EXPERT
OPINION
ON DRUG DISCOVERY
informa
2.297

Taylor & Francis
Taylor & Francis Group

Expert Opinion on Drug Discovery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iedc20

Knowledge graphs and their applications in drug discovery

Finlay MacLean

EXPERT OPINION ON DRUG DISCOVERY
https://doi.org/10.1080/17460441.2021.1910673

Taylor & Francis
Taylor & Francis Group

REVIEW

Knowledge graphs and their applications in drug discovery

Finlay MacLean

Target Identification., BenevolentAI, United Kingdom of Great Britain and Northern Ireland

ABSTRACT

Introduction: Knowledge graphs have proven to be promising systems of information storage and retrieval. Due to the recent explosion of heterogeneous multimodal data sources generated in the biomedical domain, and an industry shift toward a systems biology approach, knowledge graphs have emerged as attractive methods of data storage and hypothesis generation.
Areas covered: In this review, the author summarizes the applications of knowledge graphs in drug discovery. They evaluate their utility; differentiating between academic exercises in graph theory, and useful tools to derive novel insights, highlighting target identification and drug repurposing as two areas showing particular promise. They provide a case study on COVID-19, summarizing the research that used knowledge graphs to identify repurposable drug candidates. They describe the dangers of degree and literature bias, and discuss mitigation strategies.
Expert opinion: Whilst knowledge graphs and graph-based machine learning have certainly shown promise, they remain relatively immature technologies. Many popular link prediction algorithms fail to address strong biases in biomedical data, and only highlight biological associations, failing to model causal relationships in complex dynamic biological systems. These problems need to be addressed before knowledge graphs reach their true potential in drug discovery.

ARTICLE HISTORY
Received 10 January 2021
Accepted 26 March 2021

KEYWORDS
Biomedical knowledge graphs; drug repurposing; drug repositioning; heterogeneous information networks; graph machine learning; network embeddings; knowledge graph embedding; network pharmacology; network medicine

1. Introduction

Despite an explosion of data being generated, significant scientific and technological advancements, and industry initiatives on efficiency, the pharmaceutical industry has suffered a 'drug drought', in which investment in research and development (R&D) dramatically increased [1] without a marked increase in annually approved drugs [2]. Whilst the number of new chemical entities to reach the market has steadily increased over the last decade [2], this has been accompanied by a marked increase in R&D expenditure [1]. Business value is derived principally from research and development (R&D) yielding a positive return on investment. The return on investment on R&D for the top 12 pharmaceutical companies fell from 10% in 2010 to 2% in 2019 [3], and the cost to develop a drug rose almost two-fold to 2 USD billion USD [3]. The pharmaceutical industry is increasingly looking toward new disruptive methods to reduce the failure rate, increase the speed of development, and ultimately reduce the cost of research.

1.1. The decades of data

In the past two decades, there has been an explosion of data generated in the biomedical domain. Motivated by the landmark *The Human Genome Project* [4] earmarking the beginning of the genetic revolution, subsequent projects, such as the *1000 Genomes Project* [5] and the UK Biobank [6] have provided comprehensive analyses of the human genome. Next generation sequencing (NGS) technologies have dramatically lowered the cost of sequencing, and both genome-wide and phenotype-wide association studies are now frequently used to connect underlying genetic variation to complex phenotypic traits. Public repositories such as the *Sequence Read Archive* [7] and *European Nucleotide Archive* [8] have provided researchers with access to large databases of DNA sequencing data. Combined with the use of expression quantitative trait locus studies, technologies such as Mendelian randomization and colocalization have helped to identify causal loci and genes. In transcriptomics, the *Human Protein Atlas* [9] have generated atlases of gene expression of tissues, the brain, diseases, blood and cells, whilst the *Genotype-Tissue Expression (GTEx)* [10] project quantified expression over common tissues. The *Library of Integrated Network-Based Cellular Signatures (LINCS)* project collected gene expression profiles in response to genetic and chemical perturbagens [11]. In oncology, projects such as *The Cancer Genome Atlas* [12] and *Cancer Cell Line Encyclopedia* [13] have provided detailed molecular characterizations of cancers, whilst RNA-interference and CRISPR-Cas9 technologies identified genetic dependencies in cancers [14]. In epigenomics, the *Human Epigenome Atlas* [15] has cataloged genome-wide epigenetic markers in all major tissues. Multiple databases have documented non-coding RNAs and their regulatory targets [16,17], and their association with diseases [18].

1.2. The era of systems biology

This immense accumulation of biomedical data led to a paradigm shift in the pharmaceutical industry, moving

CONTACT Finlay MacLean finlay.maclean@benevolent.ai BenevolentAI, 4-8 Maple St, Bloomsbury, London W1T 5HD, United Kingdom of Great Britain and Northern Ireland.

Article highlights

- Knowledge graphs provide an elegant solution to the 'data problem' in the pharmaceutical industry, integrating and harmonizing the ever-growing number of multimodal data sources.

- Representing biological systems as knowledge graphs has allowed for the exploitation of graph theory and powerful graph machine learning methodologies, well suited to the target-based systems biology approach to drug discovery.

- The most common application of knowledge graphs in the pharmaceutical industry is in early stage drug discovery and repurposing, particularly in identification of pathogenic genes and drug targets.

- Biomedical knowledge graphs have yielded noteworthy repurposing candidates for COVID-19 directly leading to clinical validation and emergency use authorization.

- The predictive power in many graph machine learning techniques comes mainly from connectivity, and not network proximity, introducing a significant bias in link prediction tasks.

- This connectivity bias is further exacerbated when training on literature-derived knowledge graphs whose degree distribution diverges from that of the underlying biological system. Mitigation strategies are needed.

This box summarizes key points contained in the article.

from phenotype-based discovery to a target-based approach. The previous phenotypic approach paid less attention to the mechanism of action of the drug, focusing more on the desired phenotypic outcome. Even in Phase I on clinical trials, this approach was unable to determine the mechanism of action of a drug [19]. The collective zeitgeist of the industry moved toward a systems biology-focused approach, and with it the fundamental aim of drug discovery changed. It now aimed to first understand the complex biological systems within our cells and how their dysregulation leads to disease, and finally develop methods to selectively target these systems.

Systems biology describes the computational modeling of molecular systems, drawing from many disciplines including computer science and physics. One of the fundamental principles of biological life is that the accumulation of simple, locally-acting components leads to complex structures and systems. Systems biology follows this principle, describing a complex biological system as a network of simple biological components (analogous to those in electronic circuits) and simulates how the system changes in response to certain stimuli. These systems consist of intra- and intercellular interactions amongst molecules that govern biological functions, and whose dysregulation leads to disease. Their networks often contain multi-scale elements, ranging from molecular components to tissues; both physical and abstract entities, ranging from proteins to phenotypic outcomes. Networks also contain diverse types of interactions between entities, such as inhibitions, activations, associations and causal interactions.

1.3. Knowledge graphs

These technological and scientific advancements have undoubtedly deepened our understanding of human biology and disease. But why has this innovation not been reflected in an increase in profitable R&D in pharmaceutical companies? Some speculate that the so-called low-hanging fruit of drug discovery have been picked [19]. A tightening of drug safety regulations in response to the thalidomide tragedy have certainly upped the criteria in receiving approval for a drug [20]. One possible answer is simply how difficult it is to ingest, harmonize and use the multitude of data that has been generated. Despite significant initiatives to 'digitally transform' *Novartis*, their CEO, Vas Narasimhan, has remarked on the difficulty to clean and link their heterogeneous data [21]. This should not be read as an individual failure of Novartis, rather, it reflects the difficulties faced by the entire industry. The difficulties in data integration are further reflected in the industry-academic initiatives to standardize scientific data such as the *FAIR (Findability, Accessibility, Interoperability, and Reusability) Data Principles* [22].

