Computational Approaches for Drug Repositioning: Towards a Holistic Perspective based on Knowledge Graphs

Marina Boudin*
Univ. Bordeaux, INSERM, BPH, U1219
Bordeaux, France
marina.boudin@u-bordeaux.fr

ABSTRACT

Drug development is a costly and time consuming activity. The traditional process relies on extensive experimental efforts to map out the relevant part of the chemical space. Data about molecules, diseases, genes and other entities are present on many isolated databases, be that internal or external and in heterogeneous formats. They either require costly and inflexible data integration, or time-consuming workflows. Computational approaches, and more recently artificial intelligence based techniques, have emerged as a promising alternative for reducing the development cycle through drug repositioning. Knowledge bases are used to predict new links between old drugs and new targets. We present below the overall approach adopted for my PhD thesis, for a more holistic knowledge graph-based drug repositioning that aims to discover hidden or missing links between existing drugs and targets for which no known treatment is available. Currently, eight data and knowledge resources have already been integrated into the designed knowledge graph.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Machine learning approaches**; • **Applied computing** → **Life and medical sciences**.

KEYWORDS

Knowledge Graphs, Entity embedding, Link prediction ,Drug repositioning

ACM Reference Format:
Marina Boudin. 2020. Computational Approaches for Drug Repositioning: Towards a Holistic Perspective based on Knowledge Graphs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3418510

1 PROBLEM

The current Covid-19 crisis highlights the need to rapidly identify new treatments for emerging diseases. However, drug discovery (DD) faces difficulties in finding new effective drugs in a timely manner. Traditional DD relies on extensive experimental efforts to map out the relevant part of the chemical space. This process of developing new drugs is time consuming and very expensive because they have to be tested in clinical trials. So expensive that specialized companies set aside the treatment of some diseases such as rare diseases because of their development cost. Clinical trials for drugs are standardized into four main phases by regulatory authorities: (i) Phase I consists of evaluating the drug toxicity, (ii) during Phase II, the drug is tested on a small group of people to find the minimal efficient dose and possible side-effects, (iii) Phase III evaluates the drug efficacy against a given disease. Eventually, Phase IV studies are subject to post-marketing surveillance. Many new developed drugs fail more than succeeding as 60% of evaluated drugs fail at phase II. An alternative to traditional DD is drug repositioning or repurposing (DR). The principle of this approach is to find new uses for existing drugs or drugs that did not succeed in the last phases of clinical trials. DR is thus faster and cheaper than usual drug development because it is based on drugs that have already been approved or for which some steps in the development life cycle have already been validated before.

2 STATE OF THE ART

The first cases of DR were found at random. Given the potential of this kind of discovery, researchers have investigated different techniques to reproduce DR. As a matter of fact, there is a substantial effort that aims to propose computational techniques for DR, taking into account relevant characteristics of potential drugs for their repositioning. Most of the time, computational methods make use of a single characteristic. For example, Stokes *et al.* [10] discovered a new antibiotic by using compound chemical structures through deep learning methods able to recognize compounds with a potential activity against *E. coli*.Indeed, the current increasing availability of disease and drug-related information (*e.g.*, phenotypic and omics data) makes it relevant to use AI-based methods to expand DR possibilities. In that case, either ML or deep learning strategies are used to mine large knowledge bases in order to predict possible drug target links. A recent mainstream representation of such knowledge bases is knowledge graphs (KG) [9]. In KG, data are described by the means of (subject, predicate, object) triplets. This format is adapted to represent network-like data, which makes it suitable for DD. Indeed, understanding network-based perspectives of disease mechanisms is a key asset for DD. Network-based approaches have shown promise in predicting novel targets and new uses for existing drugs [11].The main issue in using KG for DR is the insufficient use of existing information about drugs. Indeed, a big challenge in exploring data in the biomedical domain is fragmentation, as data about molecules, diseases, genes and other

[*] Supervised by Gayo DIALLO & Fleur MOUGIN

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CIKM '20, October 19–23, 2020, Virtual Event, Ireland
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3418510

entities are present on many isolated databases (silos), be that internal or external and in many different formats. They either require costly and inflexible data integration efforts, or time-consuming workflows by performing multiple queries on each separate data source. This is due to the fact that this information is heterogeneous and distributed across multiple knowledge resources. KG as a heterogeneous data integration paradigm may be a solution to this issue.

3 APPROACH

Our approach aims to design and develop an holistic KG-based approach for DR which integrates as many drugs and diseases-related data and knowledge resources as necessary. Our main hypothesis is that KGs as data silos unlocker could be used to integrate and interlink an holistic drug and disease-related data from both the Linked Open Data cloud and other freely available knowledge resources. This would enable powerful deep learning-based prediction of hidden links between drugs and targets.The novelty of the work undertaken in the PhD thesis is to introduce the use of natural compounds in the KG based DR as 30% of marketed drugs are based on them and they are the basis of traditional medicines. To the best of our knowledge, no existing work has integrated them in such a knowledge graph. Yet, those compounds may represent a huge source of information.

