Computer Science and Information Systems 21(3):1055–1075 https://doi.org/10.2298/CSIS230530027G

A Comprehensive Review of the Data and Knowledge Graphs Approaches in Bioinformatics*

Ylenia Galluzzo

Department of Engineering, University of Palermo, Italy
ylenia.galluzzo01@unipa.it

Abstract. The scientific community is currently showing strong interest in constructing knowledge graphs from heterogeneous domains (genomic, pharmaceutical, clinical etc.). The main goal here is to support researchers in gaining an immediate overview of the biomedical and clinical data that can be utilized to construct and extend KGs. A in-depth overview of the available biomedical data and the latest applications of knowledge graphs, from the biological to the clinical context, is provided showing the most recent methods of representing biomedical knowledge with embeddings (KGEs). Furthermore, this review, differentiates biomedical databases based on their construction process (whether manually curated by experts or not), aiming to offer a detailed overview and guide researchers in selecting the appropriate database for their research considering to the specific project needs, available resources, and data complexity. In conclusion, the review highlights current challenges: integration of different knowledge graphs and the interpretability of predictions of new relations.

Keywords: Biomedical Knowledge Graph, Knowledge Graph Embeddings, Text Mining, Graph Neural Network

1. Introduction

Knowledge graphs are an area of great interest in both academia and industry, because they facilitate information extraction (facts and hypotheses) by to well-defined interconnections between relevant entities (abstract and concrete) within a given domain. In addiction, they are equally interesting for understanding how to form new relationships through the use of data semantics and linkages. Originally, knowledge graphs are represented as the knowledge base graphs in the Resource Description Framework (RDF). Information (Resource, a Property and a Property value) is represented through assertions forming SPO triples (subject, predicate, object) which express direct and complex relationships between different resources [63]. Knowledge graphs can also be described as an ontology. An ontology is a data model that represents knowledge about a specific through sets of relations among concepts within a domain and instances of objects representing the topic. The Web Ontology Language (OWL) serves as a markup language for expressing ontologies. RDF and OWL have become crucial standards within the Semantic Web. In 2012, Google popularized knowledge graphs with the introduction of its "Google Knowledge

* This is an extended version of a conference paper presented in a workshop K-GALS organized in conjunction with the conference ADBIS 2022.

Graph" [22]. This system uses Knowledge Vault, which combines probabilistic knowledge extracted from Web content with prior knowledge derived from existing knowledge repositories enabling users to receive relevant information based on the search queries. Currently, numerous open knowledge bases or ontologies have been published, including WordNet [71], DBpedia [6], Wikidata [24] etc. and industry Knowledge Graph (eg. Google, Microsoft's Bing, Facebook, eBay, IBM etc.) [77].

There are many papers summarising the current state of research on knowledge graphs. One of the most recent contributions is by Hogan et al. (2021) [45] who provide a comprehensive introduction to knowledge graphs. The authors compare existing data query models and languages, and summarize methods for creating, evaluating, and publishing knowledge graphs.

Knowledge plays an important role in reasoning-driven Natural Language Processing (NLP) tasks. Indeed, knowledge graphs have emerged as an important tool for addressing various NLP problems, such as Question Answering (KGQA) [49, 68, 83]. Semantics in information can help in extrapolating information that is more *semantically close* to the query. Structured knowledge is also a key element in conversational AI where virtual assistants (e.g. Alexa, Siri or Cortana) answer questions in an advanced way (open questions), as opposed in a more advanced manner to common chatbots programmed only to responde to strictly controlled questions (closed questions). Recently, research works have focused on collecting different techniques for constructing Knowledge Graphs (KG) and their application [134]. In particular, KGs have various application perspectives across different domains such as medical, financial, cybersecurity, news and education, social network de-anonymization, classification, geoscience.

Although several surveys on knowledge graph embeddings in general [16, 19, 112] and specifically on the biological topic [73] have been published over the past few years, this paper aims to explore and summarise the most recent advances in the application of KGs, providing a concise overview of the topic. The goal is to distinguish between biological and clinical domains, and highlight potential issues that may arise from careless construction of KGs, as well as providing information on how KGs support semantic knowledge. Current applications of the latest NLP models for creating clinical KGs are shown in this context. Furthermore, we introduce the most recent and promising future research paths (e.g., the use of multimodal approaches and Simplicial neural networks) in the fields of biomedical and precision medicine. Additionally, the paper expands on the study of usable resources for constructing a biomedical KG.

As this paper is an extended version of the conference paper [28] presented at a K-GALS workshop organized in conjunction with the ADBIS 2022 conference, new resources have been introduced for the construction of knowledge graphs (KGs). Moreover, it conducts an in-depth analysis of the methods and data currently used in biochemical and clinical applications.

2. Knowledge Graphs in Bioinformatics

The application of KGs in the field of biomedical data for decision support spanes from clinical to the biological applications. One of the earliest and most renowned rule-based systems for medical diagnosis is MYCIN [107] which has a knowledge base of 600 rules. There is a close connection between KG and biomedical NLP. On one hand, this connec-

tion allows for the enhancement of the amount and representation of data in KG. On the other hand KG enables improvements in predictions for solving NLP tasks (e.g. named entity recognition (NER) [57] and relation extraction [44]).

Relationship extraction systems are crucial for identifying connections between a wide variety of topics. For instance, they are needed to assess the relationship between non-pharmacological variables and COVID-19 pandemic as well as to support policy-making on COVID-19 in public health [125].

A knowledge graph in the biomedical field is used to connect a vast amount of interrelated information: genes with biological processes, molecular functions, and cellular components; genes with phenotype or interaction with other genes; drugs with the diseases they treat; genes responsible for diseases; generic symptoms related to diseases, etc. Using graphs as a representation of biomedical data seems to be the most natural solution for modelling objects of this type. In Fassetti et al. [25], graphs are used for the identification of features that characterize and at the same time discriminate gene expression among sets of healthy/diseased samples. This is accomplished through the identification of patterns within the graphs belonging to the sample sets with complementary health statuses. In the Table 1, most of the knowledge databases (KB) that are used for constructing and integrating knowledge in the context of biological and clinical data are listed. The data coverage and complexity are specified for each KB, along with their last update (release) dates.