Knowledge graphs (KGs) have proven to be attractive methods to store biomedical data due to their capacity to model complex data structures. Observing Figure 1, one can clearly see the parity between the biological pathway [23] and its approximation in a graph database. Formally a KG can be described as a labeled multi-graph [24]. The graph consists of entities; commonly referred to as nodes or vertices, and relationships connecting two entities; commonly referred to as edges, facts, or links. The two nodes constituting an edge or relation are often respectively called the head and tail or source and target nodes. Across many diverse domains and industries, KGs have been used in question-answering,

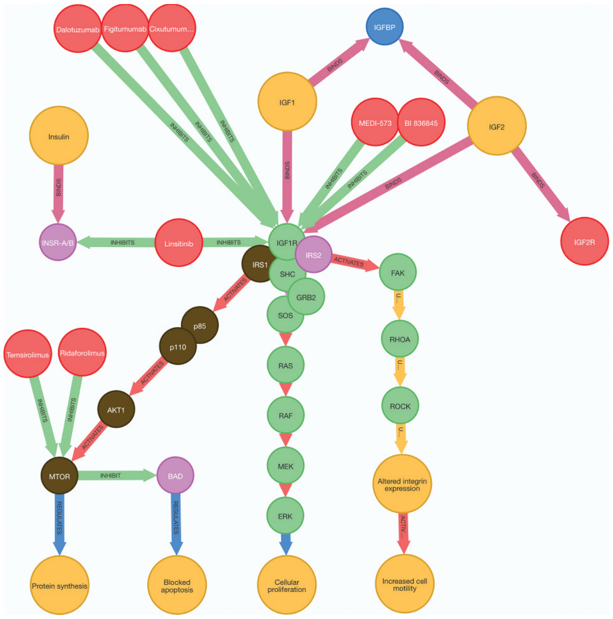

Figure 1. Representing biology as a heterogeneous information network.

A KG representation of the canonical insulin receptor signaling cascades using the Neo4j graph database [27]. Image was created by querying a KG to reconstruct the signaling pathway of Iams *et al.* [23].

recommendation, and information retrieval systems. Arguably the most famous of commercial question-answering systems is *Watson*, developed by IBM to beat human experts at the quiz show *Jeopardy* [25]. In terms of recommendation systems, *Pinterest* have famously used a KG of *user-likes-pin* to recommend new pins to their user-base [26].

In the field of drug discovery, one of the earliest notable attempts to integrate multiple structured biomedical databases was the work of Himmelstein *et al.*, developing Hetionet to prioritize drugs for repurposing [28**] and genes associated with disease [29]. Other KGs include *OpenBioLink*, principally used to benchmark link prediction models [30], and the work of Womack *et al.* [31]. Whilst the integration of structured databases has proven its utility, others have derived biological relationships from literature. The *Global Network of Biomedical Relationships* [32*] screened 24 million research articles to create a disease-gene-chemical KG consisting of 2 million thematically-labeled edges. Biomedical KGs can contain a multitude of multimodal data spanning transcriptomics, proteomics, genomics, phenomics, drug pharmacology, chemistry, and ontological information. The schema of the *Drug Repurposing Knowledge Graph* [33] exemplifies the heterogeneity of data common in KGs for drug discovery. The majority of large-scale biomedical KGs are based on semantic web technologies, the largest of which is *Bio2RDF* [34]. One of the defining features of semantic KGs is their extensibility, as demonstrated by projects, such as *Chem2Bio2RDF*, which combined *Bio2RDF* with a chemogenomic semantic graph [35].

There seems to be no agreed-upon definition of a KG. Some constrict its name to only literature-derived graphs. Others even go further, using KG to refer only to graphs that use semantic technologies to represent the data. In this article, we define a KG as any heterogeneous information network, regardless of the technology used and the provenance of the data it represents.

1.4. Graph machine learning on knowledge graphs

Marshall Nirenberg famously stated that science progresses best using simple assays to rapidly generate large data sets [19]. Whilst large-scale and genome-wide screens are certainly the gold standard of systematic drug discovery, their high costs often prohibit their use only for all but the most common (and thus profitable) of diseases. Machine learning has demonstrated its potential as a complementary approach: rapidly and inexpensively generating data in unexplored areas of the biological and chemical space. In particular, graph-based machine learning (GML; also referred to as geometric machine learning) methods have shown promise in this field. By representing biological systems as KGs, it has allowed for the exploitation of graph theory and powerful network science algorithms; drawing new insights into this otherwise silo-ed data. GML has been applied to systemically screen compounds for new interactions, and shed light upon areas unknown of the human interactome.

GML uses the topological structure of the network to classify properties of nodes, predict the existence of edges, and detect communities [36*]. It assumes that functionally or structurally similar nodes will be more highly connected within the graph. In a disease-gene association network, comorbidities will share a higher number of associated genes, inferring their functional similarity. In contrast, two diseases that have no intersection of associated genes are unlikely to exhibit functional homogeneity. Traditional approaches such as Common Neighbor, Jaccard's Index, Adamic/Adar Index and Katz compute similarities between nodes based on local neighborhoods [37], however failed to utilize the global neighborhood and topology of the graph. In recent years, embedding-based GML has become the norm. Embedding strategies encode the continuous neighborhood information of a node, graph substructure, or entire graph into a discrete low-dimensional latent vector [38]. To refer back to the previous exemplary disease-gene association network, a comorbidity of two diseases will encode both diseases with embeddings (vectors) that are mathematically similar in latent vector space, since they are proximal within the network. For a comprehensive survey of graph embeddings, we defer the reader to these comprehensive reviews [34*] [38]

Embedding strategies have now developed to encode heterogeneous graphs, often referred to as knowledge graph embeddings (KGEs). Aside from representing nodes as latent vectors, a low-dimensional representation of each relationship type is also learned. A wonderful myriad of methods have now been employed to generate KG embeddings. A review of these methodologies is outside the remit of this manuscript. We refer the reader to the comprehensive review of Rossi *et al.* [24] and repository of KGE models [39].

2. Applications of knowledge graphs

KGs have emerged as an effective method of information representation in drug discovery. Modeling biological systems as graphs has facilitated the use of powerful network-based algorithms; encoding the continuous global or local neighborhood of nodes into discrete latent vectors. The vectors are then used in a downstream machine learning task. Supervised downstream tasks include link prediction (pairwise prediction between two nodes), and node classification (classification of one node). Embeddings may be also used in unsupervised tasks such as community detection (the detection of neighborhoods of nodes via clustering). These methods make the guilt-by-association assumption; that functionally or structurally similar biological entities are likely to have similar properties, have high network proximity, and have a small distance between node embeddings in vector space.

2.1. Drug repurposing: a link prediction task

The overwhelming majority of applications using KGs are framed as link prediction tasks. Our knowledge of biology is incomplete, and the resulting information networks are sparsely populated. Link prediction on incomplete networks aims to systematically complete these networks, in which predicted edges represent biological interactions or associations that are currently unknown, unexplored or yet to be validated. Figure 2 illustrates the training process of a link prediction KG embedding model.

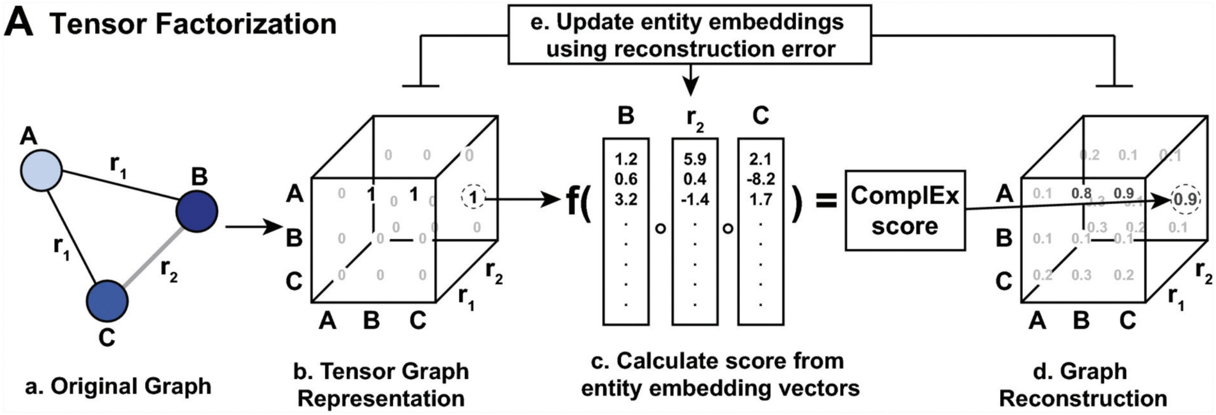

Figure 2. Prototypical link prediction training process.

Tensor factorization KG embedding model reproduced from Paliwal *et al*. [49] Step (a) shows the original graph consisting of three entities (A, B, and C), and two relations (r1 and r2). Step (b) shows the latent vector representation (embeddings) of the nodes and relations. Step (c) shows a downstream scoring function of a triplet of source node, relation, and target node embeddings. Step (d) shows the predictions of edge existence for all edges in the original graph.

