On Molecular Biological Knowledge Graphs

Sameh K. Mohamed[1,2,3]

[1] Data Science Institute
[2] Insight Centre for Data Analytics
[3] National University of Ireland Galway
sameh.mohamed@insight-centre.org

Abstract. Knowledge graph became a popular means for modelling data of interconnected entities *i.e.* linked data. They are adopted in many industrial and academic applications. In this work, we focus on the use of knowledge graphs in the field of molecular biology, where the linked data is centred around proteins and their associations with other biological entities. Recently, this type of knowledge graphs is becoming popular to support predicting different biological associations *e.g.* drug-protein targets, gene-disease associations, protein-protein interactions, etc. In this work, we explore the process of building and using these knowledge graphs. We first discuss available knowledge sources of molecular biology knowledge and the process of processing these sources to generate knowledge graphs. We then discuss various tasks and applications that can use these knowledge graphs to predict different biological associations. Finally, we provide an example for building and using knowledge graphs in molecular biology. We build a SwissProt-based knowledge graph of different protein associations, and we show by experiments that knowledge graph embedding models can achieve high accuracy in predicting the different protein associations compared to random baselines.

Keywords: Knowledge Graphs · Molecular Biology · Link Prediction.

1 Introduction

Knowledge graphs are graphs structured knowledge bases that model linked data in a form of a graph, where graph nodes represent knowledge entities and its edges represent the relations between them. In recent years, knowledge graphs became a popular means for modelling linked data in many domains, including general human knowledge [1], biomedical information [2] and language lexical information [3]. They are now used to support different applications such as enhancing semantics of search engine results, biomedical discoveries [4], powering question answering and decision support systems [5].

Recently, knowledge graphs were introduced to the field of molecular bioinformatics, where they modelled the knowledge about molecular biology entities *e.g.* proteins and their associations with other biological entities *e.g.* disease, pathways, biological functions, etc. They were then used to provide insights about the biological activities of proteins, and their associations with other biological

entities [6]. Previously, protein linked data were modelled using uni-relational networks, where the different computer-based predictive approaches utilised various network similarity measures [7] and casual network embeddings [8] to learn new biological association within these networks. However, the recent advances in the design of knowledge graph embedding models [9] allowed them to excel in the task predicting biological associations, where they outperform other models including classical network embeddings models [4].

Despite their usefulness and proven success, knowledge graphs and their embedding models are still in an early stage of adoption in the molecular biology domain, where were only used in a limited set of tasks [10]. Therefore, in this work, we explore the different possible uses of knowledge graphs in modelling molecular biology knowledge to solve different biological association tasks. The rest of our study is structured as follows:

(A) section 2 provides a basic introduction to different topics addressed in this study such as knowledge graphs, molecular biology, and knowledge graph embedding techniques.
(B) we discuss the process of building knowledge graphs for molecular biology data, and a set of expert curated molecular biological knowledge bases in section 3.
(C) We discuss the different possible applications for the knowledge graphs and their embedding models in learning different biological associations in section 4.
(D) We propose a new knowledge graph based on SwissProt knowledge base that is centred around proteins and we perform comparative evaluation for a set of knowledge graph embedding models on the proposed knowledge graph to predict protein associations to different biological entities, and we report the outcome results in section 5.

2 Background

In this section, we introduce concepts and terminologies that we use during the rest of the study.

2.1 Knowledge Graph Embedding

Knowledge graph embedding models learn a low rank vector representation of knowledge entities and relations that can be used to rank knowledge assertions according to their factuality. KGE models are trained in a multi-phase procedure, where their objective is to effectively learn a vector representation of entities and relations that can be used to score and rank possible knowledge facts.

First, a KGE model initialises all embedding vectors using random noise values. It then uses these embeddings to score the set of true and false training facts using a model-dependent scoring function. The output scores are then passed to the training loss function to compute training error. These errors are

used by optimisers to generate gradients and update the initial embeddings, where the updated embeddings give higher scores for true facts and less for false facts. This procedure is performed iteratively for a set of iterations *i.e.* epochs in order to reach a state where embeddings provide the best possible scoring for both true and false possible facts.