Biomedical databases, in general, play a crucial role in scientific research contributing to drugs development, disease diagnosis and treatment, and understanding biological processes. Although some of these databases are not structured as knowledge graphs (KG), integrating them into a knowledge graph can maximize their potential. By connecting information from different databases and linking data of diverse nature (e.g., clinical data with genetic information), enables researchers to uncover hidden relationships and connections, potentially leading to the discovery of new associations and insights in biomedical research. Therefore, Table 2 lists the most well-known and utilized databases in biomedical research. In addition, the Table 2 distinguishes databases based on their creation methodology: manually curated (by experts), automatic extraction systems or mixed methodology (automatic and manually curated). This type of distinction is particularly important in the biomedical context. A manually curated biomedical database is often considered superior to one created through an automatic methods in certain contexts. Databases curated manually by experts are regarded as the best in terms of precision, reliability, contextualization, and continuous information updates. However, it is also important to note that this process is slow, expensive and requires the collaboration of industry-specific experts. The choice of information generation methods for populating these databases depends heavily on the resources available and the complexity of the data to be evaluated.

miRNA (microRNA) databases are crucial as they offer fundamental information about microRNA sequences, their functions, their interactions with target genes, and their involvement in biological and pathological processes. Currently, there are not many KG that utilize this type of data in conjunction with generic biomedical databases. The Table 3 displays the currently available human miRNA databases, for the same reasons as for general biomedical databases. These miRNA databases may not be structured as knowledge graphs but can still be used to discover new associations and insights regarding the roles

of miRNA in various biological and pathological contexts. Integrating miRNA databases into a knowledge graph provides a broader context and a more integrated approach to understanding the roles of miRNA in gene regulation and biological networks. This approach, along with the use of this data, can lead to the identification of biomarkers, and a better understanding of diseases. Using genetic, molecular and other specific KG information containing details about Human miRNA can be valuable in the context of precision medicine accelerating research and development of customized therapies.

In literature, although using different methods and algorithms, graphs are mostly used to solve common problems: making inferences about biomedicine, creating alternative ways to represent graphs on the same knowledge domain and extending information extraction.

A large part of research is currently devoted to the identification of similar entities within a KG. Embeddings generated using neural networks are used to calculate knowledge-based similarities between, for example, drugs, proteins and diseases [127].

Many ways are used to extend knowledge in the biomedical domain to discover latent information or missing information in KGs.

Completion of the knowledge graph (KGC) aims to complete the structure of the knowledge graph by predicting the missing entities or relationships in the knowledge graph and extracting unknown facts. KGC technologies may involve the use of traditional methods, such as rule-based reasoning and the probability graph model (Markov logic network). Recently, KGC techniques use methods of learning through embeddings representation: methods based on semantic correspondence models, based on learning of representation and other methods based on neural network models.

The use of models based on a generative approach to learn the embeddings of entities and relationships allows to generate hypotheses regarding the relationships associated with a connection score between graph embeds through multiple techniques: tensor factorisation (DistMult model [9]) and latent distance similarity (TransE model [124]). This type of techniques are used in polypharmacy, to evaluate the side effects that are caused by the interaction of drug combinations [70, 76].

2.1. Example of construction of knowledge graph

Constructing knowledge graphs from heterogeneous biomedical databases (see Table 1, Table 2, Table 3) involves several complex steps, such as data integration, ontology alignment, and semantic integration. To effectively navigate these challenges, it is essential to first define the objectives of the knowledge graph construction project. For instance in the context of cancer research, the objective may be to integrate diverse biomedical datasets to facilitate knowledge discovery, data-driven insights, and personalized medicine.

The following steps can guide the construction of such KG with the focus on cancer.

1. **Identify and selection relevant biomedical databases.** For example, to construct a KG specialized on cancer, we need to consider database containing genomic data (e.g., TCGA), drug data (e.g., DrugBank), molecular pathway databases, diseases-gene associations data (e.g, COSMIC, DISEASES). It's important to note that the chosen databases are mostly created through manual curation by experts (see Table 2, Table 3). In specific contexts, a manually curated biomedical database is frequently regarded as superior to one generated through automated methods, as mentioned previously.

Table 1. A list of 13 knowledge data sources that are useful in the biomedical context, is provided

Knowledge database	Description	Coverage	Last Release
STRING [101]	database of known and predicted protein-protein interactions	67.592.464 proteins from 14.094 organisms.	August 2021
iDISK [87]	Dietary Supplements (DS) Knowledge base	4.208 DS concept, 495 drugs, 776 diseases, 985 symptoms, 605 therapeutic classes, 17 system organ classes, and 137.568 DS products.	February 2020
Hetionet [43]	biomedical knowledge assembled from 29 different databases (genes, compounds, diseases, etc.)	47.031 nodes of 11 types and 2.250.197 edges of 24 types.	February 2017
DRKG [50]	biological knowledge graph relating genes, chemical compounds, biological processes, drug side effects, diseases, and symptoms.	100.000 entities of more than 12 types. 6.000.000 relationships of more than 100 types.	in 2020
KEGG [79]	reference knowledge base for integration and interpretation large-scale molecular data sets (genomic, chemical and health information)	563 pathway maps, 47.296.502 genes, 9.010 organisms, 2.640 human diseases, 12.136 drugs etc.	May 2023
PharmGKB [130]	knowledge on actionable gene-drug associations and genotype-phenotype relationships	759 drugs, 1761 genes, 213 pathways, 227 diseases, 200 clinical guidelines, and 993 drug labels.	in 2023
Gene Ontology (GO) [2]	describes knowledge of the biological domain: molecular function, biological process, cellular component	7.554.638 annotations 1.519.515 gene products 5.291 species.	May 2023
UniProtKB [3]	collection of annotated functional information on proteins	Swiss-Prot: 569.516 seq, 205.866.895 amino acids; TrEMBL: 249.308.459 seq, 86.853.323.495 amino acids.	May 2023
Reactome [31]	knowledge graph that focuses on biological pathways and their relationships	95.164 proteins, 102.459 complexes, 90.807 reactions, 22.050 pathways; 11.278 human proteins.	March 2023
OncoKB [10]	precision oncology knowledge base, consolidating biological and clinical data on genomic alterations in cancer	Memorial Sloan Kettering (MSK), provides accurate information about the biological and clinical implications of over 5.000 cancer gene alterations.	May 2023
OGB-Biokg [47]	biomedical knowledge graph constructed by Stanford University (associations between proteins, e.g., physical interactions, co-expression, homology or genomic neighborhood etc.)	5 types of entities: diseases (10,687 nodes), proteins (17,499), drugs (10,533 nodes), side effects (9,969 nodes), protein functions (45,085 nodes).	April 2023
Bioteque [26]	a resource of biological knowledge graph embeddings	12 types of biological entities (e.g. genes, diseases, drugs) and 67 types of relationships.	July 2022
NCBI [91]	knowledge graph representing gene-related information, including gene sequences, gene structures, functional annotations, and genetic variations.	242.554.936 GenBank sequences	April 2023

Table 2. A list of 24 biomedical databases, valuable for biomedical research yet not explicitly structured as KGs, is provided

