✗	✗	✗	✗	✗	✗	✗
IntAct [30]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
InterPro [21]	S	PR	✓	✗	✗	✗	✗	✗	✗	✓
PharmaGKB [10]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
TTD [17]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Supertarget [9]	S	DR	✓	✓	✗	✗	✗	✗	✗	✗
Cellosaurus [2]	S/U	CL	✗	✗	✗	✗	✗	✓	✗	✓
MESH [1]	S/U	CL	✗	✗	✗	✓	✗	✗	✗	✓
OMIM [1]	S	GD	✓	✗	✗	✗	✗	✗	✗	✓

While the majority of the reviewed databases specialize in data focused on proteins, the UniProt database is the most popular source for protein related data as it has the highest coverage of expert-curated protein annotations [7]. The UniProt database consists of two parts SwissProt (expert-curated) and TrEMBL (lower confidence annotation). It also use a protein id naming system known as "UniProt Accessions". Other databases such as KEGG and CTD databases use "Gene Id Numbers" as ids for proteins where they define unique proteins based on their source genes. The HPA [36] and STRING [37] databases use yet another a different gene-based id system for proteins. Although all these databases have intersection between their reported protein annotations, they do not have a one-to-one mapping between their ids, therefore merging their annotations can be complicated. Similarly, databases that provide data on drugs such as the DrugBank [15], SIDER [16], CTD [19] and KEGG [14] databases also use different id systems for drugs which often does not have a one-to-one mapping for some of their common entities.

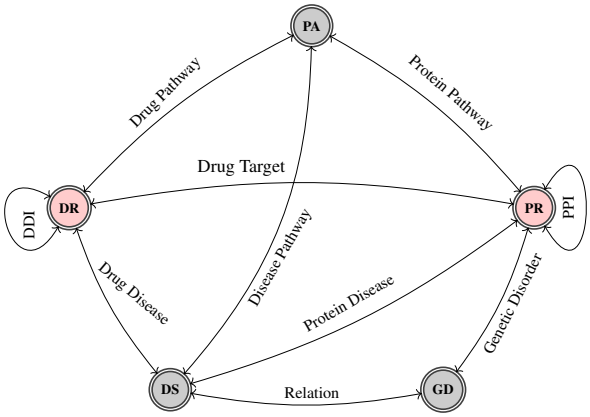

Figure 1: The schema of BioKG main biological entities and their connections. Abbreviations in this illustration are the same as in Table 1.

While resources like Bio2RDF and RDF versions of open biological databases aim to resolve these problems, their main objective was to integrate the biological databases with semantic web technologies. This led to the development of biological RDF graphs that have complex ontological information. These graphs, however, still have issues when used for training relational learning models due to their use of different id systems, variable quality and dense ontological data that is largely irrelevant to training predictive models, which the presented work attempts to remedy.

2.2 Relational Learning in Bioinformatics

In recent years, relational learning models (RLMs) became a popular method in many bioinformatics predictive tasks where they outperform other state-of-the-art approaches in various tasks [24]. They use knowledge graphs to model complex biological systems and they then learn feature representations of entities and relationships to provide accurate and scalable predictions. For example, Zitnik et. al. [40] have modelled drug–drug interactions and their associated side-effects as a knowledge graph and they applied a graph convolutional network model to predict the polypharmacy side-effects of drugs. This work has also shown that such an approach outperforms previous state-of-the-art methods in terms of the predictive accuracy. Furthermore, other studies have shown that modelling biological data with knowledge graphs and using knowledge graph embedding models *e.g.* TransE [4], ComplEx [35], TriVec [25], etc, to predict biological relationships is effective in tasks such as drug target interaction prediction [26] and tissue-specific protein function prediction [22].