Of all of the applications of KG-based link prediction within the field of drug discovery, one of the most promising is drug repurposing (DR). The aim of DR is to identify new indications and conditions for existing drugs. Framed as a systems biology problem, the aim is to predict the likelihood of edges within a biological graph. With only 10% new molecules reaching the clinic [40*], drug repurposing has proven to be a lucrative method of drug discovery; mitigating a proportion of the risk by focusing on drugs that already have established safety profiles. In recent years, almost one-third of the drugs that receive approvals are repurposed [41]. Many of the most successful repositions have been largely serendipitous: their unplanned side effects fortuitously provide benefit to other patient populations with other conditions [41]. Examples of repurposed drugs include the dihydrotestosterone inhibitor, finasteride. Finasteride was originally developed for treatment of prostate cancer and showed moderate efficacy [42] however after hair growth was noted on laboratory rats, the drug was repositioned for treatment of androgenetic alopecia [43]. Whilst finasteride serves as a poster child for how side effects can be beneficial, it also serves as a stark reminder that the majority of side effects are maleficial. Systemic inhibition of dihydrotestosterone, for example, has been associated with permanent sexual dysfunction and cognitive impairment [44]. Unlike it’s serendipitous counterpart, network-based DR has the potential to differentiate between beneficial and maleficial phenotypic outcomes; intelligently and systematically identifying new indications for existing drugs. Multiple approaches encompass network-based DR; on-target repurposing, off-target repurposing and target-agnostic repurposing. Each approach predicts a different relation between drug, gene and disease in a KG (see Figure 3).

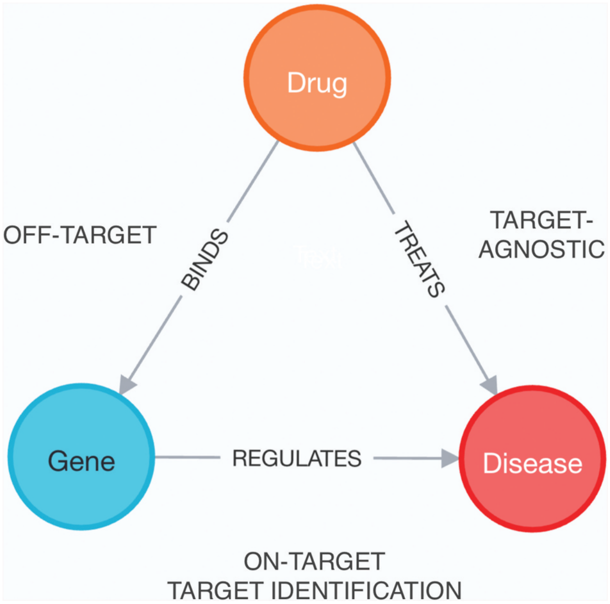

Figure 3. Approaches to network-based drug repurposing and discovery.

Edges of multiple types can be predicted to indicate the therapeutic viability of a repurposable drug to treat a disease. Different edge types correspond to different approaches to drug repurposing. On-target repurposing describes the prediction of novel therapeutic genes, which are the known on-targets of drugs. In off-target repurposing, the off-targets of a drug are predicted, one of more of which will regulate a disease. In target-agnostic repurposing, the gene through which the drug acts is not explicitly provided.

2.1.1. Target-agnostic drug repurposing

To identify repurposable drug candidates for new indications, many methods predict drug-treats-disease edges in pharmacological KGs. One example was the work of Himmelstein *et al*. who applied a degree-normalized pathway model to highlight repurposable drugs for epilepsy. The model was applied to their *hetionet* KG, consisting of genes, diseases, tissues, pathophysiologies and multimodal edges [26**]. Biomedical data in these graphs is sparse. To overcome the sparsity problem, Poleksic developed a compressed sensing technique, demonstrating superior performance over the original pathway-based implementation [45]. One of the drawbacks of pathway-based approaches is the high computational cost, limiting their use to relatively small KGs. Womack *et al*. demonstrated how *node2vec*, a popular random-walk method, was more performant and with significantly lower computational overheads [31]. KGEs have also been applied to this prediction task, including Sosa *et al*. whose model exploited the confidence scores of edges in a literature-derived KG [46]. In target-agnostic drug repurposing, neither the drug target nor disease associated genes are implicitly provided to the models and thus obfuscate the mechanism of action of the drug.

2.1.2. On-target repurposing and target identification

In contrast to target-agnostic approaches, target-based drug repurposing approaches are attractive alternatives. Diseases can be described as phenotypic manifestations caused by genomic perturbations. These genomic perturbations cause further dysregulation of genes and pathways. The aim of target-based drug discovery is to develop a compound to either directly or indirectly target one of these disease-causing or disease-associated genes. Similarly, the aim of on-target drug repurposing is to identify a preexisting drug to target one of these disease-causing or disease-associated genes. Both methods require one or more targets through which to act. GML methods have been widely applied to prioritize pathogenic genes. Himmelstein *et al*. applied the previously cited hetionet KG to predict disease-causing genes for multiple sclerosis [29]. Even embedding strategies trained on relatively small heterogeneous networks have demonstrated their superiority over baseline approaches, such as Xu *et al*. who trained a multipath random walk model on a network of gene-phenotype, protein–protein interactions and phenotypic similarities [47].

KGE approaches have recently been applied to target identification. Pitalla *et al*. used a relation-weighted RotatE model to predict drug targets for Parkinson's disease [48]. Notably, the model outperformed *OpenTargets*, the leading initiative for target identification which includes genetic, pharmacologic, pathway, multi-omics data. Similarly, Paliwal *et al*. developed *Rosalind*, a tensor factorization-based KG embedding trained on *BenevolentAI*'s proprietary biomedical knowledge to predict therapeutic targets for rheumatoid arthritis [49]. Rosalind outperformed *OpenTargets* alongside other GML approaches. Top predicted genes were experimentally validated in an in vitro assay using patient-derived cells. Five genes were determined to be promising for further preclinical research.

The utility of network-based approaches to target identification is well demonstrated by the above (both academia and industry-driven) projects. The above methods are focused on identifying pathogenic protein-coding genes. Whilst protein drug targets remain the central focus of drug discovery, the role of non-coding RNA (ncRNA) in disease is becoming more apparent [50]. Researchers have started to exploit KGs and GML to predict ncRNA-disease associations. Ji *et al*. constructed a KG consisting of micro-RNA, circular-RNA, long non-coding RNA (lncRNA), proteins and diseases, and used a matrix factorization embedding model to predict miRNA-disease associations for three common cancers [51]. GML has also been applied to predict lncRNA-disease associations. In a similar project, Zhou *et al*. built a KG similar to that used by Ji *et al*., and trained a higher-order preserving matrix factorization model [52]. They validated their model by predicting disease-related lncRNAs for three excess death rate cancers.

2.1.3. Off-target repurposing and drug-target interaction

The *one drug, one gene, one disease* paradigm of drug discovery has passed. Many diseases are now understood to be multifactorial; caused by the combination of the effects of multiple genes. Most drugs are now estimated to bind to between 10 and 100 targets [48]. Polypharmacology is a promising paradigm in drug discovery, assuming drugs act through multiple genes associated with one or more patho-mechanism. Our knowledge of the pharmacogenomic space is sparsely populated [53], mainly limited to the disease-associated genes of interest, and genes common in safety panels (essential genes and those associated with undesired phenotypes). Genome-wide screens would be of great utility to understand the polypharmacology of existing drugs (off-target drug repurposing), and novel compounds (drug discovery). Due to the cost of experimental screens, many researchers have developed *in silico* methods to quickly and inexpensively screen for drug-target interactions. Link predictions methods have been used to predict drug-binds-target edges in pharmacological KGs. A random walk-based approach, *DTINet*, was used to identify the novel inhibitory action of three approved drugs

on cyclooxygenase proteins [54]. The pharmaceutical industry is often interested in determining chemicals with little to no available interaction data: the so-called cold-start problem. Lim *et al*. developed a collaborative filtering approach tailored specifically to chemicals with few interactions [55]. In addition to the necessary pharmacogenomic data, network-based approaches have used genomic, chemical, pharmacological [54,56], side effects [57], diseases, pharmacokinetics, and proteomics [58]. KG embeddings have also been applied, screening approved drugs for off-target interactions with COVID-19 associated genes [33].