2.2 Molecular Biology

Molecular biology is the study of the nature and activities of biological macromolecules such as proteins. This includes their structure, interactions and their involvement in the different biological processes. Proteins are one of the main macromolecules that sustain living organisms. Proteins are made of small building blocks known as amino acids that are structured according to specific nucleotide sequences in the genome. However, this linear structure of proteins is non-functional unless it undergoes specific folding steps inside our cells. Once secreted by the cellular organelles, proteins start all the different tasks in our body.

In the context of cancer, proteins control crucial roles in the cells, as the proliferation, migration of cells, angiogenesis, and metastasis. In normal conditions, those proteins are in an inactive state, however, if triggered, potentially by a mutation in case of cancer, they start a cascade of downstream activations that favor the growth and proliferation of cancers. Those cascades of proteins that are linked to each other to form what is known as signaling pathways. Generally, pathways in the living systems control the various biological processes such as including organs activities, metabolism, therapeutic activities of drugs, other disease related processes, etc.

3 Building Biological Knowledge Graphs

In this section, we discuss the process of building biological knowledge graph focused which is centred around proteins. We first explore the currently available biological knowledge sources, and we then discuss how to integrate them together to build both generic and task specific knowledge graphs.

Biological knowledge graphs are built using a multiple-step procedure. First, we process available biological knowledge bases of different specialities to generate associations between their biological entities. Secondly, we apply labels to the associations according to the types of their domain and range entities. Finally, labelled associations are integrated together by joining association through common entities to construct the joint knowledge graphs of all associations.

In the following, we discuss this process in detail, and we discuss popular biological databases, their properties, and common issues that arise during building biological knowledge graphs.

Table 1. Comparison between popular biological knowledge bases. The term "GO" refers to gene ontology annotations, "S" refers to structured data and "U" refers to unstructured data.

Database	Specialisation	Associations	Format
UniProt	Proteins	Diseases, Drugs, GO, Antibodies, Pathways	S, U
Reactome	Pathways	GO, Complexes, Pathways	S
KEGG	Pathways	Diseases, Chemicals, Drugs, Pathways	S
DrugBank	Drugs	Proteins, Pathways	S, U
Gene Ontology	GO	Proteins	S
CTD	Chemicals	Drugs, Proteins, GO, Phenotypes	S, U
SIDER	Drugs	Side effects, Indications	S
HPA	Proteins	Antibodies, Tissues, Cell lines	S, U

3.1 Biological Knowledge Sources

Recently, biological knowledge bases became available in many forms including biological paper abstracts [11], raw biological experimental data [12], curated annotations [13,14], etc.

In our study, we focus on the databases that offer structured data that can be easily processed to generate knowledge graphs. Table 1 summarises specialisation and properties of popular biological knowledge bases. In the following, we discuss a subset of these knowledge bases that represent the most popular curated databases for molecular biology data:

1) **UniProt**: The Universal Protein Resource (UniProt)[4] [13] is a comprehensive resource for protein sequence and annotation data. It includes protein centered information for various species. The information in the UniProt knowledge base is divided into two main sources: manually curated data *i.e.* SwissProt and computer predicted data *i.e.* TrEMBL. Both sources include information about proteins of different species, and this information is provided in combination of both structured and unstructured data formats. The UniProt protein identifiers are considered one of the most adopted within all biological knowledge bases, where it can be easily converted to other protein identifiers. Information in the UniProt knowledge base include protein names, sequences and sub sequence information. It also include protein associations to other biological entities from other database like drugs (DrugBank [15]), disease (OMIM [16]), gene ontology annotations (GO [17]), pathways (Reactome [18]) and protein families and motifs (PFam [19]). UniProt also includes curated protein-protein interaction from the InAct database [20].

