Drug repurposing using knowledge graph embeddings with a focus on vector-borne diseases: a model comparison

Diego López Yse[1][0009−0002−8614−1922] and Diego Torres[1,2,3][0000−0001−7533−0133]

[1] Universidad Austral, Buenos Aires, Argentina
[2] LIFIA, CIBPBA-Facultad de Informática, UNLP, La Plata, Argentina
[3] Depto. CyT., Universidad Nacional de Quilmes, Bernal, Argentina

Abstract. Vector-borne diseases carried by mosquitoes, ticks, and other vectors are among the fastest-spreading and most extensive diseases worldwide, mainly active in tropical regions. Also, in the context of the current climate change, these diseases are becoming a hazard for other climatic zones. Hence, drug repurposing methods can identify already approved drugs to treat them efficiently, reducing development costs and time. Knowledge graph embedding techniques can encode biological information in a single structure that allows users to operate relationships, extract information, learn connections, and make predictions to discover potential new relationships between existing drugs and vector-borne diseases. In this article, we compared seven knowledge graph embedding models (TransE, TransR, TransH, UM, DistMult, RESCAL, and ERMLP) applied to Drug Repurposing Knowledge Graph (DRKG), analyzing their predictive performance over seven different vector-borne diseases (dengue, chagas, malaria, yellow fever, leishmaniasis, filariasis, and schistosomiasis), measuring their embedding quality and external performance against a ground-truth. Our analysis found that no single predictive model consistently outperformed all others across all diseases and proposed different strategies to improve predictive performance.

Keywords: Machine Learning · Knowledge Graphs · Knowledge Graph Embeddings · Drug repurposing · Vector-borne diseases · Biotechnology

1 Introduction

Vector-borne diseases are infections spread by the bite of infected arthropod species, such as mosquitoes and ticks. Additionally, certain vectors will have climate change as an ally for their habitat to grow even more, extending their habitable regions beyond the current ones [38]. This phenomenon seems to have materialized recently with autochthonous cases of chikungunya in Italy (2007) and France (2010 and 2014) and dengue in Spain, France, and Italy (2019 and 2020) [41].

For these reasons, the inability to continue responding to such dynamic diseases with traditional drug development methods becomes clear. The traditional drug discovery path is a lengthy and expensive process, estimated to take 14 years, and cost approximately USD 1.8 billion to develop one drug [39]. As a consequence, the process of drug repurposing appears as an alternative to face this challenge.

Drug repurposing refers to identifying new indications for existing, discontinued, or under-development drugs as a way to maximize return on assets that were initially designed with different patient populations in mind. Repurposing drugs is attractive since, generally, drugs approved for an indication are more likely to be safe in a new indication and different patient population [6]. However, this process presents difficulties since, in biomedical research, the data is mainly fragmented and stored in databases that are generally not linked, hampering progress [23]. One way to tackle this problem is by using knowledge graphs, which can exploit diverse, dynamic, large-scale collections of data [24].

A Knowledge graph represents a network of real-world entities—i.e., objects, events, situations, or concepts, and illustrates their relationships [35]. Any object, place, or person can be an entity (or node), and an edge defines the relationship between entities. Knowledge graphs are represented as a collection of ”head entity” - ”relation” - ”tail entity” triples (h, r, t), where entities correspond to nodes and relations to edges between them [48]. This representation can help to gain a contextualized understanding of data, helping to drive automation and process optimization, improve predictions, and enable an agile response to changing environments [5]. In this regard, knowledge graphs can support many biomedical applications [34] by representing diseases, compounds, genes, side-effects, and other related concepts as nodes (or entities), and establishing relationships between them, like for example ”Ibuprofen, treats, Fever”. Here, ”Ibuprofen” and ”Fever” are entities that are connected by the relationship ”treats”. The link or relationship is directed from the entity ”Ibruprofen” to the entity ”Fever”, and is labeled as ”treats”. Entities can have types that further describe them and group them into categories. Following the example, ”Ibuprofen” is a type of Compound, whereas ”Fever” is a type of Disease. This conceptualization allows data integration from various domains, making knowledge graphs a highly flexible data structure [18]. Many biomedical knowledge graphs exist, systematically curating information and facilitating the discovery of new knowledge. Moreover, open-access knowledge graphs like Hetionet [23], PharmKG [49], and DRKG [26] integrate data from genes, drugs, and diseases.