Database	**Main scope**	**Manually curated?**
SCOPe [12]	protein structural relationships	Mostly manually
Protein Data Bank (PDB) [1]	archive of 3D structure data for large biological molecules	Yes
CATH [96]	hierarchical classification of protein domains	Mixed with other methods
InterPro [81]	functional analysis of protein sequences by classifying them into families and predicting domain presence and important sites	Mixed with other methods
The Human Protein Reference Database [58]	for each protein in the human proteome integrate information pertaining: domain architecture, post-translational modifications, interaction networks and disease association.	Yes
Bgee [7]	gene expression patterns across multiple animal species	Yes
HGNC [105]	relation between gene symbol and their corresponding entries in other database	Mixture with other methods
DrugBank [116]	molecular information about drugs, their mechanisms, their interactions and their targets	Yes
Supertarget [41]	integrates drug-related information associated with medical indications, adverse drug effects, drug metabolism, pathways and (GO) terms for target proteins	Yes
SIDER [62]	collects information on drug classification and side effects and links to further information, e.g. drug-target relations	No
OFFSIDES [104]	database of drug side-effects	No
TWOSIDES [104]	database drug-drug-effect	No
STITCH [102]	database of known and predicted interactions between chemicals and proteins	Yes
SIGNOR [102]	causal relationships between human proteins, chemicals of biological relevance, stimuli and phenotypes	Yes
SMPDB [51]	database containing pathways found in model organisms such as humans, mice, E. coli etc.	No
ChEMBL [30]	chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs	Yes
ChEBI [40]	dictionary of molecular entities focused on ‘small’ chemical compounds	Yes
PubChem [59]	chemical database	Yes
TISSUE [90]	tissue expression, proteomics and transcriptomics screens	Mixture with other methods
Brenda Tissue Ontology [13]	collection of enzyme functional data	Not specified
Disease Ontology [93]	ontology for human disease	Not specified
Cell Ontology [55]	repository for biomedical ontologies	Not specified
DISEASES [82]	integrates disease-gene associations, cancer mutation data, and genome-wide association studies	Mixed with other methods
COSMIC [27]	Catalogue Of Somatic Mutations In Cancer, associate genes with the related cancer type	Yes

Table 3. A list of 8 databases containing Human miRNA information, valuable for biomedical research yet not explicitly structured as KGs, is provided

Database	**Main Goal**	**Manually curated?**
mirCancer [120]	contains associations between miRNAs and human cancers.	Mixed with other methods
miRanda [52]	miRNA-target interactions.	Not specified
miRBase [60]	miRNAs sequences and annotations, associated with names, keywords, genomic locations, and references.	No
miRNASNP [34]	contains miRNA-related mutations.	Mixed with other methods
miRTarBase [48]	miRNA-target interactions, including those implicated in cancer. Experimentally validated miRNA-target interactions.	Mixed with other methods
DIANA-TarBas [108]	miRNA-target interactions, including those related to cancer. It offers information on miRNA regulation of target genes and associated functional effects.	Yes
OncomiRDB [109]	miRNA expression profiles in various cancer types, along with their putative target genes and functional implications.	Yes
TCGA [4]	Cancer Genomics Databases. It provides miRNA dysregulation patterns in cancer and their potential roles as biomarkers or therapeutic targets.	Yes

2. **Develop an integration strategy.** Use Extract, Transform, Load (ETL) pipelines or data ingestion tools to extract data from these heterogeneous sources. Each source may have its own data format (e.g Json, XML, RDF etc.), and schema.
3. **Ontology Alignment.** Within cancer research, critical ontologies include the Human Phenotype Ontology (HPO) [36], which catalogues phenotypic abnormalities linked to genetic diseases, and the Disease Ontology (DO) [93] providing a standardized vocabulary for disease classification. Complementary ontologies encompass the Gene Ontology (GO) [2], delineating molecular functions, biological processes, and cellular components, as well as the Cell Ontology [55], characterizing cell types and anatomical structures. At this step, it is necessary to establish correspondences or mappings between entities of different ontologies. Furthermore, it is important to establish common standards and formats to facilitate comparison and alignment. Software platforms like e.g BioPortal can help researchers to systematically assess the efficacy and accuracy of ontology alignment methodologies in biomedical domains.
4. **Mapping data into graph-based data model.** The goal of this step is to develop and define a standardized schema for our data, pinpointing crucial data entities and their relationships. RDF (Resource Description Framework) triples, can link together diverse elements such as genes, proteins, diseases, drugs, and pathways. This approach fosters a comprehensive comprehension of intricate biological systems by forging significant links between different components.
5. **Semantic Integration.** Employ semantic reasoning techniques (e.g, Rule-based Reasoning, Semantic Similarity etc.) to deduce relationships and uncover novel insights from the heterogeneous data.
6. **Quality check.** Define robust quality metrics to quantitatively evaluate the comprehensiveness, precision, and coherence of the constructed knowledge graph. Conduct thorough validation procedures to verify data integrity and adherence to ontology mappings, employing systematic validation protocols. In this step, for example is necessary to solicit feedback on the accuracy, relevance, and completeness of the mappings, incorporating expert insights to refine the validation process, ensuring its scientific validity and applicability in the field.

2.2. Methods of Knowledge Graph Embedding

A very common application that has grown recently is the creation of entity embeddings or assertions on KGs by training deep learning networks such as autoencoders from inputs constructed by KG nodes [69, 133]. The purpose of representing graphs in a high-dimensional space to a low-dimensional space is to capture the essence of a graph while preserving its intrinsic (global and / or local) structure in the form of a dense vector representation, both of arcs [9] and of single nodes [35]. This approach has been used to analyze knowledge graphs across different domains enabling, starting from compressed and *meaningful* information, the application of classification techniques and the development of predictors, which can aid researchers in identifying associations between diseases and bio-molecules [95], discovering new treatments for existing drugs [43] and addressing other related problems.

In the following section, we will explore how the most recent techniques of *representative learning* are applied in the biochemical and clinical context.

Application on Biochemical Data. One of the studies that facilitated the creation of a comprehensive KG in the biomedical context is PharmKG [130]. This study aggregated multi-omics data, including disease-related words, gene expression and chemical structure information, while preserving biological and semantic features through the latest KGE approaches. A significant aspect of the study focused on drug-related topics, particularly addressing drug reuse and adverse reactions, crucial for preventing harm to patients. Additionally, the study investigated potential drug-drug interactions (DDIs), drug-protein (DPIs), drug-disease, drug-target interactions (DTIs) by modelling the problem of predicting links between graph nodes with KGs. In the study by Zhu et al. [132], the authors provided a detailed overview of existing drug knowledge bases and their applications. This work used datasets containing key properties of drugs (DrugBank [116] and SuperTarget [41]) as well as datasets containing the main information about chemical compounds (PubChem [59] and ChEMBL [30]). In the study by Lin et. al [65], a Bio2RDF created from DrugBank was employed to evaluate the relationships between a potential drug and its neighbours. This evaluation is based on GNN models applied to the biological KG. Innovative work in drug research includes the development of the BioDKG-DDI model [86], that aims to identify DDI interaction relationships to support experimental work in drug development laboratories. BioDKG-DDI uses an innovative self-attention mechanism on DNNs (deep neural networks)to attenuate multi-features embeddings including molecular structures, drug structures, and drug similarity matrix. Furthermore, recent advancements in DNNs are employed to predict the search for similar drugs, e.g. to identifying molecules with antibiotic properties through graph-based retro-synthesis [66]. Drug-protein interactions are another area of interest for researchers. In BridgeDPI [118], convolutional CNN and feed-forward network layers are used to encode SMILES representation of drug and protein sequences. GNNs are then employed, as in other similar works in the literature, to build *bridge* nodes between interactions and predict new connections between nodes.