In Fig. 1, we provide an basic graph schema of the mainly investigated relationships between biological entities and their related information at the molecular and pharmacological level. These relationships include the previously mentioned drug protein targets and drug side-effects along with other relationships such as disease associated genes, protein associated pathway, etc. In the context of predicting each of these relationships, relational learning models usually build a knowledge graph centred around the two end of the relationships where it usually include other relationships in graph schema. For example, in the prediction of relationships between drugs and proteins, RLMs are usually training on a knowledge graph that has information about drugs such as the ATC class, chemical structure groups, etc [26]. It also includes information about proteins such as protein–protein interactions, protein related pathway, gene ontology annotations, etc. All this information has to be fused in one knowledge graph centred around drug–protein relationships to enable RLMs to efficiently model and predict new drug–protein interactions. However, there is no existing, publicly available data set that would enable this with sufficient coverage, which is another gap the presented work covers.

3 BIOKG KNOWLEDGE GRAPH

In this section, we discuss the contents of BioKG knowledge graph and the details of the pipeline to build these contents. We also provide statistics of its covered entities and relations.

3.1 Processing the Original Data Sources

The BioKG knowledge graph is built through a two-phase procedure as shown in Fig. 2. This procedure includes parsing open biological database to intermediate structured formats, then integrating these formats to obtain the BioKG contents. In the following, we discuss this procedure where we describe materials and techniques used in each phase.

3.1.1 Parsing Sources. The BioKG consists of a set open biological databases (identified in Table 1) which contain different types of biological data. The criteria used for choosing these databases depend on three factors: entity coverage, data quality, and integration with other databases. In terms of coverage, the UniProt and KEGG databases are the most popular sources for protein data as they have the highest coverage of proteins/genes especially in humans. This can be shown by the wide adoption of UniProt ids as baseline references for proteins in multiple open biological databases such as Reactome, Phosphosite, etc. We also selected a wide range database to cover the different types of biological entities such as proteins, drugs, pathways, expressions and other entities as illustrated in Fig. 1. In terms of quality, we focus on databases that have expert-curated data and we exclude data generated by inference technique to ensure high quality data. We, therefore, only include the SwissProt part of UniProt which contain only reviewed protein entries and we exclude the inferred data parts of databases. We also selected databases that have better integration to ensure the connectivity of the different parts of BioKG knowledge graph.

For each of the selected databases, we parse the database contents into a structured tabular format. This format allows for more dynamic representations of the included data which is often modelled in higher dimensionality and/or more complex formats than knowledge graph triplets. For example, the protein tissue expression data parsed from the human protein atlas (HPA) is associated with different expression levels. This data is stored in an intermediate format in the form (<protein>, expressed_in, <tissue>, <expression_level>) where protein, tissue and tissue levels are variable depending on the different entries. This format is incompatible with knowledge graph triplets, so, in the final phase it is converted to triplets by excluding the expression level column data. We have developed automated parsers for each of the included databases which consume

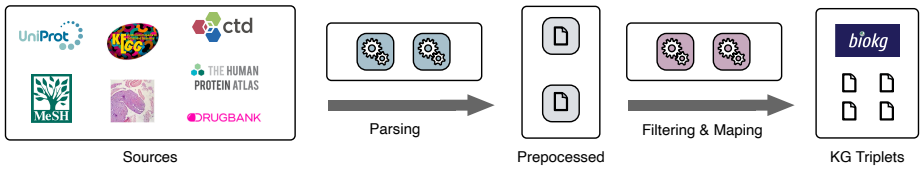

Figure 2: An Illustration of the processing pipeline to build the BioKG data.