In knowledge graphs terms, drug repurposing can be stated as a link prediction problem. This way, the aim is to predict unknown links between drug and disease entities, where a link between a drug–disease entity suggests that the drug treats the disease [17]. But traditional machine learning approaches applied to solving these challenges have many constraints, including dimensionality and incompleteness, sparsity, and heterogeneity [1].

While the native representation of a knowledge graph is high-dimensional, bringing high computation and space costs, there are methods to project the information into a lower-dimensional latent space that best preserves the graph structure [18] to perform tasks like link prediction more efficiently. Graph embedding methods convert graph data into a low dimensional space where the structural information and properties are maximumly preserved [10], allowing us to state the drug repurposing problem as a link prediction challenge in a vectorized space.

Differences between the various embedding algorithms are related to three aspects: (i) how they represent entities and relations, (ii) how they define the scoring function, and (iii) how they optimize the ranking criterion that maximizes the global plausibility of the existing triples [16]. Some of the embedding models that show state-of-the-art performance in knowledge graph completion relate to translation-based models that treat relations as translation operators over the entities in an embedding space [42] (e.g., TransE, TransR, TransH, UM). Alternatively, semantic matching models that use semantic similarity between entities and relations in the embedding space are commonly used for the task (e.g., DistMult, RESCAL, and ERMLP).

In recent years, embedding methods on knowledge graphs have been mainly focused on accelerating the drug repurposing process for Covid-19 [48], working retrospectively after the health crisis materialized. Reacting in hindsight over these challenges and developing solutions to face an individual disease instead of multiple or groups of diseases limit the impact these technologies might have on saving and improving human lives. Still, few efforts anticipate future health threats in a systematic way, particularly those that will be brought by climate change. Vector-borne diseases are becoming a worldwide hazard since climate alteration is the primary driver of the activity and migration of these vectors.

This paper analyzes the predictive performance of seven popular knowledge graph embedding models (TransE, TransR, TransH, UM, DistMult, RESCAL, and ERMLP) over seven different vector-borne diseases (dengue, chagas, malaria, yellow fever, leishmaniasis, filariasis, and schistosomiasis), measuring their embedding quality and external performance against a ground-truth. Our goal is not to develop the best overall model but to examine the differences in performance between models and diseases to identify gaps, trends, and opportunities in building a comprehensive drug repurposing system. We also validate all predictions against ground truth, analyzing the overall results and exploring further options to enhance our proposal.

This article is structured as follows: Section 2 describes related studies focused on drug repurposing using knowledge graphs. Section 3 introduces the main characteristics of knowledge graphs and the most promising ones related to our research. Section 4 describes our methodological approach to evaluate knowledge graph embedding models. Section 5 details our evaluation method to validate the predictions generated in section 4. Section 6 analyzes and interprets our findings, explaining the significance of our results. Finally, section 7 summarizes our work and establishes routes to enhance our proposal even further.

2 Related Work

Link prediction on knowledge graphs is the task of predicting unseen relations between two existing entities. Recent approaches to this activity rely on knowledge graph embedding methods [47], which learn a mapping from nodes and edges to a continuous vector space, preserving the proximity structure of the knowledge graph to run machine learning algorithms.

Some of the most popular translation-based embedding models used for drug repurposing include (i)TransE [31], which embeds entities and relations in the same vector space, and variants like (ii)TransH and (iii)TransR [13], designed to differently project entities depending on each relation type, meaning that they assign an entity with different representations when involved in various relation types.

Some of the main semantic matching embedding models for drug repurposing include (i)RESCAL [18], which applies a tensor to express the inherent structure of a knowledge graph and uses the rank-d factorization to obtain its latent semantics, (ii)DistMult [16], which simplifies the computational complexity of RESCAL and improves performance, (iii)UM [9], which models relation types similarly as entities and requires less parameters when the number of relation types grows, and (iv)ERMLP [2], a multi-layer perceptron based approach that uses a single hidden layer and represents entities and relations as vectors.