Many papers in scientific literature in recent years have focused on the study of graph models applied to KG for the discovery of Drug-Target Interactions (DTIs) [5, 56, 84, 100, 110, 113, 114]. Drug-Target Interactions (DTIs) are the interactions between drugs and molecular targets in the human body. The interest in this type of interaction is justified by the fact that it is now essential to understand how drugs act in the body, how they bind to molecular targets and how they affect biological processes. These types of interactions can determine the effectiveness of a drug in treating a disease and may highlight its unwanted side effects. In pharmaceutical research it is crucial to have a comprehensive picture of DTIs interactions in a KG in order to develop safer, more effective and targeted drugs for the treatment of complex diseases. A new DT2Vec+ [5] approach to the computational reprocessing of drugs to predict new drug-target interactions (DTIs) was proposed in 2023, which showed promising results. DT2Vec+ was created by integrating and mapping drug-target-disease triplet association graphs. The heterogeneous graph in DT2Vec+ with “drug”, “target” and “disease” entities has been mapped to low-size vectors using node embedding principles to create specific characteristics for each entity. The authors also tested the new method on DTI tasks to propose drugs targeted at specific cancer biomarkers. Another approach for DTI predictions is KG-DTI [113]. The KG is constructed using 29,607 positive drug-target pairs. To extract the built-in features, KG-DTI, uses the DistMult embedding strategy instead. KG-DTI is then applied to recommend drugs for AD (Alzheimer’s disease) by targeting apolipoprotein E.

The results presented by the authors of KG-DTI show that seven of the top ten drugs recommended for AD are supported and validated by clinical practice and literature. KG2ECapsule [100] employs entity representations obtained by recursively propagating takeovers from the receptive fields of attention-based entities, similar to DTI-GAT (Drug-Target Interaction prediction with Graph networks attention) [110]. In particular, DTI-GAT uses the attention mechanism on graphs to facilitate the topological interpretation of DTI, assigning a different attention weight to each node in KG. The accuracy rate of DTI-GAT reaches 93.75, on enzyme dataset (BRENDA [92]), surpassing that of other prediction methods. An innovative approach for DPI is employed by the authors of TransDTI [56] (Transformer-Based Language Models for Estimating DTIs). They use transformer-based language models to classify interactions between drug-target pairs as active, inactive, and intermediate. The results presented by the TransDTI authors suggest that transformer-based linguistics effectively predict new drug-target interactions from sequence data. In 2022, GCHN-DTI [114], introduced a heterogeneous network created from various data sources including drug-target interactions, drug-drug interactions, similarities between drugs, target-target interactions and similarities between targets. In GCHN-DTI utilizes a graph convolution approach for the DTI task. The method employs an attention mechanism between convolutional graph layers to combine the embedding of nodes of each layer. GCHN-DTI demonstrates superiority over several state-of-the-art methods. One knowledge graph embedding approach that integrates and works well on DDI (Drug-Drug Interaction), DTI (Drug-Target Interaction) and PPI (Protein-Protein Interaction) is ConvE-Bio [84]. While ConvE-Bio serves as a powerful tool for predicting biomedical relationships it currently faces limitations related to processing large graphs. As presented in the Table 4 several noteworthy works focus on solving different tasks. For "diseases diagnosis" a recent tool based on the study gene association information and cofunctional gene modules is MLA-GNN(multi-level attention graph neural network) [121]. MLA-GNN achieves state-of-art performance on transcriptomic data [4] and proteomic data (COVID-19). The authors also employ an innovative mechanism to try to identify the genes most involved in model analysis and prediction. Another type of task studied by researchers in this context is the application of knowledge graph-based "disease-gene" prediction. The GenePredict-KG model [29] is developed for this purpose by integrating several datasets. Despite achieving results that surpass state-of-the-art performance, the method suffers from several limitations related to class imbalance.

In order to provide a comprehensive overview of current research, it is important to mention the application of Neural networks known as Hyperbolic Graph Neural Networks (HGNN) to the DISEASE dataset, based on the SIR disease spreading model [11], demonstrating excellent results in link prediction. However, recent advancements have shown that these new sophisticated neural networks have been outperformed by Simplicial neural networks (SNNs) for link prediction [15], achieving better results in terms of ROC AUC on the same dataset.

In the following Table 4 we will succinctly present, the latest applications of knowledge graph embedding in the context of biochemical data and the tasks they aim to address.

Application on Clinical Data. Studies of KG-based recommendation systems built from electronic medical records (EMRs) aim to enhance medical decision-making for improved

Table 4. Latest KGs constructions and Graph Neural Network applications in the biochemical field

Model	Task	Dataset	Year
BioDKG-DDI [86]	DDI	DrugBank [116], SIDER [62], KEGG [79], PubChem [59] and OFFSIDES [104]	2022
BridgeDPI [118]	DPI	BindingDB [32],C.ELEGANS and HUMAN datasets [67], DUD-E [75]	2022
RetroGNN [66]	Drug Discovery	Zinc15 database [99]	2020
HGNN [11]	Link Prediction, Node Classification	Disease, PubMed	2019
ConvE-Bio [84]	DDI, DTI, PPI	DrugBank [116], Human Protein [88]	2023
DT2Vec+ [5]	DTI	CTD [20], DrugBank [116], ChEMBL [30]	2023
KG-DTI [113]	DTI, DTP	DrugBank [116]	2021
KG2ECapsule [100]	DDI	DrugBank [116], OGB-Biokg [47], KEGG [79]	2023
DTI-GAT [110]	DTI	SuperTarget [41], DrugBank [116], KEGG [79], BRENDA [13]	2021
TransDTI [56]	DTI	ChEMBL [30], Kiba [103]	2022
GCHN-DTI [114]	DTI	DrugBank [116]	2022
KG-COVID-19 [85]	ML tasks, queries	PharmGKB [130], Therapeutic Target Database (TTD) [14], ChEMBL [30], GO [2], STRING [101], IntAct Molecular Interaction Database [80]	2021
MLA-GNN [121]	Disease diagnosis	TCGA [4], COVID-19 [111]	2022
KGMultiple Ontologies [78]	Gene-Disease Association	Uniprot [18], OMIM [38], Orphanet [115]	2020
GenePredict-KG [29]	Gene-Disease Association	STRING [101], SIDER [62], DrugBank [116], Human Phenotype Ontology (HPO) [36] Genotype-Tissue Expression (GTEx) [128] Gene Ontology Annotation (GOA) [17] Mammalian Phenotype (MP) [97] Mouse Genome Informatics (MGI) [23], PubChem [59], OMIM [38]	2023
GNBR [98]	Drug repurposing	Orphanet [115], OMIM [38], UMLS [8] DrugCentral [106]	2019
Compact Walks [46]	Pathways discovery	Hetionet [43], ROBOKOP [74]	2022

patient care. Creating KGs from medical record texts containing a patient's treatment history (medical diagnoses, therapies, etc.) presents a cost-effective approach compared to building KGs based on deeper biological aspects (relationships between genes, diseases, chemical composition of drugs, etc.) that require more attention. Research in this area is still in its early stages and recent advancements in information extraction models (such as LSTM, BERT, and NER models) enable the extraction of meaningful information from unstructured data, enriching biomedical knowledge bases with non-trivial connections [39] [64] [33]. Recently, Zhang et al. [129] demonstrated the effectiveness of attention mechanisms and convolutional graphs techniques in creating embedded KGs features enhance the classification and generation of radiological reports in order to improve diagnosis and support physicians in their work.