2) **Reactome**: REACTOME[5] [18] is an open-source, open access, manually curated and peer-reviewed pathway database. The REACTOME database

[4] https://www.uniprot.org/
[5] https://reactome.org/

provides associations between proteins and protein complexes, reactions and pathways. It also includes links between its base pathways and the gene ontology annotations. The REACTOME database uses UniProt unique identifiers *i.e.* accession codes as the unique identifiers for its proteins, while in some cases it uses other identifier systems when UniProt ids are not available.

3) **KEGG**: The Kyoto encyclopedia of genes and genomes (KEGG)[6] [14,21] is a database resource for high-level functions and utilities of the biological systems. It includes information about genes (and their corresponding proteins), diseases, drugs, chemicals, pathways and some other biological entities and concepts related to them. The database uses the Entrez Gene DB [22] based identifiers for its proteins proteins, and it uses its own identifier systems for all other biological entities. The KEGG database includes a rich set of biological associations including protein-drug, protein-disease, protein-pathway, drug-pathway and drug-disease associations.

4) **Gene Ontology**: The gene ontology (GO) database [17] is large resource for protein functions in living systems. It provides detailed annotations for proteins biological processes, molecular functions and cellular locations. The GO database represents the most popular resource for modelling protein activities, where provides associations between proteins and their activities in a hierarchical structure. The database also contains a rich ontology that describes the different relations between the protein annotations.

5) **DrugBank**: The DrugBank database [15] is resource for drug target data. It includes different types of information about drugs and chemicals including their structure, target proteins, related therapeutic pathways, etc. The resource also classifies drugs according their development state *e.g.* under development or approved, and it includes information about the drugs' expected indications along with their dosages and known drug drug interactions. The DrugBank database is considered one of the largest available resources for curated data about drugs and their protein targets.

3.2 Knowledge Integration

After processing the biological knowledge resources, the data extracted from them is usually formatted as sets of pairs of entities with unlabelled associations. Since knowledge graphs require labelled relations, these data associations need to be labelled according to their corresponding semantics. They are then merged together through common entities to form a graph. In the following, we discuss both association labelling and merging in the context of building biological knowledge graphs.

[6] https://www.kegg.jp/kegg/

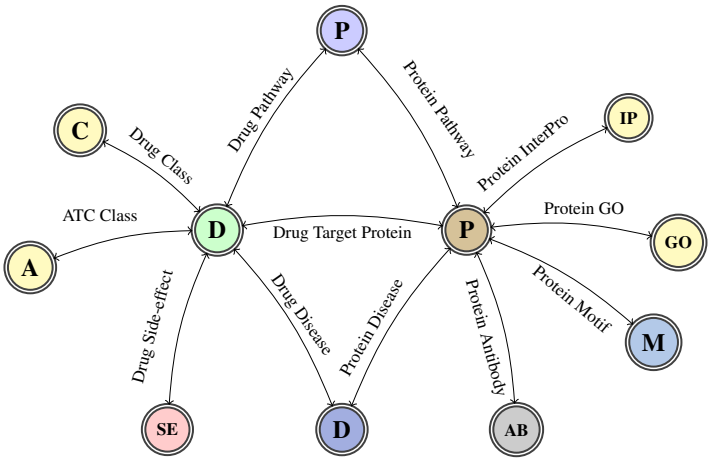

Fig. 1. Graph schema for an example molecular biological knowledge graph.

1) **Association labelling**: Biological associations extracted from knowledge sources are combined to form a graph of interconnected entities. The easiest for of association labelling is to use a combination of both relation source and destination entity databases to generate a unique label for each relation *e.g.* "*protein pathway*" as in Fig. 1. However, in many cases sources or destinations of the same extracted association pairs can belong different types biological entities. For example, the associations between UniProt and Reactome entities are commonly considered protein-pathway relations. However, the Reactome entities in these relations can often refer to protein complex or reactions, which then significantly changes the semantics of the association. In another example, the UniProt database provide associations between its proteins and the InterPro [23] entities, where these entities represent a wide range of sequence-based patterns like domains, repeats, protein families, etc. Therefore, Their associations with the UniProt proteins represent different semantics. In this case, using different types of associations labels that can be obtained by amending the destination entity type to the relation type *e.g. "protein InterPro domain"* is essential to enrich the semantics of the outcome knowledge graph.