4 METHODOLOGY

Following this approach, Figure 1 presents the designed workflow and illustrates details given below.

4.1 Drug and disease-related knowledge resources

Our approach is based on the following assumption: the more data and knowledge, the better for DR. To this end, it was first necessary to select knowledge resources dealing with the entities of interest, *i.e.*, drugs, targets, genes and diseases. Molecules which are binding with drugs are called targets. These molecules can be proteins or nucleic-acids. Drug-target pairs extracted from DrugBank[1] are the data core. We then selected resources that could be connected with DrugBank either directly or indirectly. SMILES (Simplified Molecular Input Line Entry Specification) structures, drug-drug interactions and ATC (Anatomical Therapeutic Chemical) codes were extracted from DrugBank. SIDER[2] was also chosen because it contains information about drugs with their side-effects and indications. As natural compounds (*e.g.*, herbs and plant extracts) are also very relevant for DR, we selected the NPASS[3] knowledge base that provides pairs of natural compounds and targets. UniProt[4] and NCBI Gene[5] were used to connect targets and genes. Diseases were extracted from PharmGKB[6] and the Human Phenotype Ontology (HPO)[7]. The integration of freely available resources is challenging when trying to link them to each other because of the lack of a standardized representation and relationships.

4.2 Data and knowledge integration

The first step for the harmonization and integration is to identify a way to link each resource to the other, ensuring that they share common information. Some efforts, such as the Linked Open Data Cloud [7] in general and Bio2RDF [3] in the biological domain, have already been conducted in binding knowledge bases between each other so that reusing their results is possible. In a second time, resources are formatted in triplets to facilitate the integration into a unique KG. These triplets compose the knowledge graph and describe the relationships standing between entities. Building links between knowledge bases is time-consuming due to missing information about entities and links to other knowledge bases. To find the best methodology for this process, a collaboration with the University of Texas, Austin, US (Pr. Ying Ding) has been initiated.

4.3 Knowledge graph embedding

The next step of the approach involves to test the hypothesis through ML techniques. All along of this work, multiple tests will be carried out to assert the progress of the KG integration. To apply learning techniques on KGs, entities must have a representation in accordance with their topology in the graph. Embedding techniques based on a vector representation of each entity according to their position in the KG being processed will thus be investigated. Many ways to obtain these embedding vectors exist [6, 8]. For example, the node2vec [6] Python library learns embedding vectors from a random walk between entity links. Random walk algorithms begin a path from each node and walk from there during a precise number of nodes provided as an hyperparameter. The node2vec algorithm thus learns any entity representation from the graph and computes similarity between entities. However, the goal is to predict missing links from the KG and node2vec is lacking such a prediction pipeline. Unlike node2vec, the PyKEEN Python library [1] implements a prediction pipeline in addition to the embedding step. Indeed, PyKEEN contains a wide range of embedding algorithms such as TransE, TransR and TransD. The designed workflow (Figure 1) uses both node2vec and PyKEEN. Firstly, the KG entity similarities are computed with node2vec in order to reduce the number of links and nodes. Similarities with a score of at least 90% are kept and stored in a triplet according to the *is_similar_to* predicate. This process reduces PyKEEN computation time within the embedding pipeline and thus generates predictions more rapidly. Then, similarities and connections between the main entities are embedded through PyKEEN and provided as input to the prediction algorithm. The pipeline needs supplementary knowledge for the relations of interest (*e.g.*,*has_target* between a drug and a target) and additional inputs containing all the entities to be tested.With the embedding techniques playing a major role, the KG must be well represented in the vector space to be able to obtain good prediction scores. A collaboration with the University of Technology TalTech, Tallinn, Estonia (Pr. Sadok Ben Yahia) was started to elaborate a suitable embedding techniques for very large KGs.

[1] https://www.drugbank.ca/
[2] http://sideeffects.embl.de/
[3] http://bidd2.nus.edu.sg/NPASS/
[4] https://www.uniprot.org/
[5] https://www.ncbi.nlm.nih.gov/gene
[6] https://www.pharmgkb.org/
[7] https://hpo.jax.org/app/