The Table 5 presents several noteworthy research studies from last three years, which incorporate clinical data of various types (e.g, images, ontologies, etc.). It is noteworthy that currently only MKGs [117] constructs KGs that encompass both biomedical and clinical data. As a result, it has the potential to address numerous tasks such as drug-drug interactions (DDIs), drug-protein interactions (DPIs), classification of nodes, and more.

Table 5. Latest KGs constructions and applications in the clinical field

Model	Task	Dataset	Year
ClinicalBERT [89]	link prediction	PubMed abstract, MIMIC-III database [54]	2021
SMR [33]	link prediction	MIMIC-III [54], DrugBank [116], ICD-9 ontology [94]	2020
RR-KG [129]	generation of radiological reports	U-RR dataset [21]	2020
MaKG [123]	generation of radiological reports	IU XRay [21]and MIMIC CXR [53]	2022
MKGs [117]	several	real world data (EMH, EHR etc), UMLS [8], ROBOKOP [74], DrugBank [116], UniProt [18], InterProt [81], SIDER [62], GO [2], KEGG [79], Therapeutic target database [14] etc.	2023

3. Open research problems

3.1. Construction and integration of knowledge graphs

Biomedical knowledge graphs are typically curated manually by expert researchers. One such example is COSMIC [27], constructed by a group of domain experts who associated genes with related cancer types based on literature. However, the field of biological knowledge is constantly evolving, necessitating scalable intelligent systems capable of integrating real-time updates. Addressing this challenge involves not only updating knowledge bases but also ensuring the reliability of knowledge representations in KGs and their relationships. One widely studied technique for enhancing KG reliability involves aligning entities from different KGs based on their similarity. Recently, Xiang et al. [119] introduced a method that incorporates ontology hierarchies and class disjunctions to improve entity alignment accuracy and avoid mismapping.

Research has also shown that the quality of available knowledge graphs directly impacts the accuracy of knowledge graph embedding (KGE) predictions. Low data quality can propagate into embedding models, leading to decreased prediction accuracy [73]. Missing knowledge and integration errors in KGs can further exacerbate this issue, perpetuating incorrect and misleading domain knowledge. This is particularly problematic in the biomedical domain, where inaccuracies can have significant consequences.

3.2. Performance

Complex biological systems are often represented as graphs, but the exploration, training and prediction techniques applied to these graphs require significant of resources and time leading to limited scalability. While knowledge KGEs address some aspects of this problem by operating with linear time and space complexity, the challenge of dynamically encoding new entities into the graph remains unresolved. KGEs rely heavily on prior knowledge of embeddings for each type of information in the knowledge base, allowing them to maintain both local and global information. However, this dependence on prior knowledge presents scalability issues that propagate into the prediction process.

3.3. Explainable predictions

Lack of interpretability is a recurring problem with deep learning models [37]. Which becomes particularly concerning given the increasing use of neural networks in decision-making within biomedical applications. Efforts have been made in this regard to address this issue. For example, CrossE [126] explores the process of explaining graph search paths using embeddings to interpret link prediction. In the context of KGE, learning meaningful embeddings through specific optimisation techniques often leads to predictions that are difficult to interpret. In data analysis, GNN models frequently employed in the biomedical domain, generate relevant information for each data node, thereby enhancing interpretability to some extent. Recent efforts have aimed to impose constraints during training to make KGE models partially interpretable (e.g. type constraints and basic relation axioms) [61, 72]. Extracting information from NLP presents challenges for constructing reliable KG in the health domain. Complex models used for understanding natural language still have many issues [42]. Importantly, biases inherent in extracting information from EMHs should not be underestimated. Inevitably these biases in the data will propagate to some extent in the results of the predictions. Therefore is crucial for research to prioritize the development of more reliable and explainable models for the healthcare sector.

4. Discussion and conclusion

This survey aims to present the latest models and strategies to use knowledge graphs in the biomedical context. Their use has become increasingly widespread in recent years, with current research focusing on enhancing the outcomes derived from their application in biomedicine. As outlined in this survey, many knowledge graphs are typically constructed from data sources, which are either manually curated by experienced researchers or generated through sophisticated NLP techniques (NER, relation extraction). We subsequently pointed out the potential errors that this approach may introduce in the biomedical context, during KGs construction. In this regard, with the aim of ensuring the future research in the biomedical domain is increasingly reliable and accurate, this review delves into the detailed construction methods of biological, chemical and clinical databases (see Table 2 and Table 3). The differences between the types of entities used in the biomedical knowledge bases and their "size" are noted in Table 1.

The process of extending knowledge in KGEs can indeed be addressed by the low-dimensional representation of the characteristics of each entity and/or relation within the graph. This compressed and representative representation of the knowledge graph can help identify potential inconsistencies during the integration process the graphs and partially resolve some problems associated with errors in knowledge graph construction caused by misaligned entities. KGEs are currently highly active area of research, due to their ability to provide a generalisable context on the KG and probabilistically deduce new relations missing in the existing graph structure. This characteristic has accelerated the discovery of new drugs in many studies by evaluating the interaction between properties of molecules present in the KG. The importance of KG feature representation, as discussed, underscore its effectiveness in constructing increasingly comprehensive KGs.

A recent advancement in research involves the construction of a multimodal knowledge network, where additional information is incorporated into the KG to enhance rea-

soning. This approach utilizes a combination of various interaction features among KG entities to improve predictions (e.g. on drug repositioning) [122]. The multimodal approach has also been recently applied in precision medicine, where detailed knowledge and a specific focus are essential for creating KGs that represent and generate *reliable* knowledge [131].

In conclusion, this discussion highlights the current open challenges in the use of KGs in the biomedical field, emphasizing the need to improve the interpretability and quality of biomedical KG data in order to increase confidence within the community regarding predictions and thereby support advancements in specialised medicine.