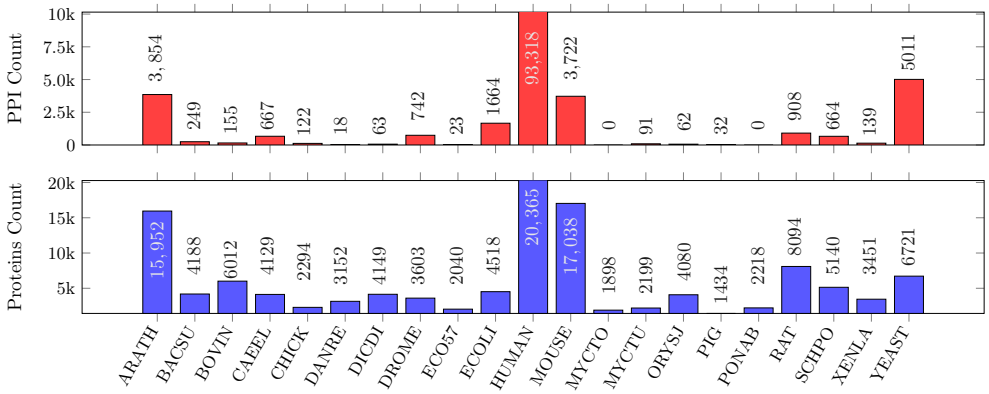

Figure 3: Statistics of proteins in the BioKG and their protein-protein interactions (PPIs) categorized by their species. The PPIs for each species are only considered where both proteins are from the same species.

and parse the contents of the database and output intermediate data files.

3.1.2 Compiling BioKG Contents. The BioKG knowledge graph contents is compiled from the intermediate formats generated by the biological databases' parsers. The compilation process mainly involves to three procedures: id , filtering and building triplets. The mapping of ids is the process of converting ids of entities of the same type to the same id system. We execute this process using the BioDBLinker module that discuss in details in Section 4. For example, all drugs in the BioKG use the DrugBank ids while all proteins use the UniProt ids. This ensures the connectivity of nodes coming from heterogeneous source in the BioKG knowledge graph.

Data filtering process in the building of BioKG contents is aimed to satisfy to objectives: high quality data and focus on drug discovery biological applications. The high quality data is obtained by filtering intermediate data formats to only include the expert curated parts of the included databases. The data is also filtered to only include biological entities and phenomenons related to drugs and their related protein targets' activities such as illustrated in Fig. 1.

3.2 Structure of the Knowledge Graph

The BioKG knowledge graph compiles biological data from different sources in a graph format with focus on data on proteins and chemical drugs. The contents of BioKG knowledge graph can be categorized into three categories: links, properties and metadata. Links represent the connections between the different biological entities, while properties represent the annotations associated to entities. Each biological entity type has its own set of links and properties that describe its activities in biological systems. Fig. 1 illustrates the main biological entities included in the BioKG and the relationships between them.

In the following, we discuss the contents included under each of the three categories (links, properties and metadata) in the BioKG knowledge graph.

3.2.1 BioKG Links. The links part of the BioKG data is the core part of BioKG which models the relationships between the biological entities as illustrated in Fig. 1. The number of instances of each of the relationships in BioKG is illustrated in Fig 4 and they are described as follows:

- **protein–protein interactions (PPIs)**. BioKG contains 113451 PPIs of 21 selected species as illustrated in Fig 3. These interactions are extracted from the respective protein entries from SwissProt and IntAct databases.
- **drug–protein interactions (DPIs)**. BioKG contains different types of DPIs such as drug–protein targets, carriers, enzymes and transporters. The DPIs in BioKG are focused on human proteins and they are extracted from the DrugBank and KEGG databases exclusively, and they are considered in two forms: unified relation and separate relation for each type (exclusively from DrugBank). The DPI relations (27781 instances) in the BioKG links are the union of all the separate instances of the drug-carries, drug-transporters, drug-targets, drug-enzymes relationships combined with other drug-protein interactions extracted from the KEGG database.