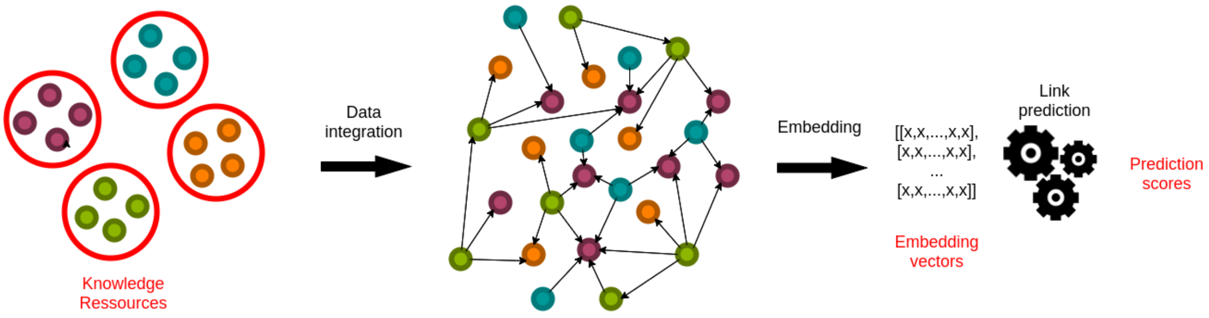

Figure 1: Overall workflow of the proposed approach.

4.4 Validation and preliminary tests

Potential drugs that this work will be able to identify need to go through a testing step. A collaboration with the reference centers on rare diseases in Bordeaux (France) has been launched to this end. They have developed new innovative techniques to test whether a drug is capable of acting on a disease within a few days. This collaboration will also allow to improve the project and share knowledge about DR for finding some rare diseases related treatments.

5 PRELIMINARY RESULTS

Data and knowledge integration is still in progress but a fair amount of this process has already been completed (Figure 2). Currently, the KG accounts 152,086 nodes and 680,757 links.

The workflow applied to drugs, drug similarities and targets showed some promising initial results. After applying node2vec on drugs, 42,021 similarity links were extracted. Drug-target pairs have been separated into two datasets: a test dataset and a train dataset containing 20% and 80% of data, respectively. Next, an hyperparameter optimization of TransE has been applied to drug similarities combined with drug-target connections. The model with the best results has been used for (hidden and missing) link predictions. The embedding algorithm provides a model with all the embeddings and evaluation summaries. For links prediction, two additional inputs are provided in addition to the previous step: an input including all the drug and target names and another input with the relation of interest (*i.e.*, *has_target*). The link prediction process generates predictions containing all the possible combinations with their prediction scores. Scores with the highest number correspond to the best model predictions. So far, the *hit@10* parameter (*i.e.*, the good target appears in the first 10 predictions) reaches 20% for the embedding step. The link prediction results showed some encouraging results. For example, a predicted link connected the drug “Spironolactone” (DrugBank ID: DB00421) and the target “Glutathione synthetase” (DrugBank ID: BE0000185) with a score of 23.62. In a study aiming to reduce acute lung injury, Barut *et al.* studied the effect of spironolactone administration [2]. An increased production of reduced glutathione (being a product of glutathione synthetase) has occurred in the presence of spironolactone. Currently, the evaluation of the generated results is done manually by looking up the scientific literature and ongoing clinical trials asserting the drug-target links plausibility. This process needs to be automated to scale up.

6 CONCLUSION AND FUTURE WORK

AI-based link prediction over KGs is a powerful approach that is producing interesting results in various domains. Our preliminary test has proven to be promising. KGs applied to biomedical data are useful to find new associations and meaningful information. Embedding and link prediction methods are evolving. For example, Liang *et al.* worked on a network of biomedical entities and their relations, and created a cascade learning framework to predict links from a KG with usual embedding algorithms like node2vec, LINE and TransE among others [8]. The technique accuracy reached 93%. Biomedical information is distributed across a wide range of knowledge bases. More data are to be integrated, such as gene regulations and SMILE structures of drugs, which require dedicated processes to be included in the KG. In the precision medicine domain, an existing effort, the SPOKE study[8], has been conducted to highlight the heterogeneous characteristics of biomedical data and tries to integrate and interlink various knowledge bases. It already integrates SIDER, DrugBank, GWAS Catalog and CHEMBL, among others. For DD in particular, Chen *et al.* have used semantic linked data to assess drug-target associations [4]. Our present work is different to those in particular by the fact that it takes into account knowledge derived from natural products. Indeed, a major novelty is the inclusion of such natural products, which have often proved to be a critical starting point in drug design accounting for about 30% of the approved drugs [5].Embedding techniques are used to predict hidden or missing semantic links between drugs and targets leading to the formulation of new hypotheses to be explored with clinical trials or scientific literature. As of perspective, we are considering to improve the workflow and to reduce the number of steps. Further, an automated generated hypothesis validation is expected. A particular emphasis will be made on DD for rare disease treatments.