References

1. Protein data bank: the single global archive for 3d macromolecular structure data. Nucleic acids research 47(D1), D520–D528 (2019)
2. The gene ontology resource: enriching a gold mine. Nucleic acids research 49(D1), D325–D334 (2021)
3. Uniprot: the universal protein knowledgebase in 2021. Nucleic acids research 49(D1), D480–D489 (2021)
4. 53, D.C.C.B.R..J.M.A..K.A..P.T..P.D..W.Y., 68, T.S.S.L.D.A.: The cancer genome atlas pan-cancer analysis project. Nature genetics 45(10), 1113–1120 (2013)
5. Amiri Souri, E., Chenoweth, A., Karagiannis, S., Tsoka, S.: Drug repurposing and prediction of multiple interaction types via graph embedding. BMC bioinformatics 24(1), 1–17 (2023)
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. pp. 722–735. Springer (2007)
7. Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S.S., De Farias, T.M., Moretti, S., Parmentier, G., De Laval, V.R., Rosikiewicz, M., et al.: The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Research 49(D1), D831–D847 (2021)
8. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32(suppl_1), D267–D270 (2004)
9. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26 (2013)
10. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al.: Oncokb: a precision oncology knowledge base. JCO precision oncology 1, 1–16 (2017)
11. Chami, I., Ying, Z., Ré, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. Advances in neural information processing systems 32 (2019)
12. Chandonia, J.M., Guan, L., Lin, S., Yu, C., Fox, N.K., Brenner, S.E.: Scope: improvements to the structural classification of proteins–extended database to facilitate variant interpretation and machine learning. Nucleic acids research 50(D1), D553–D559 (2022)
13. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., Schomburg, D.: Brenda, the elixir core data resource in 2021: new developments and updates. Nucleic acids research 49(D1), D498–D508 (2021)
14. Chen, X., Ji, Z.L., Chen, Y.Z.: Ttd: therapeutic target database. Nucleic acids research 30(1), 412–415 (2002)
15. Chen, Y., Gel, Y.R., Poor, H.V.: Bscnets: block simplicial complex neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 6333–6341 (2022)

16. Choudhary, S., Luthra, T., Mittal, A., Singh, R.: A survey of knowledge graph embedding and their applications. arXiv preprint arXiv:2107.07842 (2021)
17. Consortium, G.O.: Gene ontology consortium: going forward. Nucleic acids research 43(D1), D1049–D1056 (2015)
18. Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research 47(D1), D506–D515 (2019)
19. Dai, Y., Wang, S., Xiong, N.N., Guo, W.: A survey on knowledge graph embedding: Approaches, applications and benchmarks. Electronics 9(5), 750 (2020)
20. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T.C., Mattingly, C.J.: The comparative toxicogenomics database: update 2019. Nucleic acids research 47(D1), D948–D954 (2019)
21. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23(2), 304–310 (2016)
22. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610 (2014)
23. Eppig, J.T.: Mouse genome informatics (mgi) resource: genetic, genomic, and biological knowledgebase for the laboratory mouse. ILAR journal 58(1), 17–41 (2017)
24. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing wikidata to the linked data web. In: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13. pp. 50–65. Springer (2014)
25. Fassetti, F., Rombo, S.E., Serrao, C.: Discovering discriminative graph patterns from gene expression data. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing. pp. 23–30 (2016)
26. Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M., Aloy, P.: Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. Nature Communications 13(1), 5304 (2022)
27. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al.: Cosmic: somatic cancer genetics at high-resolution. Nucleic acids research 45(D1), D777–D783 (2017)
28. Galluzzo, Y.: A review: Biological insights on knowledge graphs. In: New Trends in Database and Information Systems: ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5–8, 2022, Proceedings. pp. 388–399. Springer (2022)
29. Gao, Z., Pan, Y., Ding, P., Xu, R.: A knowledge graph-based disease-gene prediction system using multi-relational graph convolution networks. In: AMIA Annual Symposium Proceedings. vol. 2022, p. 468. American Medical Informatics Association (2022)
30. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(D1), D1100–D1107 (2012)
31. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al.: The reactome pathway knowledgebase 2022. Nucleic acids research 50(D1), D687–D692 (2022)
32. Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J.: Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research 44(D1), D1045–D1053 (2016)
33. Gong, F., Wang, M., Wang, H., Wang, S., Liu, M.: Smr: medical knowledge graph embedding for safe medicine recommendation. Big Data Research 23, 100174 (2021)

34. Gong, J., Tong, Y., Zhang, H.M., Wang, K., Hu, T., Shan, G., Sun, J., Guo, A.Y.: Genome-wide identification of snps in microrna genes and the snp effects on microrna target binding and biogenesis. Human mutation 33(1), 254–263 (2012)
35. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)
36. Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., et al.: The human phenotype ontology: semantic unification of common and rare disease. The American Journal of Human Genetics 97(1), 111–124 (2015)
37. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 1–42 (2018)
38. Hamosh, A., Scott, A.F., Amberger, J., Valle, D., McKusick, V.A.: Online mendelian inheritance in man (omim). Human mutation 15(1), 57–61 (2000)
39. Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., El Asri, B.: Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update 1, 100042 (2021)
40. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., Steinbeck, C.: Chebi in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research 44(D1), D1214–D1219 (2016)
41. Hecker, N., Ahmed, J., Von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., Gilson, M.K., Bourne, P.E., Preissner, R.: Supertarget goes quantitative: update on drug–target interactions. Nucleic acids research 40(D1), D1113–D1117 (2012)
42. Helwe, C., Clavel, C., Suchanek, F.M.: Reasoning with transformer-based models: Deep learning, but shallow reasoning. In: 3rd conference on automated knowledge base construction (2021)
43. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 6, e26726 (2017)
44. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 541–550 (2011)
45. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., d Melo, G.: Knowledge graphs [j]. synthesis lectures on data semantics and knowledge (2021)
46. Hou, P.Y., Korn, D.R., Melo-Filho, C.C., Wright, D.R., Tropsha, A., Chirkova, R.: Compact walks: Taming knowledge-graph embeddings with domain-and task-specific pathways. In: Proceedings of the 2022 International Conference on Management of Data. pp. 458–469 (2022)
47. Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J.: Open graph benchmark: Datasets for machine learning on graphs. arXiv preprint arXiv:2005.00687 (2020)
48. Huang, H.Y., Lin, Y.C.D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., et al.: mirtarbase update 2022: an informative resource for experimentally validated mirna–target interactions. Nucleic acids research 50(D1), D222–D230 (2022)
49. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: Proceedings of the twelfth ACM international conference on web search and data mining. pp. 105–113 (2019)
50. Ioannidis, V.N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., Karypis, G.: Drkg-drug repurposing knowledge graph for covid-19. arXiv preprint arXiv:2010.09600 (2020)