Many computational methods have been developed to perform *in silico* screens. The main advantage of network-based approaches is they do not require the 3D structure of the protein. Deep learning approaches such as *DeepPurpose*, using only the primary amino acid sequence of the protein and SMILES string of the molecule, have shown to be competitive methods of DTI prediction [59]. If the 3D structure of the protein is known, molecular docking studies provide an unparalleled level of information of how a drug interacts with the binding pocket of a protein. Deep learning has also been successfully applied to molecular docking, as exemplified by the commercialization of *Atomnet* by *Atomwise* [60]. Until recently, these approaches were limited to proteins with a known 3D structure. A deep learning method AlphaFold, that uses amino acid sequence as input, has recently demonstrated accuracy comparable to experimental techniques such as X-ray crystallography [61]. This may widen the screening possibilities of structure-based approaches, overshadowing network-based approaches to DTI prediction.

2.1.4. COVID-19: A case study in network-based repurposing

Unlike it's serendipitous counterpart, network-based drug repurposing is still waiting to see its first compound reach the market.

True validation that this method is an effective tool in drug discovery will come only once drugs identified are approved for their new indication, and a systematic review of the methodology has been conducted. Arguably, the most mature and substantial efforts to identify repurposable drugs have been focused on finding a therapeutics to target the SARS-CoV-2 coronavirus or treat the associated COVID-19 disease.

Network-based methods need a network on which to train, in this case capturing data pertaining to SARS-CoV-2. Reese *et al*. [62] integrated multiple structured databases into their biomedical KG. Next, they integrated datasets pertaining to COVID-19 (Zhou *et al*. [63], CORD-19 [64]). Ioannidis *et al*. [33] combined the preexisting KGs [26**], [30*] with additional databases. To predict the likelihood that an approved drug would treat COVID-19, the researchers trained a TransE embedding model on the KG, and then computed the distance scores between approved drugs and COVID-19 and similar coronaviruses, and drugs and COVID-19-associated genes. Hseih *et al*. extended the KG of Ioannidis *et al*. [33], integrating a SARS-CoV-2-specific graph into the original graph via transfer learning [65]. To discover drugs that can functionally target SARS-CoV-2-associated host genes, protein and drug embeddings were used to predict therapeutic effectiveness.

A multitude of network medicine approaches have been applied to identify existing drugs to palliate or treat COVID-19. The majority of these studies focus on predicting drugs that prevent viral entry (targeting viral genes), viral replicative mechanisms, or suppression of the host inflammatory response (both targeting host genes). One of the most noteworthy efforts was produced by Gysi *et al*. [66], who integrated host-host, host-viral, and host-drug protein interaction networks, using an ensemble of predictive models to predict 81 potential candidates to treat SARS-CoV-2. Their method successfully predicted that the SARS-CoV-2 could manifest in brain tissue and have neurological comorbidities, which have since been validated [67]. Other researchers have similarly used network proximity of drug targets to viral proteins in protein–protein interactomes [63]. *BenevolentAI* used a proprietary literature-derived KG to identify baricitinib, a drug used to treat rheumatoid arthritis, to treat patients with bilateral COVID-19 pneumonia [68]. Since then, more than 12 clinical trials have been conducted, including by the drug's proprietor, *Eli Lilly*. The janus kinase inhibitor has now been granted emergency use authorization (EUA) from the FDA for the treatment of hospitalized patients with COVID-19 [69]. The methodology employed by *BenevolentAI* is yet to be published, and drug authorization of baricitinib is only temporary. True validation for the use of KGs in drug repurposing will come only after (i) the FDA approval of drugs surfaced through KG and GML methods, and (ii) such methods have been subjected to the scientific method. Nevertheless, the numerous aforementioned studies certainly suggest that the inclusion of KG-based methods into drug discovery workflows could be of great benefit.

2.2. Additional link prediction applications

The utility of the method is reflected in its adoption in industry. Whilst the above applications of both drug repurposing and early drug discovery have been employed in industry, many other applications remain as academic research projects. KGs and GML techniques have been widely used in academia and applied to further pharmacological and multi-omic prediction tasks, including prediction of protein–protein interaction [70–72], polypharmacy side effect [73], disease side effects [74], and drug–drug interactions [75]. An exhaustive summary of all of these is out of the remit of this review. We refer the reader to the review of Su *et al*. [76].

2.3. Node classification applications

Whilst most applications of KGs in drug discovery are framed as link prediction tasks, node classification has also demonstrated its utility. Node classification describes the process in which a model is trained on features derived from a

subset of nodes with a labeled property, and subsequently predicts the likelihood that unlabeled nodes possess this property. To the best of our knowledge, there are few examples of the adoption of these methods by industry or public biological databases, however demonstrate the diverse range in which KGs can be applied to drug discovery.

2.3.1. Protein function

Understanding protein function is one of the earliest prerequisites in the drug discovery process. Proteins with similar sequences tend to exhibit similar functions [77]. Also, proteins with similar sequences tend to interact with similar proteins within protein–protein interaction (PPI) networks. Thus, protein nodes with high network proximity tend to share protein function. The seminal paper for the *node2vec* model [78], a semi-supervised random walk model, showed how node embeddings and a downstream multi-label classifier could be effectively used to predict protein function, using the *BioGRID* PPI network [79]. Whilst *node2vec* used only a homogeneous network of PPIs, others have extended their work, including other forms of both graphical and complementary information. *DeepGo* uses both sequence and PPI networks to generate features for each protein [80]. Nariai *et al*. demonstrated that the integration of PPI networks, gene expression, protein motif information, gene knockout phenotype data and protein localization information yielded greater performance than homogeneous PPI networks [81]. Proteins do not execute all of their functions at all times, and in all tissues in which they are expressed [82]. Researchers developed *OhmNet*, using *node2vec* to generate embeddings based on tissue-specific PPIs, demonstrating improved performance over methods employing tissue-agnostic networks.

Node classification is far from the only computational method for predicting protein function. Protein function is directly related to its 3D structure. A deep learning method *AlphaFold*, that uses amino acid sequence as input, has recently demonstrated accuracy comparable to experimental techniques such as X-ray crystallography [61]. Such advances may facilitate the overshadowing of network methods by deep learning methods which learn functions from 3D topology.

2.3.2. *Essential genes*

PPI networks underpin the majority of intracellular communication. Many researchers have utilized topological properties of these networks to identify the most important key regulators. Hub genes; the genes most important within a submodule, are often selected on the basis of node degree (also called degree centrality). It is now widely reported that hub genes correspond to essential genes [83]. Essential genes have more recently been associated with other topological characteristics, such as high betweenness centrality [84], a measure of the centrality of a node between submodules. If one imagined a network resembling an hourglass, it is likely that nodes with high degree centrality would exist within the top and bottom 'glass' bulb subnetworks, whereas genes with high betweenness centrality are those close to the bottleneck between bulbs, acting as the detrimental linchpin in communication between submodules.

Based on the assumption that essential genes are topologically distinct from their non-essential counterparts, researchers demonstrated how PPI networks can be used to predict the essentiality of genes [85]. Node embeddings based on PPI networks were generated via a biased random walk. A binary classifier was trained on existing known essential genes, using node embeddings as features. To explore the biological functions of the essential genes, they clustered genes by their node embedding, and performed GO functional enrichment analysis on the genes in these clusters. They found notable correlation with known important processes such as RNA splicing, ribosome biogenesis and golgi vesicle transport.

2.3.3. *Antitumor activity*

Many foods are known to be rich in compounds with antitumor activity [86], however the cancer suppressive or preventative potential of all compounds has not been assessed experimentally. Veselkov *et al*. computed node embeddings for compounds by using random walks on a human PPI network [87]. Compounds were represented within the protein embedding space, by using the known targets of the respective compound as starting nodes for the random walk. Using known anticancer drugs as the training set, a support vector machine classifier was used to predict which food compounds were candidates for cancer prevention or treatment.

2.4. *Neighborhood detection applications*

Neighborhood or community detection describes the process of clustering the latent vectors of nodes. Following the *guilt by-association* principle, clusters in the high dimensional latent space of the embeddings correspond to biologically relevant communities within the KGs.

2.4.1. *Visualization as validation*

It is commonplace in graph machine learning for researchers to use neighborhood detection to validate the biological relevance of their model. The majority of embedding approaches encode functionally or structurally similar nodes close to each other in the embedding space. By clustering the embeddings in high dimensional space, and mapping to 2D space via dimensionality reduction, researchers are able to visualize nodes of interest, and identify to what extent their model has successfully encoded the biological network. The model *gene2vec* [88] generated embeddings based on pan-genome gene co-expression and clustered using t-SNE. Coloring nodes by gene expression, the clustered graph successfully identified tissue-specific gene clusters. Goh *et al*. created a bipartite graph of diseases and their genetic associations, and then performed functional clustering of the diseases to create the human disease network. Neoplastic diseases constituted the largest cluster, with notable clusters of comorbidities such as diabetes and obesity, hypertension, asthma, and atherosclerosis [89] (see Figure 4).