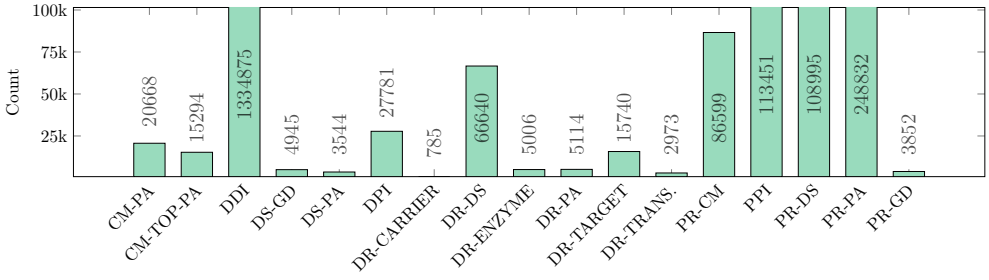

Figure 4: Statistics of the links part of the BioKG knowledge graph. DPI and PPI refer to drug-protein interactions and protein-protein interactions respectively, CM denotes complexes and other abbreviations follow definitions in Table 1.

- **drug–drug interactions (DDIs)**. drug–drug interactions represent the interactions between drugs which often happens because of the prescription of drug combinations *i.e.* polypharmacy. The data of DDIs in BioKG is collected exclusively from the DrugBank database where there are 1334875 instances of DDIs relationships in BioKG links.
- **protein relations to genetic disorders**. Proteins are the products of genes, and the protein–genetic disorder relations capture the associations between proteins and the disorders of their origin genes. There are 3852 associations between proteins and genetic disorders in BioKG which are extracted from the SwissProt database and their links to the OMIM genetic disorder database.
- **protein relations to diseases**. BioKG contains 108995 instances of relations between diseases and their associated proteins which is extracted from the KEGG database. All disease ids are set the Medical Subject Headings (MeSH) format and all protein ids are set to UniProt format to comply with the rest of BioKG.
- **protein–pathway associations**. The involvement of proteins in specific pathways is captured in protein–pathway relation in BioKG where there are 248832 instance of protein–pathway associations. These instances are collected from the UniProt, KEGG, Reactome and DrugBank databases.
- **disease–genetic disorder associations**. There are 4945 disease–genetic disorder relationships in BioKG which are extracted from the KEGG and MESH databases.
- **disease–pathway associations**. Associations between diseases and specific protein interaction chains (pathways) which describe the disease or describe another biological process related to it. BioKG contains 3544 disease–pathway associations exclusively extracted from the KEGG database.
- **drug–pathway associations**. There are 5114 associations between drugs and pathways in the BioKG which are exclusively extracted from the DrugBank database.
- **complex related associations**. Complexes are composites of proteins which represent a set of physically interacting proteins. BioKG contain data on complexes and their member proteins and associated pathways which is extracted exclusively from the Reactome database.

3.2.2 Properties. The properties part of BioKG contains the associations between the previously discussed biological entities and their different attributes as illustrated in Fig. 5. For example, protein attributes include their association to gene ontology entries and their sequence annotations. On the other hand, properties of drugs in BioKG are their associated side-effects, indications and ATC classification codes.

BioKG also contains other types of properties for pathways, disease and genetic disorders where these properties are often a categorization of these entity types into groups based on different type-specific criteria.

3.2.3 Metadata. The metadata part contains data about biological entities names, types, synonyms, etc. This part of the data is not meant to be used in the training of relational learning models, and it does not contain any attributes or important associations for biological entities. Our objective, however, in this part is to maximize the richness of metadata on each of the included biological entities to facilitate analysing their related insights and to allow for tracking history of changes of ids and synonyms of biological entities' databases entries.

4 BIODBLINKER

In this section, we discuss the motivation behind the BioDBLinker library and its implementation as well as its usage.

4.1 Motivation

Many biological knowledge bases contain overlapping or partially overlapping data. In order to extract the unique set of relations between a given set of entities it is therefore necessary to remove this duplication. This process is made more difficult by the fact that some data sources use different identifiers for the same entity. To overcome this issue it is necessary to parse entity ids into a unified id system. As we have found that this is a recurring step required when generating biological knowledge graphs we undertook to develop a library which could be reused for this purpose which we believe would be useful to others in the community when building biological knowledge graphs.

Current methods for mapping biological entities include online services which require manual data entry, or mapping files which require writing scripts to process mapping inputs. Our approach main objective is to tackle these issues by providing offline services for the mappings which can be used automatically/programmatically in various application.