51. Jewison, T., Su, Y., Disfany, F.M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., et al.: Smpdb 2.0: big improvements to the small molecule pathway database. Nucleic acids research 42(D1), D478–D484 (2014)
52. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., Marks, D.S.: Human microrna targets. PLoS biology 2(11), e363 (2004)
53. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
54. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data 3(1), 1–9 (2016)
55. Jupp, S., Burdett, T., Leroy, C., Parkinson, H.E.: A new ontology lookup service at embl-ebi. SWAT4LS 2, 118–119 (2015)
56. Kalakoti, Y., Yadav, S., Sundar, D.: Transdti: Transformer-based language models for estimating dtis and building a drug recommendation workflow. ACS omega 7(3), 2706–2717 (2022)
57. Karampatakis, S., Dimitriadis, A., Revenko, A., Blaschke, C.: Training ner models: knowledge graphs in the loop. In: The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece, May 31–June 4, 2020, Revised Selected Papers 17. pp. 135–139. Springer (2020)
58. Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al.: Human protein reference database—2009 update. Nucleic acids research 37(suppl_1), D767–D772 (2009)
59. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al.: Pubchem substance and compound databases. Nucleic acids research 44(D1), D1202–D1213 (2016)
60. Kozomara, A., Griffiths-Jones, S.: mirbase: integrating microrna annotation and deep-sequencing data. Nucleic acids research 39(suppl_1), D152–D157 (2010)
61. Krompaß, D., Baier, S., Tresp, V.: Type-constrained representation learning in knowledge graphs. In: The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14. pp. 640–655. Springer (2015)
62. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic acids research 44(D1), D1075–D1079 (2016)
63. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specification, w3c recommendation 22 february 1999 (1999)
64. Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T.H., Wang, S., et al.: Real-world data medical knowledge graph: construction and applications. Artificial intelligence in medicine 103, 101817 (2020)
65. Lin, X., Quan, Z., Wang, Z.J., Ma, T., Zeng, X.: Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In: IJCAI. vol. 380, pp. 2739–2745 (2020)
66. Liu, C.H., Korablyov, M., Jastrzebski, S., Włodarczyk-Pruszyński, P., Bengio, Y., Segler, M.H.: Retrognn: Approximating retrosynthesis by graph neural networks for de novo drug design. arXiv preprint arXiv:2011.13042 (2020)
67. Liu, H., Sun, J., Guan, J., Zheng, J., Zhou, S.: Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics 31(12), i221–i229 (2015)
68. Lukovnikov, D., Fischer, A., Lehmann, J., Auer, S.: Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th international conference on World Wide Web. pp. 1211–1220 (2017)
69. Ma, T., Xiao, C., Zhou, J., Wang, F.: Drug similarity integration through attentive multi-view graph auto-encoders. arXiv preprint arXiv:1804.10850 (2018)

70. Malone, B., García-Durán, A., Niepert, M.: Knowledge graph completion to predict polypharmacy side effects. In: Data Integration in the Life Sciences: 13th International Conference, DILS 2018, Hannover, Germany, November 20-21, 2018, Proceedings 13. pp. 144–149. Springer (2019)
71. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
72. Minervini, P., Costabello, L., Muñoz, E., Nováček, V., Vandenbussche, P.Y.: Regularizing knowledge graph embeddings via equivalence and inversion axioms. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10. pp. 668–683. Springer (2017)
73. Mohamed, S.K., Nounu, A., Nováček, V.: Biological applications of knowledge graph embedding models. Briefings in bioinformatics 22(2), 1679–1693 (2021)
74. Morton, K., Wang, P., Bizon, C., Cox, S., Balhoff, J., Kebede, Y., Fecho, K., Tropsha, A.: Robokop: an abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics 35(24), 5382–5384 (2019)
75. Mysinger, M.M., Carchia, M., Irwin, J.J., Shoichet, B.K.: Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. Journal of medicinal chemistry 55(14), 6582–6594 (2012)
76. Nováček, V., Mohamed, S.K.: Predicting polypharmacy side-effects using knowledge graph embeddings. AMIA Summits on Translational Science Proceedings 2020, 449 (2020)
77. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. Queue 17(2), 48–75 (2019)
78. Nunes, S., Sousa, R.T., Pesquita, C.: Predicting gene-disease associations with knowledge graph embeddings over multiple ontologies. arXiv preprint arXiv:2105.04944 (2021)
79. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic acids research 27(1), 29–34 (1999)
80. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al.: The mintact project—intact as a common curation platform for 11 molecular interaction databases. Nucleic acids research 42(D1), D358–D363 (2014)
81. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al.: Interpro in 2022. Nucleic Acids Research 51(D1), D418–D427 (2023)
82. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J.: Diseases: Text mining and data integration of disease–gene associations. Methods 74, 83–89 (2015)
83. Purkayastha, S., Dana, S., Garg, D., Khandelwal, D., Bhargav, G.: Knowledge graph question answering via sparql silhouette generation. arXiv preprint arXiv:2109.09475 (2021)
84. Qu, X., Cai, Y.: Conve-bio: Knowledge graph embedding for biomedical relation prediction. In: 2023 International Conference on Intelligent Supercomputing and BioPharma (ISBP). pp. 10–13. IEEE (2023)
85. Reese, J.T., Unni, D., Callahan, T.J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K.A., Good, B.M., Balhoff, J.P., Fontana, T., et al.: Kg-covid-19: a framework to produce customized knowledge graphs for covid-19 response. Patterns 2(1), 100155 (2021)
86. Ren, Z.H., Yu, C.Q., Li, L.P., You, Z.H., Guan, Y.J., Wang, X.F., Pan, J.: Biodkg–ddi: predicting drug–drug interactions based on drug knowledge graph fusing biochemical information. Briefings in Functional Genomics 21(3), 216–229 (2022)
87. Rizvi, R.F., Vasilakes, J., Adam, T.J., Melton, G.B., Bishop, J.R., Bian, J., Tao, C., Zhang, R.: idisk: the integrated dietary supplements knowledge base. Journal of the American Medical Informatics Association 27(4), 539–548 (2020)
88. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5), 742–754 (2010)