[8] https://spoke.rbvi.ucsf.edu/

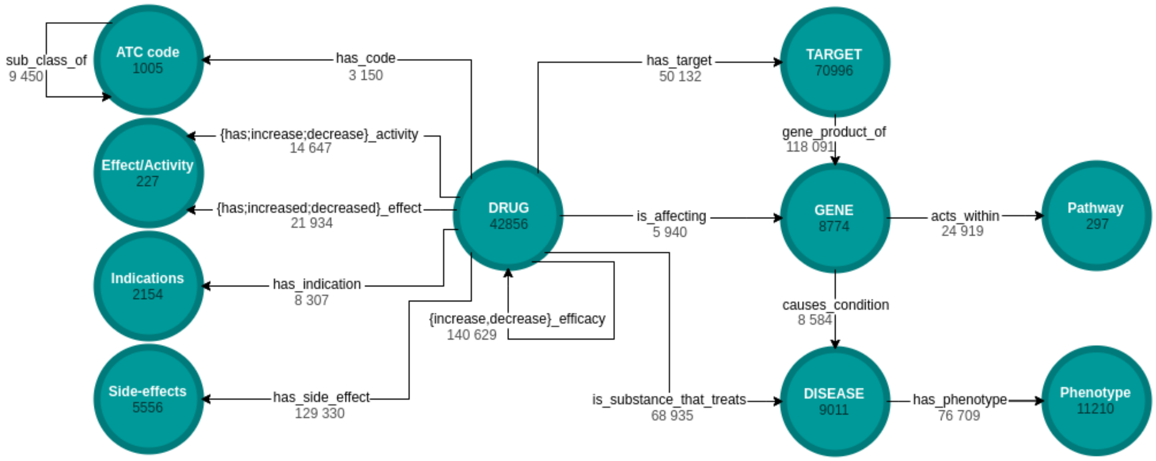

Figure 2: Schema of the current knowledge graph with the number of nodes and relations. Primary entities are represented in uppercase. The numbers in gray under the links correspond to the number of relations of this kind. Numbers under entities indicate the number of such nodes in the knowledge graph.

7 AKNWOLEDGEMENT

We would like to thank Digital Public Health graduate program of University of Bordeaux for granting Marina Boudin the three years funding for the Phd thesis.

REFERENCES

[1] Mehdi Ali, Hajira Jabeen, Charles Tapley Hoyt, and Jens Lehman. 2020. The KEEN Universe: An Ecosystem for Knowledge Graph Embeddings with a Focus on Reproducibility and Transferability. (2020). arXiv:2001.10560 http://arxiv.org/abs/2001.10560

[2] Figen Barut, V. Haktan Ozacmak, Inci Turan, Hale Sayan-Ozacmak, and Erol Aktunc. 2016. Reduction of Acute Lung Injury by Administration of Spironolactone After Intestinal Ischemia and Reperfusion in Rats. *Clinical and Investigative Medicine* (Feb 2016), E15–E24. https://doi.org/10.25011/cim.v39i1.26326

[3] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. 2008. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics* 41, 5 (Oct 2008), 706–716. https://doi.org/10.1016/j.jbi.2008.03.004

[4] Bin Chen, Ying Ding, and David J. Wild. 2012. Assessing Drug Target Association Using Semantic Linked Data. *PLOS Computational Biology* 8, 7 (07 2012), 1–10. https://doi.org/10.1371/journal.pcbi.1002574

[5] Ya Chen, Christina de Bruyn Kops, and Johannes Kirchmair. 2019. Resources for Chemical, Biological, and Structural Data on Natural Products. In *Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research*, A. Douglas Kinghorn, Heinz Falk, Simon Gibbons, Jun'ichi Kobayashi, Yoshinori Asakawa, and Ji-Kai Liu (Eds.). Springer International Publishing, Cham, 37–71. https://doi.org/10.1007/978-3-030-14632-0_2

[6] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, San Francisco, California, USA, 855–864. https://doi.org/10.1145/2939672.2939754

[7] Tom Heath and Christian Bizer. 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology* 1, 1 (Feb 2011), 1–136. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

[8] Xiaomin Liang, Daifeng Li, Min Song, Andrew Madden, Ying Ding, and Yi Bu. 2019. Predicting biomedical relationships using the knowledge and graph embedding cascade model. *PLOS ONE* 14, 6 (2019), e0218264. https://doi.org/10.1371/journal.pone.0218264

[9] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics* 36 (Jan 2016), 1–22. https://doi.org/10.1016/j.websem.2016.01.001

[10] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackerman, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. 2020. A Deep Learning Approach to Antibiotic Discovery. *Cell* 180, 4 (Feb. 2020), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

[11] Sergei Starikov Peter A.C.'t Hoen Dorien J.M. Peters Marco Roos Kristina M. Hettne Tareq B. Malas, Roman Kudrin. 2019. Semantic Knowledge Graph Network Features for Drug Repurposing. *Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences* 2042 (2019). http://ceur-ws.org/Vol-2042/paper12.pdf