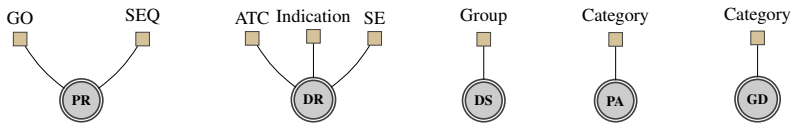

Figure 5: Illustration of BioKG main biological entities and their associated attributes in the properties part of the BioKG data.

4.2 Mapping Generation

BioDBLinker library provides a mapping generator class to generate the mappings required for the linker classes. The mapping generator parses the data in a number of formats from a number of different biological data sources to extract mappings from these data sources to other biological data sources in a unified manner. The source files for the mappings are downloaded from their respective biological data source at runtime allowing the mappings to be updated as new versions of the source files are released as required. For example, we use the UniProt mapping file to building mappings from/to UniProt ids and other 21 identifier systems. Similarly, we build mappings from the mapping files provided by the KEGG database to map its ids with other related databases.

4.3 Usage

BioDBLinker provides 3 main functions: (1) linking entity ids and names, (2) linking from entity ids to ids in other data sources and (3) retrieving ids for a given entity type in a specific database. The ids in a linker class can be accessed as properties of the class. When converting to names or other data sources a list of ids are passed to the function and a list of lists of names or mapped ids is returned, this allows for the case where an entity has multiple names or can be mapped to multiple entities in the target data sources.

4.4 Coverage

The BioDBLinker module covers 5 main data sources in relation to the BioKG knowledge graph, UniProt/SwissProt, Drugbank, KEGG, SIDER and Human Protein Atlas (HPA). Each of these data sources map to a number of other biological data sources, the BioDBLinker covers the mappings from/to each of the main mentioned databases to their respective associated databases.

5 BENCHMARKS

In this section, we discuss a set of four benchmarks that we provide with the BioKG. These benchmarks are focused on drug target discovery and drug-drug interactions related effects. In the following, we discuss the properties of each of our proposed benchmark datasets.

DDI-MINERAL Benchmark. The DDI-MINERAL benchmark consists of a set of drug-drug interactions and their associations to abnormalities of minerals levels in the human body where we focus on four elements: potassium, calcium, sodium and glucose. The benchmark contains 56017 drug-drug interactions of 922 drugs which is associated to an increased or decreased risk of an abnormal level of a mineral. For example, the interaction between the Canagliflozin and Miglitol drugs is associated with an increase of the risk of hypoglycemia (the condition in which your blood sugar (glucose) level is lower than normal).

The benchmark is formatted in triplets form where each entry represents a (drug, condition, drug) triplet. Each condition in the entries consist of two parts:risk modifier and risk type. The risk modifier is a basic increase or decrease flag while the risk type part denotes the risk type such as hyperkalemia, hyponatremia, etc.

This benchmark can be used in the assessment of relational models in the context of drug-drug interactions and their associated side-effects. The current popular polypharmacy side-effects benchmark provided by Zitnik et. al. [40] is based on a rather outdated TWOSIDES dataset [34] and it has a non-specific range of diverse side-effects. Our benchmark, on the other hand, is based on the DrugBank database (a recent, continually updated and more comprehensive resource) and focuses on a more specific set of DDI risks (e.g. the anomalies of minerals levels). This supports training more up to date and specific predictive models.

DDI-EFFICACY Benchmark. The DDI-EFFICACY benchmark is another drug-drug interaction based benchmark which is focused on the relation between these interaction and the therapeutic efficacy of the interacting drugs. The benchmark consists of 136127 drug-drug interactions of 3342 drugs and their effect (increase or decrease) on the therapeutic efficacy of the interacting drugs.

Similar to the DDI-MINERAL Benchmark benchmark, this benchmark provides a dataset which is focused on polypharmacy side-effects in relation to the drug efficacy. This benchmark can be used to assess the ability of relational models to predict such specific side-effects of interactions between drugs.