2) **Merging biological associations**: Combining biological knowledge graphs is performed by joining common entities in different association types. However, due to the different types of identifiers in biological data, same entities are often referred to using different codes. This can intuitively fixed by mapping entities to one identifier system (*e.g.* UniProt for proteins). However, different resources have different coverage ratios of biological entities, where mapping to a different identifier system can result in a loss of information. Also, the mapping between different database is not not guaranteed to be a

one-to-one mapping, where some databases include different identifiers for the same entities that represents different states of these entities.
On the other hand, the identifiers within the same biological databases are regularly updated in their continues review and update process, where famous resources like Reactome and UniProt have witnessed changes to large volume of their identifiers in the recent years. This change affect other databases that use old identifiers, therefore, they are not linked to the newly introduced entities. This issue can be resolved by updating outdated links using the logs of the identifiers updates.

4 Use Cases in Molecular Biology

In this section, we present a set of potential application for knowledge graphs in the field of molecular biology. We focus our enquiry on the applications that can be achieved using link prediction on knowledge graphs to learn different associations between biological entities.

4.1 General Predictive Models in Molecular Biology

The research in molecular biology demands large resources to execute laboratory experimentation of potential biological discovery. This process is also non-scalable and often operate on single or few number of investigated entities. Due to the advances of computer-based predictive models, researcher have benefited from the scalable simulation-based models to assist their in-lab experiments to predict new links between the different biological entities as in the following tasks:

1) **Protein Molecular and Biological Activities**: Proteins have different biological and molecular functions in the living systems. These activities are modelled in the gene ontology database, where protein functions and activities are provided as annotations to proteins. This encouraged multiple works to develop methods for predicting unknown protein activities using the known gene ontology annotation. These works use multiple approaches such as sequence analysis, matrix decomposition [24] and network similarity [25].

2) **Gene Disease Associations** Predicting the gene-disease associations has witnessed rapid developments in recent years to identify unknown links between proteins and diseases. Similar to protein activity prediction, the developed gene-disease associations prediction models use different methods such as matrix decomposition, network propagation and boosted tree regression [26].

3) **Drug Protein Targets and their side effects** The study of drug targets has become very popular with the objective of explaining mechanisms of actions of current drugs and their possible unknown off-target activities. Knowing targets of potential clinical significance also plays a crucial role

in the process of rational drug development. With such knowledge, one can design candidate compounds targeting specific proteins to achieve intended therapeutic effects. However, a drug rarely binds only to the intended targets, and off-target effects are common. This may lead to unwanted adverse effects, but also to successful drug re-purposing, *i.e.* use of approved drugs for new diseases [27]. Large-scale and reliable prediction of drug-target interactions (DTIs) can substantially facilitate development of such new treatments. Various DTI prediction methods have been proposed to date. Examples include chemical genetic and proteomic methods such as affinity chromatography and expression cloning approaches. These, however, can only process a limited number of possible drugs and targets due to the dependency on laboratory experiments and available physical resources. Computational prediction approaches have therefore received a lot of attention lately as they can lead to much faster assessments of possible drug-target interactions [**?**].

4.2 Knowledge Graph Embedding in Molecular Biology Tasks

In all of the aforementioned examples, knowledge graph embedding models [9] are a natural fit, where they can predict links between the different biological entities in the biological knowledge graphs. They learn the representation of the association and its corresponding entities by training on the known example then they can provide prediction over the unknown links. In this context, knowledge graph embedding models are known to excel and provide state-of-the-art results in learning links between nodes in a labelled graph [9]. For example, Zitnik et. al [6] showed that using knowledge graph embedding models provide state-of-the-art results in the task of predicting drug polypharmacy side effects.

Despite their high predictive accuracy [9], knowledge graph embedding models are in early adoption stages in molecular biology. The current works in predicting biological associations mostly depends on network-based predictive models, where the use of labelled graph is still not widely utilised [8].