2.4.2. *Cancer driver gene detection*

Another notable application of neighborhood detection has been in the detection of cancer drivers. Cantini *et al*. used a consensus model of 5 popular community detection algorithms to identify communities in transcription factor and miRNA co-targeting networks, PPI and gene co-expression networks for both tumor and healthy tissue [90]. Next, they compared communities detected in the multiple tumor and healthy tissue networks to identify both genes and associated functions only prevalent in cancer tissue. These candidate cancer drivers included known oncogenes and potential new oncogenic drivers.

3. The inherent biases in link prediction

3.1. *Degree bias and literature bias*

One of the most challenging problems in link prediction in biological networks is mitigating the inherent bias caused by the degree distribution of a graph being unrepresentative of the underlying distribution of the network. In contrast to a bias-free link prediction model, which infers the likelihood of an edge by the proximity of nodes within the network, a model biased by degree infers the likelihood of an edge by the connectivity of nodes. A biased model would predict a disease and gene are connected based solely on how many independent connections each node has in the graph, without considering the biological rationale; how well-connected the disease and gene are to *each other*. The magnitude of this problem is relative to the disparity between degree distribution of the training graph and degree distribution of the true underlying network. Literature bias is often responsible for this over-representation of a handful of well-researched nodes, causing the degree distribution to not represent the true distribution of the underlying biology.

Most embedding-based models applied to link prediction tasks are based on the *guilt-by-association* assumption; that structurally or functionally similar nodes will be encoded near each other in the *n*-dimensional space of the embedding. These methods assume that the downstream classifier or score function (see Figure 2) use network proximity as the driving force behind the model's predictive power. In reality, the overwhelming majority of predictive power comes predominantly from node degree-based features, with local topology attributable to a small portion of performance [91**]. In

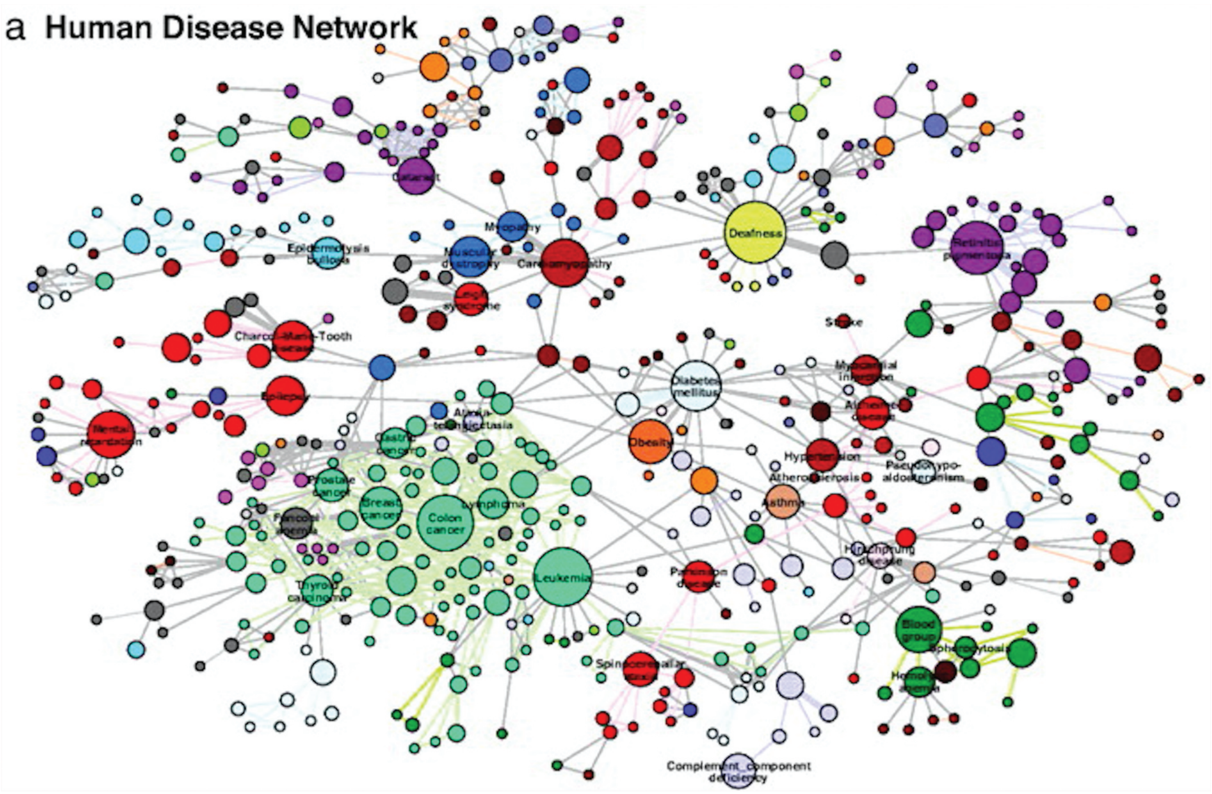

Figure 4. Visualizing biological networks.

Illustration of the *human disease network* constructed by Goh *et al.* [89]. Nodes represent Mendelian traits and disorders, and edges indicate that the two diseases share at least one genetic association. Reproduced in part from [89] with permission of PNAS. Copyright (2007) National Academy of Sciences, U.S.A.

other words, nodes which are highly connected are statistically more likely to be connected to other nodes, and thus have a higher prior probability of connection. To clarify, in a drug-treats-disease prediction task, imatinib would be highly predicted to treat diabetes type 1, mainly due to their connectivity, whilst the probability that marizomib treats diffuse intrinsic pontine glioma would be low, due to the rarity of both disease and drug. This issue is especially problematic, when the prediction task is focused on predicting edge existence for low-degree nodes, for example, drug repurposing for rare diseases.

Training a model that depends on degree bias is not inherently problematic. If one imagines a PPI graph that accurately represents the biological system, hub genes have a higher connections, and thus are more likely to be connected to any other protein at random. A simplistic model using degree-derived features would accurately predict the connectivity of this network. The problem lies, however, when the degree distribution diverges from that of the underlying biological network. In this example, nodes that are over-represented in the graph will receive high probabilities, despite a lack of biological evidence that the node is central in the network. Simplistic models that depend on node degree fail to overcome this disparity between network and graph. In contrast, models that use network proximity are considerably more robust, and predictions are less likely to vary when the training graph distribution differs from the biological network. As we often do not know the ground truth, one can simply test a model on an external graph with a different biological distribution (but captures the same information). Researchers [26**] demonstrated how a degree-biased model achieved a near perfect score (AUROC of 0.979) when tested on the same graph distribution that it was trained on (a *drug-treats-disease* graph based on *DrugBank* [92]). However, when tested on a separate graph (a *drug-treats-disease* graph based on *DrugCentral* [93]), the model achieved a score close to random (AUROC of 0.541). In contrast, their model that utilized network proximity achieved similar scores across both graphs (AUROC of 0.974 and 0.855 on *DrugBank* and *DrugCentral*, respectively).

Literature-derived networks often have vastly different degree distributions to that of the underlying network. They rely on the extraction of biological relationships stated in academic manuscripts. The connectivity of nodes in these networks is therefore largely governed by the quantity of research that is conducted in that area, and not by the underlying biology. For example, more than three quarters of protein research focuses on the 10% of proteins that were known before the genome was mapped, even though many proteins have now been linked to disease [94]. There is no discernible biological difference between well-researched and under-researched genes, save their level of research interest and tool availability. Studies have shown that therapeutic opportunities [95] and essentiality [96] of well-researched genes are similar to those in the unstudied 'dark' genome. Systematic genome-wide screens effectively eliminate research bias, and provide a much more accurate proxy for true distribution of biological networks. Comparisons between networks based on systematic screens and literature highlight the disparity between literature-derived KGs and the biology they aim to represent [91**]. This example should heed as a warning to researchers or machine learning practitioners performing link prediction on graphs. One must stay vigilant to this bias, as it is often possible to achieve seemingly satisfactory results based solely on the degree of the network; without considering the topology of the network nor the similarity of nodes. *[paragraph removed – see comments]*

3.2. Mitigation strategies

These results highlight the need for effective mitigation strategies to remove the reliance on degree-based features; instead encouraging models to learn network topology to infer the existence of edges. Whilst far from commonplace in literature, researchers have now addressed the problem of degree bias, mitigating the bias at distinct points in the model training process.