Regarding model predictions validation, while several studies [19] compare predicted drugs against a ground truth like ClinicalTrials.gov [46], others add validation steps (e.g., gene set enrichment analysis to further validate the predicted drug candidates [47]).

Most of the latest efforts in this field focus on predicting drug candidates for individual diseases, specially Covid-19 [17], without extending the analysis to other types of diseases. These approaches not only limit the scope of the proposals (can the predictive models perform well on other diseases?) but also misses identifying potential dynamics between diseases (do the predictive models respond similarly to related diseases?).

Additionally, most studies concentrate on validating the top-n predicted drugs against ground truth [14] for an apparent reason: since only a limited number of compounds can be experimentally screened, knowledge graph embedding models that achieve a high accuracy for the top predicted drug–disease combinations are preferred over models that might achieve a better overall precision but exhibit a lower accuracy among the top predictions [37]. But not analyzing the model's behaviors below those thresholds prevents us from understanding if predictive models that perform similarly within the threshold outperform others below it. Whatsmore, analyzing those patterns can help us to ensemble models that maximize predictive performance among top predictions.

3 Biomedical Knowledge Graphs

Knowledge graphs organize data from multiple sources, capturing information about entities of interest in a given domain or task, using a graph-structured data model or topology to integrate data. Contrary to relational databases, they store nodes and relationships instead of tables or documents. Find next some of the main publicly available biomedical knowledge graphs that can be used for drug repurposing:

- **Hetionet v1.0** [23] is an integrative network encoding knowledge from millions of biomedical studies. It consists of 47,031 nodes of 11 types and 2,250,197 relationships of 24 types. Data were integrated from 29 public resources to connect compounds, diseases, genes, anatomies, pathways, biological processes, molecular functions, cellular components, pharmacologic classes, side effects, and symptoms. The edges represent relationships between these nodes. It uses compiled information from databases for the following entity types: diseases from Disease Ontology [40], symptoms from Medical Subject Headings (MeSH) [30], compounds from DrugBank [45], side effects from SIDER [28], pharmacologic classes from DrugCentral [44], genes from Entrez Gene [32], anatomies from Uberon [33], pathways from WikiPathways [29], and biological processes, cellular components, and molecular from the Gene Ontology [4].
- **PharmKG** [49] is composed of more than 500,000 individual interconnections between genes, drugs, and diseases, with 29 relation types over a vocabulary of approximately 8,000 disambiguated entities. It was constructed based on 6 public databases that offered high-quality structured information, including OMIM [20], DrugBank, PharmGKB [22], Therapeutic Target Database (TTD) [11], SIDER and HumanNet [25], in combination with GNBR [36].
- **Drug Repurposing Knowledge Graph (DRKG)** [26] is a comprehensive biological knowledge graph relating genes, compounds, diseases, biological processes, side effects, and symptoms. It includes information from 6 existing databases: DrugBank, Hetionet, GNBR, String [43], IntAct [21] and DGIdb [15]. It consists of 97,238 entities belonging to 13 entity types; and 5,874,261 triples belonging to 107 edge types. For representing entities, DRKG uses an entity type identifier followed by a unique ID of the specific entity, e.g., Compound::DB00107, refers to the drug "Oxytocin" from the DrugBank database. For representing relations, DRKG uses a naming convention that combines the name of the data source, the name of the relation, and the types of head and tail entities involved, e.g., GNBR::J::Gene:Disease, refers to "a gene that has a role in the pathogenesis of a disease" from the GNBR database.

4 Methodological Approach for Evaluation

To perform a comparison of knowledge graph embedding models for drug repurposing on vector-borne diseases, we follow the protocols used in other works [48] and propose the following evaluation pipeline:

- **Select a knowledge graph**. Since the embedding models that will be evaluated must be run on top of a knowledge graph, the first step is to select a knowledge graph in line with the target problem. Because our target challenge is drug repurposing for vector-borne diseases, a knowledge graph covering those diseases with an adequate volume of triples should be selected.

• **Train and test embedding models on the knowledge graph**. After selecting a knowledge graph, we can train and test different embedding models. By doing so, we map the knowledge graph nodes and edges to a low-dimensional representation (preserving the proximity structure of the knowledge graph) to exploit it for applications like link prediction.