5 Experimental Example

In this section, we give an example of building a molecular biological knowledge graph, and using it to learn association between different biological entities using knowledge graph embedding models.

5.1 Building The SwissProt Knowledge Graph

The SwissProt database represent the manually curated part of the UniProt knowledge base. It contain different assertion about proteins in different species. In our example, we only consider data about proteins in the human body *i.e.* homo sapiens' proteins. We process the SwissProt textual entries of proteins to generate their associations with five different datasets: Reactome, InterPro, DrugBank, OMIM, and GO. We also include enzyme classes of enzyme proteins.

Table 2. Comparison in terms of the area under the ROC and precision recall curves between the evaluation results of a set of knowledge graph embedding models and a random baseline mode in predicting different protein associations on the SwissProt knowledge graph.

Ralation	Random		TransE		DistMult		ComplEx	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR
P. biological process	0.50	0.09	0.96	0.77	**0.96**	**0.88**	0.95	0.87
P. cell compartment	0.50	0.09	0.96	0.71	**0.96**	**0.84**	0.94	0.82
P. molecular function	0.50	0.09	0.96	0.77	**0.98**	**0.92**	**0.98**	**0.92**
P. antibody	0.50	0.09	0.75	0.22	**0.86**	**0.33**	0.80	0.28
P. disease	0.50	0.09	0.78	0.29	**0.94**	0.53	0.92	**0.54**
P. pathway	0.50	0.09	0.97	0.83	**1.00**	**0.98**	**1.00**	**0.98**
P. drug	0.50	0.09	0.95	0.80	**0.99**	**0.95**	**0.99**	**0.95**
AVERAGE	0.50	0.09	0.90	0.63	**0.95**	**0.77**	0.94	**0.77**

During the process of extracting protein associations with other entities, we retrieve the entries of protein related entities from their corresponding databases when possible, and we include their types as facts in the knowledge graph. After extracting protein associations, we label each association instance according to the type of the destination entity, where we extract these types from the destination databases *e.g.* InterPro and GO. The outcome of labelling does not require special merging since all the used entities use common identifier system. We have exported the outcome knowledge graph to two separate files for general and type-based assertions. We have also published these two files online on a figshare repository [7].

5.2 Experimental Setup

Data: In our experiments, we assess investigate the task of learning multiple types of protein associations to different biological associations such as biological processes, cell compartments, molecular functions, antibodies, diseases, pathways and drugs. We divide our knowledge graph into three split: training, validation and testing, where each of the validation and testing sets contain 10% of each of the investigated associations, and the training set contains the rest of all associations. Negative assertions for both testing and validation examples are generating using corrupting either the subject or the object of the associations using uniform randomly selected entities of the same type as the original entity. The ratio between positive and negative instances is one to ten respectively in both testing and validation.

[7] https://figshare.com/articles/swissprot-hsa-kg/7828601

Model settings: In our experiments, we use state-of-the-art knowledge graph embedding models such as the TransE [28], the DistMult [29], and the ComplEx [30] models to perform link prediction over the SwissProt knowledge graph. We compare these model to a random models as a baseline. We run all the models over the training data of the knowledge graph and we compare their predictive accuracy using the area under both the roc and precision recall curves as an assessment metric. A grid search is performed to obtain best hyper parameters for each model on the validation set, where the set of investigated parameters are: embeddings size $K \in \{50, 100, 150, 200\}$, margin $\lambda \in \{1, 2, 3, 4, 5\}$ for the TransE and the DistMult models, and number of negative samples $n \in \{2, 4, 6, 10\}$. All embeddings vectors of our models are initialised using the uniform Xavier random initialiser [31]. For all the experiments, we use batches of size 5000, with a maximum of 1000 training iterations *i.e.* epochs. The gradient update procedure is performed using the AMSGrad optimiser [32] with a fixed 0.01 learning rate.