Multiple embedding strategies exist to encode node neighborhoods. Researchers have applied a degree penalty to prevent over-representation of high degree nodes using a random walk embedding and skip-gram model [97]. Others have used a Bayesian method, explicitly providing the prior probability alongside the adjacency matrix to the model; preventing the encoding of node degree in the embeddings [98,99]. Many graphs have only positive edges, and negative edges are created by either randomly sampling an edge from all possible non-edges (node pairs without a connection), or by corrupting a node in a positive edge, replacing a node at random. Importantly, it has been shown by uniformly sampling, positive and negative samples do not have the same degree distribution. By sampling nodes from the global degree distribution, models are forced to differentiate edges by their network proximity, and not their network connectivity, ultimately leading to improved and less biased performance [100,101]. Whilst the above methods try to prevent information of node degree from being provided to the model, some researchers have removed reliance of degree by doing precisely the opposite. By explicitly providing the prior probability alongside network-based features to the model during training, the model relies on the degree-based features and does not learn them. During testing, a uniformly connected network is assumed, and a uniform prior is provided in place of the biased prior, yielding bias-free predictions [26**] [98]

4. Expert opinion

KGs have shown great promise in drug discovery providing an answer to the pharmaceutical industry's 'big data' problem. KGs have opened the doors to the application of graph theory to drug discovery; harnessing powerful network algorithms to systematically 'fill in' the unknown areas of the genome and draw novel insights into the genes and mechanisms that underpin disease. There has been significant research interest in KGs in both academia and industry, using them principally for target identification and drug repurposing. KGs are principally used for link prediction by employing KG embedding methods. Whilst these methods are continually evolving, they remain relatively immature. Hereinafter, we present the author's opinion on the current shortcomings of KGs, the areas in which they need to be improved, and evaluate their utility in a drug discovery project.

KGs are fundamentally question answering tools. Questions such as *does drug X treat disease Y?* and *does gene X regulate disease Y?* have demonstrably been answered. However this doesn't reflect the granularity and variety of questions asked by researchers and scientists in the drug discovery process. We need to work toward a universal and extensible system that can answer questions such as *given pathway X, which compounds agonize targets assayed in only functional assays with potency <1 mm? And given diseases with the shared pathogenic mechanism Y, which targets have failed clinical trials at Phase I or II and why?* and *for disease Z, which targets have ligands in different stages of the development process with publications and/or patents describing these compounds?* When KGs can answer these questions, their value will increase immeasurably.

KGs are fundamentally question answering tools. Questions such as *does drug X treat disease Y?* and *does gene X regulate disease Y?* Have demonstrably been answered. However, this does not reflect the granularity and variety of questions asked by researchers and scientists in the drug discovery process. We need to work toward a universal and extensible system that can answer questions such as *given pathway X, which compounds agonize targets assayed in only functional assays with potency <1 mm*? and *given diseases with the shared pathogenic mechanism Y, which targets have failed clinical trials at Phase I or II and why*? and *for disease Z, which targets have ligands in different stages of the development process with publications and/or patents describing these compounds*? Work on this topic is already underway [102]. When KGs can answer these questions, their value will increase immeasurably.

Pathology is fascinatingly complex. This complexity is often not well-represented in KGs. Many research projects use publicly accessible KGs which provide a reductive model of disease (with edges such as *drug-binds-gene, gene-associates-disease* and *drug-treats-disease*). These graphs fail to represent neither the genetic heterogeneity, nor transient nature of disease. Whilst a drug repurposing link prediction model may successfully predict *CFTR-associates-chronic_pancreatitis, ivacaftor-binds-CFTR*, and *ivacaftor-treats-chronic_pancreatitis*, these generalizations do not reflect the complexity of the disease, or the prerequisites of ivacaftor to be an effective treatment. In reality, we want to be able to use a KG to approximate the causal reasoning of a team of researchers: "chronic pancreatitis is caused by loss of function of the CFTR gene. Mutations in CFTR cause an imbalance of calcium homeostasis, leading to early protease activation, fibrosis, inflammation and abdominal pain. Ivacaftor is used to treat a subset of cystic fibrosis patients via potentiation and correction of mutant CFTR, which restores the calcium homeostasis in endothelial cells. Patients with similar loss-of-function mutations in the CFTR gene could be treated with Ivacaftor. Whilst CFTR remains the main pathomechanism of chronic pancreatitis, other possible treatments include immunosuppressants, antifibrotics, protease inhibitors, and analgesics". A KG that can deliver this level of granularity would be a fundamental asset in any drug discovery company.

KGs are mainly used in conjunction with KG embedding models. These models are based on *reasoning-by-association* (also called *guilt-by-association*). This is distinctly different from a causal model of the underlying biological mechanism.

The most informative paths of many network embedding-based models are not describing biological paths (e.g. *drug inhibits-gene-causes-disease*), but instead are describing similarities between source and target nodes (e.g. *drug-resembles-drug treats-disease* and *disease-resembles-disease-treats-drug*) [26**] [45], Moreover, most relation inference models do not capture directionality nor trend of the edge [103]. Whilst researchers have developed models that produce inference paths between source and target nodes to approximate the biology path [39,104], such paths are often not well-correlated with the underlying causal biological path. Perhaps we should strive to move away from models that simply associate biological components, and more toward models that accurately describe the underlying biological system. This problem seems endemic in the wider field of artificial intelligence. Gary Marcus and Ernest Davis echoed the problem, stating ″we need to stop building computer systems that merely get better and better at detecting statistical patterns … and start building computer systems that from the moment of their assembly innately grasp three basic concepts: time, space and causality" [105]. Whilst KG embeddings remains an over-populated area of research, with researchers competing to eek out the smallest increase in model performance, causal network reasoning remains a largely unexplored field. There have been a handful of notable network-based causal reasoning approaches that have been successfully applied to drug discovery [106–114]. We hope to see more causal models, built upon biologically-representative KGs.

In areas such as target identification, link prediction methods have demonstrated their utility in academia and industry led projects. Applications such as drug–target interaction, drug–drug interaction, and protein–ncRNA interaction remain academic exercises in graph theory, often surpassed by powerful deep learning approaches with features based solely on the physicochemical structures of the interacting entities (the power of which is exemplified by *AlphaFold*). We believe GML is best suited to the prediction of abstract entities such as diseases. Modeling physicochemical interactions should be left to structure-based approaches. Whilst it was assumed that graph embedding methods inferred edge existence via network proximity, it has become evident the overwhelming majority of their predictive power comes simply from the connectivity of the nodes, and not their local neighborhood. This issue becomes especially problematic when using literature-derived KGs, where link prediction models strive to approximate the biologically incorrect degree distribution of a literature-derived network and not that of the underlying biological system. Mitigation strategies, more appropriate evaluation metrics and less biased graphs are desperately needed to correct this problem. Network medicine is based on the assumption that we can accurately model the biological systems that govern disease; applying graph theory to describe them mathematically. Whilst we are increasingly creating more and more data pertaining to these systems, we currently cannot sufficiently model them. KGs are undoubtedly a useful framework on which to build such approaches. To be able to develop informative computational models, we must strive toward building KGs which describe the complex dynamic biological systems of the human body, how they are dysregulated in the disease state, and how therapeutics act upon the systems. Whilst the dog days of phenotypic-based drug discovery have not yet passed, the dawn of target-based discovery is certainly upon us. Biologically-representative KGs will be instrumental in the era of systems biology.

Acknowledgments

The author would like to express their gratitude to Delphine Rolando, Rachel Hodos and Dane Corneil. Their expertise in drug discovery, graph machine learning, and knowledge graphs was instrumental in writing this review. Lastly, we thank Daniel Miskell for his insight over the years.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

Funding

This manuscript was supported by BenevolentAI.

Declaration of interest

F MacLean is a full-time employee of BenevolentAI. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

ORCID

Finlay MacLean http://orcid.org/0000-0003-2779-179X

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. "Total global pharmaceutical RD spending 2012–2026,". [cited 2021 Jul 03]. Available from: https://www.statista.com/statistics/309466/global-r-and-d-expenditure-forpharmaceuticals
2. "2020 FDA drug approvals,". [cited 2021 Jul 03]. Available from: https://www.nature.com/articles/d41573-021-00002-0
3. "Ten years on: measuring the return from pharmaceutical innovation 2019,". [cited 2021 Jul 03]. Available from: https://www2.deloitte.com/us/en/pages/life-sciences-andhealth-care/articles/measuring-return-from-pharmaceutical-innovation.html
4. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. Science. 2003;300(5617):286–290.
5. 1000 G. P. Consortiumet al.. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061.
6. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK biobank. Nat Genet. 2018;50(11):1593–1599.
7. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. Nucleic Acids Res. 2010;39(suppl 1):D19–D21.
8. Leinonen R, Akhtar R, Birney E, *et al*. The european nucleotide archive. Nucleic Acids Res. 2010;39(suppl 1):D28–D31.