It is also worth noting that the types of polypharmacy side-effects included in both the DDI-EFFICACY and DDI-MINERAL benchmarks are not included at all in the current standard benchmarks such as Zitnik et. al. [40].

DPI-FDA Benchmark. The DPI-FDA consist of a set of drug target protein interactions of FDA approved drugs which is compiled from the KEGG and drug bank databases. This benchmark consist of 18928 drug protein interaction of 2277 drugs and 2654 proteins.

This benchmarks provides an extension to currently available benchmarks such as the DrugBank_FDA [38] and Yamanishi09 [39] benchmarks which have 9881 and 5127 DPIs respectively. Such an increase in the data size can enhance the training process of relational learning models and mitigate the generalization problems associated with the smaller benchmarks [26]. This extension of the number of DPIs provided in our benchmark is possible as we use the latest data releases of related biological databases unlike current benchmarks which we based on outdated versions (sometimes 10+ years old).

DPI-FDA-EXP Benchmark. The DPI-FDA-EXP is drug-protein association based benchmark which capture the effect of drugs on the expression levels of proteins in the living systems. The

benchmark contain 903429 statements on 1291 FDA approved drugs and their effects on the expression of 55196 proteins.

PPI-PHOSPHO Benchmark. Protein phosphorylation interactions is an enzymic protein–protein interaction which happens when a protein *i.e.* kinases, donates a phosphate group to another protein *i.e.* substrate. This type of PPIs is crucial for signalling in virtually any living cell. The PPI-PHOSPHO benchmark dataset is a kinase–substrate phosphorylation dataset which is based on the PhosphoSitePlus database [13] and the work of Hijazi et. al. [11]. The benchmark contains 25662 records in the format (<kinase>, PHOSPHORYLATES, <substrate>, <site>), where kinases and substrates are specific protein types represented using the proteins' UniProt ids and the site field represents the specific residue in substrate sequence which interacts with the kinase protein.

This benchmark provides a richer dataset for phosphorylation interactions which extends the currently used benchmarks [12, 28, 31] that suffer from limited coverage of kinases and substrates, and have fewer records due to dependence on older version of phosphorylation databases.

All the benchmarks can be downloaded from the biokg github repository at https://github.com/dsi-bdi/biokg/releases/download/v1.0.0/benchmarks.zip.

6 DISCUSSION

In this section, we discuss issues, lessons learned and other observations regarding the development and the use of BioKG knowledge graph in building relational learning models for biological applications.

6.1 Data Quality

In the process of building the BioKG knowledge graph, we tried to ensure the highest quality of all the data included by extracting data from curated sources exclusively. However, one of our included sources, the InterPro database for protein sequence annotations [21], is based on predicted sequence patterns using predefined rules and Markov models. We included this database as it is well-integrated with expert-curated SwissProt database and it compiles the most accurate parts of open sources for sequence annotations [21].

6.2 Availability

The implementation of the pipeline to build the BioKG contents is available at https://github.com/dsi-bdi/biokg. Downloadable BioKG contents (links, properties and metadata) are also available in the releasees section of the BioKG repository.

The BioDBLinker module is available as a Python module called *biodblinker* which can be installed using the standard *pip* process. The source code of the BioDBLinker module is also available on GitHub at https://github.com/dsi-bdi/biodblinker.

6.3 Limitations and Potential Issues

Despite the high coverage of biological entries in the BioKG, it still suffers from sparsity of data due to the unbalanced representation biological entities in open biological databases [24]. This unbalance is a result of the unbalanced research focus on specific entities, where some biological entities which are related to popular phenomenons are heavily studied, therefore, have richer database entries and annotations. This unbalance has a negative effect on relational learning models where they learn less efficient representations for the under-represented entities [24].