5.3 Results

Table 2 shows the results of our experiments, where it shows that all the knowledge graph embedding models provide a significant enhancements over the random baseline model. The results shows that the accuracy of the models vary on the different types of relations, where the best accuracy of knowledge graph embedding models is achieved on predicting protein pathway association, and their worst accuracy is reported in predicting protein associated diseases. The results also show that the DistMult model achieves best results in terms of the area under both the precision recall and roc curves with 0.77 and 0.95 respectively. On the other hand, the random baseline achieves 0.09 and 0.50 in terms of the area under the precision recall and roc curves respectively.

6 Conclusions

In this work, we have discussed the use of knowledge graphs in the field of molecular biology. We have also shown that knowledge graph are a natural fit for modelling molecular biological systems of interconnected entities. We discussed the process of building knowledge graph from popular curated biological knowledge sources, and the issues that can arise through this process and their possible solutions. We have also discussed the potential use cases of knowledge graphs in molecular biology and the current developed techniques to handle these tasks. Finally, we have presented an example experimental pipeline of processing, building and using a knowledge graph of molecular biological data. We built a knowledge graph from the SwissProt database, and we have used knowledge graph embedding models to learn different types of protein associations within the built knowledge graph. Our experiments showed that knowledge graph embedding models showed a significant and consistent enhancement in the predictive accuracy compared to a random baseline in terms of both the area under the roc and precision recall curves.

7 Acknowledgements

This work has been supported by Insight Centre for Data Analytics at National University of Ireland Galway, Ireland (supported by the Science Foundation Ireland grant 12/RC/2289). The GPU card used in our experiments is granted to us by the Nvidia GPU Grant Program.

References

1. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.
2. Michel Dumontier et. al. Bio2rdf release 3: A larger, more connected network of linked data for the life sciences. In *Proceedings of the ISWC 2014 Posters & Demos*, pages 401–404, 2014.
3. George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, November 1995.
4. Emir Muñoz, Vít Novácek, and Pierre-Yves Vandenbussche. Using drug similarities for discovery of possible adverse reactions. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*. AMIA, 2016.
5. David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.
6. Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. In *Bioinformatics*, 2018.
7. Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9:471–472, 2012.
8. Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. Network embedding in biomedical data science. *Briefings in bioinformatics*, 2018.
9. Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.
10. Mona Alshahrani, Mohammed Asif Khan, Omar Maddouri, Akira R. Kinjo, Núria Queralt-Rosinach, and Robert Hoehndorf. Neuro-symbolic representation learning on biological knowledge graphs. In *Bioinformatics*, 2017.
11. Alan R. Aronson, James G. Mork, Clifford W. Gay, Susanne M. Humphrey, and Willie J. Rogers. The nlm indexing initiative's medical text indexer. *Studies in health technology and informatics*, 107 Pt 1:268–72, 2004.
12. Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. In *Nucleic Acids Research*, 2014.
13. The UniProt Consortium. Uniprot: the universal protein knowledgebase. In *Nucleic Acids Research*, 2017.

14. Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.
15. David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36:D901–D906, 2008.
16. Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, David Valle, and Victor A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33:D514 – D517, 2002.
17. The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. In *Nucleic Acids Research*, 2019.
18. Antonio Fabregat et. al. The reactome pathway knowledgebase. *Nucleic acids research*, 44 D1:D481–7, 2016.
19. Robert D. Finn et. al. Pfam: the protein families database. In *Nucleic Acids Research*, 2014.
20. Sandra E. Orchard et. al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. In *Nucleic Acids Research*, 2014.
21. Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016.
22. Donna R. Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 39 Database issue:D52–7, 2011.
23. Alex L et. al. Mitchell. Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2019.
24. Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. In *Bioinformatics*, 2017.
25. Xiangxiang Zeng, Xuan Zhang, and Quan Zou. Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks. *Briefings in bioinformatics*, 17 2:193–203, 2016.
26. Hongyi Zhou and Jeffrey Skolnick. A knowledge-based approach for predicting gene-disease associations. *Bioinformatics*, 32 18:2831–8, 2016.
27. Anne et. al. Corbett. Drug repositioning for alzheimer's disease. *Nature Reviews Drug Discovery*, 11(11):833, 2012.
28. Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.
29. Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, 2015.
30. Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.
31. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org, 2010.
32. Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*, 2018.