9. Ponten F, Jirstrom K, Uhlen M. The human protein atlas—a tool for pathology. J Pathol. 2008;216(4):387–393.
10. GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science. 2015;348 (6235):648–660.
11. Stathias V, Turner J, Koleti A, *et al*. Lincs data portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020;48(D1):D431–D439.
12. Tomczak K, Czerwinska P, Wiznerowicz M. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol. 2015;19(1A):A68.
13. Ghandi M, Huang FW, Jane-Valbuena J, *et al*. Next-generation characterization of the cancer cell line encyclopedia. Nature. 2019;569(7757):503–508.
14. Tsherniak A, Vazquez F, Montgomery PG, *et al*. Defining a cancer dependency map. Cell. 2017;170(3):564–576.
15. Chadwick LH. The NIH roadmap epigenomics program data resource. Epigenomics. 2012;4(3):317–324.
16. Kozomara A, Griffiths-Jones S. MiRBase: integrating microrna annotation and deep-sequencing data. Nucleic Acids Res. 2010;39(suppl 1):D152–D157.
17. Volders P-J, Helsens K, Wang X, et al. Lncipedia: a database for annotated human lncrna transcript sequences and structures. Nucleic Acids Res. 2013;41(D1):D246–D251.
18. Cui T, Zhang L, Huang Y, et al. Mndr v2. 0: an updated resource of ncrna–disease associations in mammals. Nucleic Acids Res. 2018;46 (D1):D371–D374.
19. Earm K, Earm YE. Integrative approach in the era of failing drug discovery and development. Integr Med Res. 2014;3(4):211–216.
20. Rago L, Santoso B. “Drug regulation: history, present and future,” ¨. Drug Benefit Risks. 2008;2:65–77.
21. “Novartis CEO who wanted to bring tech into pharma now explains why it’s so hard,”. [cited 2020 Sep 30]. Available from: https://www.forbes.com/sites/davidshaywitz/2019/01/16/novartis-ceo-who-wanted-to-bring-tech-into-pharma-now-explains-why-its-so-hard, accessed: 2020-september-30.
22. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al*. The fair guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1):1–9.
23. Iams WT, Lovly CM. Molecular pathways: clinical applications and future direction of insulin-like growth factor-1 receptor pathway blockade. Clin Cancer Res. 2015;21(19): 4270–4277.
24. Rossi A, Firmani D, Matinata A, et al. Knowledge graph embedding for link prediction: a comparative analysis. arXiv Preprint arXiv:2002 00819. 2020.
25. Zou X. A survey on application of knowledge graph. JPhCS. 2020;1487(1):012016.
26. Gao Y, Li Y-F, Lin Y, et al. Deep learning on knowledge graph for recommender system: a survey. arXiv Preprint arXiv:2004 00387. 2020.
27. “Neo4j graph database. [cited 2021 Sep 12]. Available from: https://neo4j.com
28. Himmelstein, DS, Lizee, A, Hessler, C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6:e26726.

•• **Article of high interest - This seminal paper represents one of the earliest attempts to train a link prediction model on a biomedical knowledge graph, to answer biological questions (in this case drug repurposing). This research area has matured significantly since this manuscript. The knowledge graph they developed is rather small using today’s standards, and research interest has moved away from pathway-based models to embedding-based models, in part due to their scalability. However, all researchers and practitioners working in this space would benefit from understanding the provenance of their work. The authors also highlight the prior probability of connection problem in their manuscript. This, however, is covered in more detail in their more recent work.**

29. Himmelstein DS, Baranzini SE. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS Comput Biol. 2015;11(7):e1004259.
30. Breit A, Ott S, Agibetov A, et al. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. arXiv Preprint arXiv:1912 04616. 2019.
31. Womack F, McClelland J, Koslicki D. Leveraging distributed biomedical knowledge sources to discover novel uses for known drugs. bioRxiv. 2019;765305.
32. Percha B, Altman, RB. A global network of biomedical relationships derived from text. Bioinformatics. 2018;34(15):2614–2624.

- **Article of interest - Despite being not without their shortcomings, literature-derived knowledge graphs are popular methods of rapidly generating biological knowledge graphs. This paper provides an effective method of knowledge graph generation from publicly available data sources. Their derived biological relationships are pleasingly complex, compared to other efforts. Their evaluation of literature-derived relationships against structured databases highlights the disparity between structured and unstructured data sources, and the need for effective edge harmonization methods.**

33. Ioannidis VN, Song X, Manchanda S, et al. Drkg-drug repurposing knowledge graph for COVID-19. arXiv. 2020.
34. Belleau F, Nolin M-A, Tourigny N, et al. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008;41(5):706–716.
35. Chen B, Dong X, Jiao D, et al. Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics. 2010;11(1):255.
36. Yue, X, Wang, Z, Huang, J, et al. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics. 2020;36(4):1241–1251.

- **Article of interest - For any researcher wishing to learn the fundamentals of graph embeddings and their applications, this review is a must-read.**

37. Gao F, Musial K, Cooper C, et al. Link prediction methods and their accuracy for different social networks and network metrics. Sci Programm. 2015;2015:1–13.
38. Cai H, Zheng VW, Chang K-C-C. A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans Knowledge Data Eng. 2018;30(9):1616–1637.
39. Xia X, “Knowledge Graph Embedding Methodologies,”. [cited 2020 Jul 03]. Available from: https://github.com/xinguoxia/KGE#methodologies
40. Hodos, RA, Kidd, BA, Khader, S, et al. Computational approaches to drug repurposing and pharmacology. Wiley Interdiscip Rev Syst Biol Med. 2016;8(3):186.

- **Article of interest - This review highlights drug repurposing as a promising application of knowledge graphs. Knowledge graphs and associated graph machine learning approaches only constitute a few of the many computational approaches that have been used for drug repurposing. This manuscript provides a comprehensive summary of most other approaches.**

41. Talevi A, Bellera CL. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. Expert Opin Drug Discov. 2020;15(4):397–401.
42. Wang L, Lei Y, Gao Y, et al. Association of finasteride with prostate cancer: a systematic review and meta-analysis. Medicine (Baltimore). 2020;99(15):e19486.
43. Jain P, Jain SK, Jain M. Harnessing drug repurposing for exploration of new diseases: an insight to strategies and case studies. Curr Mol Med. 2020;20. DOI:10.2174/1566524020666200619125404
44. Ganzer CA, Jacobs AR, Iqbal F. Persistent sexual, emotional, and cognitive impairment post-finasteride: a survey of men reporting symptoms. Am J Men’s Health. 2015;9(3):222–228.
45. Poleksic A. Overcoming sparseness of biomedical networks to identify drug repositioning candidates. bioRxiv. 2020.
46. Sosa DN, Derry A, Guo M, et al. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. bioRxiv. 2019;727925.