89. Roy, A., Pan, S.: Incorporating medical knowledge in bert for clinical relation extraction. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 5357–5366 (2021)
90. Santos, A., Tsafou, K., Stolte, C., Pletscher-Frankild, S., O'Donoghue, S.I., Jensen, L.J.: Comprehensive comparison of large-scale tissue expression datasets. PeerJ 3, e1054 (2015)
91. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al.: Ncbi taxonomy: a comprehensive update on curation, resources and tools. Database 2020 (2020)
92. Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A., Schomburg, D.: The brenda enzyme information system–from a database to an expert system. Journal of biotechnology 261, 194–206 (2017)
93. Schriml, L.M., Munro, J.B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J.A., Jackson, R., Bello, S.M., Bearer, C., et al.: The human disease ontology 2022 update. Nucleic acids research 50(D1), D1255–D1261 (2022)
94. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research 40(D1), D940–D946 (2012)
95. Shen, Z., Zhang, Y.H., Han, K., Nandi, A.K., Honig, B., Huang, D.S.: mirna-disease association prediction with collaborative matrix factorization. Complexity 2017 (2017)
96. Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S., Woodridge, L., Rauer, C., Sen, N., et al.: Cath: increased structural coverage of functional space. Nucleic acids research 49(D1), D266–D273 (2021)
97. Smith, C.L., Eppig, J.T.: The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1(3), 390–399 (2009)
98. Sosa, D.N., Derry, A., Guo, M., Wei, E., Brinton, C., Altman, R.B.: A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020. pp. 463–474. World Scientific (2019)
99. Sterling, T., Irwin, J.J.: Zinc 15–ligand discovery for everyone. Journal of chemical information and modeling 55(11), 2324–2337 (2015)
100. Su, X., You, Z.H., Huang, D.s., Wang, L., Wong, L., Ji, B., Zhao, B.: Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. IEEE Transactions on Knowledge and Data Engineering (2022)
101. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al.: The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research 49(D1), D605–D612 (2021)
102. Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L.J., Bork, P., Kuhn, M.: Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic acids research 44(D1), D380–D384 (2016)
103. Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., Aittokallio, T.: Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling 54(3), 735–743 (2014)
104. Tatonetti, N.P., Ye, P.P., Daneshjou, R., Altman, R.B.: Data-driven prediction of drug effects and interactions. Science translational medicine 4(125), 125ra31–125ra31 (2012)
105. Tweedie, S., Braschi, B., Gray, K., Jones, T.E., Seal, R.L., Yates, B., Bruford, E.A.: Genenames. org: the hgnc and vgnc resources in 2021. Nucleic acids research 49(D1), D939–D946 (2021)
106. Ursu, O., Holmes, J., Bologa, C.G., Yang, J.J., Mathias, S.L., Stathias, V., Nguyen, D.T., Schürer, S., Oprea, T.: Drugcentral 2018: an update. Nucleic acids research 47(D1), D963–D970 (2019)

107. Van Melle, W.: Mycin: a knowledge-based consultation program for infectious disease diagnosis. International journal of man-machine studies 10(3), 313–322 (1978)
108. Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D., et al.: Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. Nucleic acids research 43(D1), D153–D159 (2015)
109. Wang, D., Gu, J., Wang, T., Ding, Z.: Oncomirdb: a database for the experimentally verified oncogenic and tumor-suppressive micrornas. Bioinformatics 30(15), 2237–2238 (2014)
110. Wang, H., Zhou, G., Liu, S., Jiang, J.Y., Wang, W.: Drug-target interaction prediction with graph attention networks. arXiv preprint arXiv:2107.06099 (2021)
111. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al.: Cord-19: The covid-19 open research dataset. ArXiv (2020)
112. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering 29(12), 2724–2743 (2017)
113. Wang, S., Du, Z., Ding, M., Rodriguez-Paton, A., Song, T.: Kg-dti: a knowledge graph based deep learning method for drug-target interaction predictions and alzheimer's disease drug repositions. Applied Intelligence 52(1), 846–857 (2022)
114. Wang, W., Liang, S., Yu, M., Liu, D., Zhang, H., Wang, X., Zhou, Y.: Gchn-dti: Predicting drug-target interactions by graph convolution on heterogeneous networks. Methods 206, 101–107 (2022)
115. Weinreich, S.S., Mangon, R., Sikkens, J., Teeuw, M.e., Cornel, M.: Orphanet: a european database for rare diseases. Nederlands tijdschrift voor geneeskunde 152(9), 518–519 (2008)
116. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(suppl_1), D901–D906 (2008)
117. Wu, X., Duan, J., Pan, Y., Li, M.: Medical knowledge graph: Data sources, construction, reasoning, and applications. Big Data Mining and Analytics 6(2), 201–217 (2023)
118. Wu, Y., Gao, M., Zeng, M., Zhang, J., Li, M.: Bridgedpi: a novel graph neural network for predicting drug–protein interactions. Bioinformatics 38(9), 2571–2578 (2022)
119. Xiang, Y., Zhang, Z., Chen, J., Chen, X., Lin, Z., Zheng, Y.: Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding. arXiv preprint arXiv:2105.07688 (2021)
120. Xie, B., Ding, Q., Han, H., Wu, D.: mircancer: a microrna–cancer association database constructed by text mining on literature. Bioinformatics 29(5), 638–644 (2013)
121. Xing, X., Yang, F., Li, H., Zhang, J., Zhao, Y., Gao, M., Huang, J., Yao, J.: Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. Bioinformatics 38(8), 2178–2186 (2022)
122. Xiong, Z., Huang, F., Wang, Z., Liu, S., Zhang, W.: A multimodal framework for improving in silico drug repositioning with the prior knowledge from knowledge graphs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19(5), 2623–2631 (2021)
123. Yan, S.: Memory-aligned knowledge graph for clinically accurate radiology image report generation. In: Proceedings of the 21st Workshop on Biomedical Language Processing. pp. 116–122 (2022)
124. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)
125. Yang, Y., Cao, Z., Zhao, P., Zeng, D.D., Zhang, Q., Luo, Y.: Constructing public health evidence knowledge graph for decision-making support from covid-19 literature of modelling study. Journal of Safety Science and Resilience 2(3), 146–156 (2021)
126. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 96–104 (2019)

127. Zhang, X.M., Liang, L., Liu, L., Tang, M.J.: Graph neural networks and their current applications in bioinformatics. Frontiers in genetics 12, 690049 (2021)
128. Zhang, X., Che, C.: Drug repurposing for parkinson's disease by integrating knowledge graph completion model and knowledge fusion of medical literature. Future Internet 13(1), 14 (2021)
129. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)
130. Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E.F., Yang, Y., Niu, Z.: Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in bioinformatics 22(4), bbaa344 (2021)
131. Zhu, C., Yang, Z., Xia, X., Li, N., Zhong, F., Liu, L.: Multimodal reasoning based on knowledge graph embedding for specific diseases. Bioinformatics 38(8), 2235–2245 (2022)
132. Zhu, Y., Elemento, O., Pathak, J., Wang, F.: Drug knowledge bases and their applications in biomedical informatics research. Briefings in bioinformatics 20(4), 1308–1321 (2019)
133. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), i457–i466 (2018)
134. Zou, X.: A survey on application of knowledge graph. In: Journal of Physics: Conference Series. vol. 1487, p. 012016. IOP Publishing (2020)

Ylenia Galluzzo is PhD Student in Information Communication Technology (ICT) at the University of Palermo since November 2022. My main research interests are on Knowledge Graphs, Big Data Analytics, ML, Deep Learning and Bioinformatics. I deal with methodologies for modelling and analysing data through complex networks, in the field of Big Data. Previously, I worked in the industrial sector in an R&D team where I developed skills in the field of: Natural Language Processing (NLP), Anomaly Detection and methods Supervised and Unsupervised for Topic Modeling.

Received: June 01, 2023; Accepted: May 02, 2024.