The use of BioKG and any other form of biological knowledge graphs can often lead to train-test data leakage when used without careful review of the relation between investigated phenomenons and the data in the knowledge graph. For example, the data on drug–drug combinations interactions (polypharmacy) is related to drug–protein interactions where two interacting drugs are often detected when they have the same protein target. The TWOSIDES database for instance uses DPIs to extract DDIs [34], therefore, using DPIs as extra data to support relational models in predicting DDIs can introduce indirect data leakage in such settings. We, therefore, suggest careful review of the relation between training knowledge graphs and investigated phenomenons in the testing data in biological predictive tasks to avoid such an issue.

Knowledge graphs and their embedding models are also biased towards well-studied biological entities which have better representation within the graph. Hence, the performance of models relying on biological KGs can suffer from low accuracy when executed on understudied or new biological entities which is absent from the graph. We, therefore suggest incorporating other forms of biological data such as protein sequences, protein structures and structures of chemical compounds into the predictive models to enhance their representations of the understudied entities.

6.4 Future Directions

We aim to provide updates to the BioKG in future works to keep it updated with the latest releases and changes of the source biological databases. We have recently investigated the principles and results presented here in the development of several state-of-the-art relational learning models [22–24, 26], and we aim to continue this line of work where we intend to assess the predictive accuracy and scalability of popular relational models on various other benchmarks based on the data introduced in this paper.

7 CONCLUSIONS

In this work, we proposed a new knowledge graph, BioKG, which covers a broad range of primary sources of biological data with the objective of supporting relational learning models on biological predictive tasks. The BioKG creation pipeline extracts data from open biological databases and provides them in a form of a graph of biological entities and their connections to each other along with their attributes and other related metadata. The contents of BioKG is compiled from expert-curated and popular sources to ensure high data quality and high level of integration.

We also provided a module for linking biological entities from different databases, the BioDBLinker which is based on open biological databases mappings. The module provides offline services for mapping between the different id systems for biological entities along with bidirectional name-id lookup services. The range and depth of resources covered as well as the flexibility in adding new sources arguably complements and extends currently available solutions, such as the Bio2RDF suite.

Furthermore, we have proposed a set of benchmarking datasets which can be are built from the drug–drug and drug–protein interactions data in the BioKG. These benchmarks cover different aspects related to both types of interactions and can be useful means for assessing the predictive accuracy of relational learning models in corresponding discovery tasks.

8 ACKNOWLEDGEMENTS

The work presented in this paper was supported by the CLARIFY project funded by European Commission under the grant number 875160, and by the Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland (supported by the Science Foundation Ireland grant (12/RC/2289_P2).

REFERENCES

[1] Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 43 (2015), D789 – D798.

[2] Amos Bairoch. 2018. The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of biomolecular techniques : JBT* 29 2 (2018), 25–38.

[3] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. 2008. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics* 41 5 (2008), 706–16.

[4] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*. 2787–2795.

[5] Gene Ontology Consortium. 2005. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research* 34 (2005), D322 – D326.

[6] The UniProt Consortium. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* 38 (2010), D142 – D148.

[7] The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 47 (2019), D506 – D515.

[8] David Croft and Gavin O'Kelly et. al. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* 39 (2011), D691 – D697.

[9] Nikolai Hecker, Jessica Ahmed, Joachim von Eichborn, Mathias Dunkel, Karel Macha, Andreas Eckert, Michael K. Gilson, Philip E. Bourne, and Robert Preissner. 2012. SuperTarget goes quantitative: update on drug–target interactions. *Nucleic Acids Research* 40 (2012), D1113 – D1117.

[10] Micheal Hewett, Diane E. Oliver, Daniel L. Rubin, Katrina L. Easton, Joshua M. Stuart, Russ B. Altman, and Teri E. Klein. 2002. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic acids research* 30 1 (2002), 163–5.

[11] Maruan Hijazi, Ryan Smith, Vinothini Rajeeve, Conrad Bessant, and Pedro R. Cutillas. 2020. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nature Biotechnology* 38 (2020), 493 – 502.