47. Xu B, Liu Y, Yu S, et al. A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. BMC Med Genomics. 2019;12(10):188.
48. Gaudelet T, Day B, Jamasb AR, *et al*. Utilising graph machine learning within drug discovery and development. arXiv Preprint arXiv:2012 05716. 2020.
49. Paliwal S, De Giorgio A, Neil D, et al. Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. Sci Rep. 2020;10(1):1–19.
50. Amaral PP, Dinger ME, Mattick JS. Non-coding rnas in homeostasis, disease and stress responses: an evolutionary perspective. Brief Funct Genomics. 2013;12(3):254–278.
51. Ji B-Y, You Z-H, Cheng L, et al. Predicting mirna-disease association from heterogeneous information network with grarep embedding model. Sci Rep. 2020;10(1):1–12.
52. Zhou J-R, You Z-H, Cheng L, et al. Prediction of lncrna–disease associations via an embedding learning hope in heterogeneous information networks. Mol Ther Nucleic Acids. 2020;23:277-285.
53. Zheng Y, Peng H, Zhang X, et al. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces. BMC Bioinformatics. 2019;20(23):605.
54. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8 (1):1–13.
55. Lim H, Gray P, Xie L, et al. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Sci Rep. 2016;6(1):1–11.
56. Ba-Alawi W, Soufan O, Essack M, et al. Daspfind: new efficient method to predict drug–target interactions. J Cheminform. 2016;8 (1):15.
57. Mizutani S, Pauwels E, Stoven V, et al. Relating drug–protein interaction network with drug side effects. Bioinformatics. 2012;28(18): i522–i528.
58. Wan F, Hong L, Xiao A, et al. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics. 2019;35(1):104–111.
59. Huang K, Fu T, Xiao C, et al. Deeppurpose: a deep learning based drug repurposing toolkit. arXiv Preprint arXiv:2004 08919. 2020.
60. Wallach I, Dzamba M, Heifets A. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv Preprint arXiv:1510 02855. 2015.
61. Senior AW, Evans R, Jumper J, *et al*. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577 (7792):706–710.
62. Reese JT, Unni DR, Callahan TJ, *et al*. KG-COVID-19: a framework to produce customized knowledge graphs for covid-19 response. Patterns. 2020;2(1):100155.
63. Zhou Y, Hou Y, Shen J, et al. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov 2. Cell Discov. 2020;6 (1):1–18.
64. Wang LL, Lo K, Chandrasekhar Y, *et al*. Cord-19: the covid-19 open research dataset. ArXiv. 2020.
65. Hsieh K, Wang Y, Chen L, et al. Drug repurposing for covid-19 using graph neural network with genetic, mechanistic, and epidemiological validation. arXiv Preprint arXiv:2009 10931. 2020.
66. Gysi DM, Valle ID, Zitnik M, *et al*. Network medicine framework for identifying drug repurposing opportunities for covid-19. arXiv Preprint arXiv:2004 07229. 2020.
67. Gasmi A, Tippairote T, Mujawdiya PK, et al. Neurological involvements of sars-cov2 infection. Mol Neurobiol. 202
68. Stebbing J, Phelan A, Griffin I, et al. Covid-19: combining antiviral and anti-inflammatory treatments. Lancet Infect Dis. 2020;20 (4):400–402.
69. "Baricitinib receives emergency use authorization from the FDA for the treatment of hospitalized patients with COVID-19,". [cited 2021 Jan 02]. Available from: https://investor.lilly.com/news-releases/news-release-details/baricitinib-receives-emergency-use-authorization-fda-treatment
70. Kuchaiev O, Rasajski M, Higham DJ, et al. Geometric de-noising of protein-protein interaction networks. PLoS Comput Biol. 2009;5(8): e1000454.
71. Xiao Z, Deng Y. Graph embedding-based novel protein interaction prediction via higher-order graph convolutional network. PloS One. 2020;15(9):e0238915.
72. Yang F, Fan K, Song D, et al. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics. 2020;21(1):1–16.
73. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457–i466.
74. Lim H, Poleksic A, Xie L. Exploring landscape of drug-target-pathway-side effect associations. AMIA Summits Translat Sci Proceed. 2018:132–141.
75. Zhang W, Chen Y, Liu F, et al. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics. 2017;18(1):18.
76. Su C, Tong J, Zhu Y, et al. Network embedding in biomedical data science. Brief Bioinform. 2020;21(1):182–197.
77. Sangar V, Blankenberg DJ, Altman N, et al. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics. 2007;8(1):294.
78. Grover A, Leskovec J, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, USA. pp. 855–864.
79. Stark C, Breitkreutz B-J, Reguly T, et al. Biogrid: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(suppl 1):D535–D539.
80. Kulmanov M, Khan MA, Hoehndorf R. Deepgo: predicting protein functions from sequence and interactions using a deep ontology aware classifier. Bioinformatics. 2018;34(4):660–668.
81. Nariai N, Kolaczyk ED, Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data. Plos One. 2007;2(3): e337.
82. Makrodimitris S, Van Ham RC, Reinders MJ. Automatic gene function prediction in the 2020's. Genes (Basel). 2020;11(11):1264.
83. Goymer P. Why do we need hubs? Nat Rev Genet. 2008;9(9):651.
84. Chen S-J, Liao D-L, Chen C-H, et al. Construction and analysis of protein-protein interaction network of heroin use disorder. Sci Rep. 2019;9(1):1–9.
85. Dai W, Chang Q, Peng W, et al. Network embedding the protein–protein interaction network for human essential genes identification. Genes (Basel). 2020;11(2):153.
86. Lefranc F, Tabanca N, Kiss R. Assessing the anticancer effects associated with food products and/or nutraceuticals using in vitro and in vivo preclinical development-related pharmacological tests. In: Seminars in cancer biology. Vol. 46. Elsevier; 2017. p. 14–32.
87. Veselkov K, Gonzalez G, Aljifri S, et al. Hyperfoods: machine intelligent mapping of cancer-beating molecules in foods. Sci Rep. 2019;9(1):1–12.
88. Du J, Jia P, Dai Y, et al. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics. 2019;20(1):7–15.
89. Goh K-I, Cusick ME, Valle D, et al., "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007, USA.
90. Cantini L, Medico E, Fortunato S, et al. Detection of gene communities in multi-networks reveals cancer drivers. Sci Rep. 2015;5 (1):17386.
91. Zietz M, Himmelstein DS, Kloster K, et al. The probability of edge existence due to node degree: a baseline for network-based predictions. Manubot, Tech Rep. 2020.
•• **Article of high interest - This paper provides the most comprehensive analysis of the problems arising from i) the degree imbalance in graphs with long-tailed distributions, and ii) the disparity between literature-derived biological networks and those derived from systematic screens. Of particular interest is the authors' closed form approximation of the prior probability of connection. This allows researchers and industry**

professionals to differentiate between predictions based on network connectivity and proximity at almost no computational cost.

92. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36 (suppl1):D901–D906.
93. Avram S, Bologa CG, Holmes J, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res. 2021;49(D1): D1160–D1169.
94. Edwards AM, Isserlin R, Bader GD, et al. Too many roads not taken. Nature. 2011;470(7333):163–165.
95. Oprea TI, Bologa CG, Brunak S, *et al*. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. 2018;17(5):317.
96. Hutchison CA, Chuang R-Y, Noskov VN, *et al*. Design and synthesis of a minimal bacterial genome. Science. 2016;351(6280):6280.
97. Feng R, Yang Y, Hu W, et al. Representation learning for scale-free networks. arXiv Preprint arXiv:1711 10755. 2017.
98. Kang B, Lijffijt J, Bie TD. Conditional network embeddings. arXiv Preprint arXiv:1805 07544. 2018.
99. Buyl M, De Bie T. Debayes: a bayesian method for debiasing network embeddings. arXiv Preprint arXiv:2002 11442. 2020.
100. Lerer A, Wu L, Shen J, et al. Pytorch-biggraph: a large-scale graph embedding system. arXiv Preprint arXiv:1903 12287. 2019.
101. Zheng D, Song X, Ma C, et al. Dgl-ke: training knowledge graph embeddings at scale. arXiv Preprint arXiv:2004 08532. 2020.
102. Hamilton WL, Bajaj P, Zitnik M, et al. Embedding logical queries on knowledge graphs. arXiv Preprint arXiv:1806 01445. 2018.
103. Lee B, Zhang S, Poleksic A, et al. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet. 2020;10:1381.
104. Lin XV, Socher R, Xiong C. Multi-hop knowledge graph reasoning with reward shaping. arXiv Preprint arXiv:1808 10568. 2018.
105. Bishop JM. Artificial intelligence is stupid and causal reasoning won't fix it. arXiv Preprint arXiv:2008 07371. 2020.
106. Liu A, Trairatphisan P, Gjerga E, et al. From expression footprints to causal pathways: contextualizing large signaling networks with carnival. NPJ Syst Biol Appl. 2019;5(1):1–10.
107. Rivas-Barragan D, Mubeen S, Guim-Bernat F, et al. Drug2ways: reasoning over causal paths in biological networks for drug discovery. bioRxiv. 2020.
108. Vidal M, Cusick ME, Barabasi A-L. Interactome networks and human disease. Cell. 2011;144(6):986–998.
109. Broido AD, Clauset A. Scale-free networks are rare. Nat Commun. 2019;10(1):1–10.
110. Dorogovtsev S, Mendes J, Samukhin A. Generic scale of the" scale-free" growing networks. arXiv Preprint Cond-mat/0011115. 2000.
111. Rohani N, Eslahchi C. Drug-drug interaction predicting by neural network using integrated similarity. Sci Rep. 2019;9(1):1–11.
112. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. Jama. 2020;323(9):844–853.
113. Mohs RC, Greig NH. Drug discovery and development: role of basic biological research. Alzheimers Dementia. 2017;3(4):651–657.
114. Xue S, Lu J, Zhang G. Cross-domain network representations. Pattern Recogn. 2019;94:135–148.