[12] Heiko Horn, Erwin Schoof, Jinho Kim, Xavier Robin, Martin L. Miller, Francesca Diella, Anita Palma, Gianni Cesareni, Lars Juhl Jensen, and Rune Linding. 2014. KinomeXplorer: an integrated platform for kinome biology studies. *Nature Methods* 11 (2014), 603–604.

[13] Peter V. Hornbeck, Jon M. Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research* 40 (2012), D261 – D270.

[14] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44 (2016), D457 – D462.

[15] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, Anchi Guo, and David Scott Wishart. 2011. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research* 39 (2011), D1035 – D1041.

[16] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research* 44 (2016), D1075 – D1079.

[17] Xin Liu, Feng Zhu, Xiaohua Ma, Lin Tao, Jingxian Zhang, Shengyong Yang, Yuquan Wei, and Y. Z. Chen. 2011. The Therapeutic Target Database: an internet resource for the primary targets of approved, clinical trial and experimental drugs. *Expert opinion on therapeutic targets* 15 8 (2011), 903–12.

[18] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*. www.cidrdb.org.

[19] Carolyn J. Mattingly, Glenn T. Colby, John N. Forrest, and James L. Boyer. 2003. The Comparative Toxicogenomics Database (CTD). *Environmental Health Perspectives* 111 (2003), 793 – 795.

[20] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[21] Alex L. Mitchell and Terri K. Attwood et. al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* 47 (2019), D351 – D360.

[22] Sameh K. Mohamed. 2020. Predicting tissue-specific protein functions using multi-part tensor decomposition. *Information Sciences* 508 (2020), 343–357.

[23] Sameh K Mohamed and Aayah Nounu. 2020. Predicting The Effects of Chemical-Protein Interactions On Proteins Using Tensor Factorisation. *AMIA Summits on Translational Science Proceedings* 2020 (2020), 430.

[24] Sameh K Mohamed, Aayah Nounu, and Vít Novácek. 2020. Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics* (02 2020). https://doi.org/10.1093/bib/bbaa012 bbaa012.

[25] Sameh K. Mohamed and Vít Novácek. 2019. Link Prediction Using Multi Part Embeddings. In *ESWC (Lecture Notes in Computer Science, Vol. 11503)*. Springer, 240–254.

[26] Sameh K. Mohamed, Vít Novácek, and Aayah Nounu. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 36, 2 (2020), 603–610.

[27] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* 104 (2016), 11–33.

[28] John C. Obenauer, Lewis C. Cantley, and Michael B. Yaffe. 2003. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research* 31 13 (2003), 3635–41.

[29] Rawan S. Olayan, Haitham Ashoor, and Vladimir B. Bajic. 2018. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 34 (2018), 1164 – 1173.

[30] Sandra E. Orchard, Mais G. Ammari, and Bruno Aranda et. al. 2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* 42 (2014), D358 – D363.

[31] Jiangning Song, Huilin Wang, Jiawei Wang, André Leier, Tatiana T. Marquez-Lago, Bingjiao Yang, Ziding Zhang, Tatsuya Akutsu, Geoffrey I. Webb, and Roger J. Daly. 2017. PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific Reports* 7 (2017).

[32] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* 34 (2006), D535 – D539.

[33] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Mínguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars Juhl Jensen, and Christian von Mering. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* 39 (2011), D561 – D568.

[34] Nicholas P. Tatonetti, Patrick Ye, Roxana Daneshjou, and Russ B. Altman. 2012. Data-driven prediction of drug effects and interactions. *Science translational medicine* 4 125 (2012), 125ra31.

[35] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 2071–2080.

[36] Mathias Uhlén, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernérus, Lisa Björling, and Frederik Pontén. 2010. Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology* 28 (2010), 1248–1250.

[37] Christian von Mering, Martijn A. Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic acids research* 31 1 (2003), 258–61.

[38] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36 (2008), D901–D906.

[39] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24 (2008), i232 – i240.

[40] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34 (2018), i457 – i466.