• **Perform link prediction on knowledge graph embeddings**. The embedding models are subsequently used for link prediction tasks to predict new triples or infer missing ones between non-connected nodes within the knowledge graph [37]. The result of this step is a ranked list of predictions.

• **Define a ground truth to validate predictions**. Besides measuring the internal performance of the embedding models, it becomes essential to validate the predicted results against some ground truth. For this reason, a proper source for external validation must be selected, covering the target diseases and an adequate volume of compounds that treat them.

• **Evaluate model predictions against ground truth**. Next, we validate the accuracy of each model prediction for all target diseases by measuring at which ranked position a predicted compound matches a ground truth one. An embedding model that hits all compounds existing on the ground truth source for a given disease using fewer predictions is considered better than another model that needs more predictions to hit the same number of ground truth compounds.

5 Evaluation

Knowledge graph embedding models are usually evaluated based on link prediction. For example, for a given query of the form "X, treats, dengue", the capability of a link predictor to predict the correct entities that answer the query, i.e. "metformin, treats, dengue" is measured. Nevertheless, since true negative examples are not available in our study (both the training and the test set contain only true facts), the evaluation procedure is defined as a ranking task in which the capability of the embedding model to differentiate corrupted triples from known true triples is assessed [2].

5.1 Metrics

For evaluating embedding models, we analyzed the ranking results of each one: (a)intrinsically, within the scope of the knowledge graph and its defined triples, and (b)externally against a ground truth to understand their predictive power over real-world information.

Following prior studies [8], we used two standard rank-based metrics to measure each embedding model's intrinsic performance on link prediction: Adjusted Mean Rank (AMR) and hits@k.

- **Adjusted Mean Rank(AMR)** [7] is the ratio of the Mean Rank to the Expected Mean Rank, assessing a model's performance independently of the underlying set size. It lies on the open interval (0,2), where lower is better.

- **Hits@k** measures the fraction of times when the correct or "true" entity appears under the top-k entities in the ranked list. The value of hits@k is between 0 and 1. The larger the value, the better the model works [12]. For our work, we estimated hits@1, hits@3, hits@5, and hits@10 metrics, which were calculated using Python's PyKEEN library [3].

To validate the embedding models externally, we analyzed the predicted ranked compound list against the actual treatment drugs defined in ground truth for those diseases using the following metrics:

- **First hit** is the ranking position at which compounds proposed by an embedding model match one from the ground truth database for a given disease.
- **Median hit** is the ranking position at which compounds proposed by an embedding model match 50% of the compounds from the ground truth database for a given disease.
- **Last hit** is the ranking position at which compounds proposed by an embedding model match all the compounds from the ground truth database for a given disease.

For all these metrics, the smaller the value, the better, meaning that a model with lower "first", "median", or "last hit" values compared to another one, matches real-world compounds using fewer predictions.

5.2 Data

Although several embedding models were identified in literature review [16], only a subset was selected due to computational constraints related to high training costs. For this reason, we evaluated seven popular knowledge graph embedding models, covering both translational distance (TransE, TransR, TransH, and UM) and semantic matching (DistMult, RESCAL, and ERMLP). These models were applied to seven vector-borne diseases (dengue, chagas, malaria, yellow fever, leishmaniasis, filariasis, and schistosomiasis) on DRKG dataset. We used ClinicalTrials.gov as our source of ground truth and evaluated each model's performance. All the code for this study is available via open-source licensing on GitHub at: `https://github.com/dlopezyse/Drug-Repurposing-using-KGE`

Following the methodological approach described in section 4, first, we selected DRKG as our knowledge graph due to the volume of triples and the high representation of compound-disease interactions.

Next, we used PyKEEN 1.10 pipeline to train, test and validate all embedding models with 50 epochs, Marging Ranking as the loss function, and random seed = 1234 as the only predefined parameters. We configured Google Colab GPUs to run the models, and no hyperparameter optimization was performed due to computational constraints. All models were evaluated using Adjusted Mean Rank(AMR) and hits@k measures.

Within DRKG, we focused on the GNBR database (the most extensive for our target problem) and defined "GNBR::T::Compound:Disease" as the target relation to predict compounds that treat the mentioned diseases. We performed link predictions for each target disease using all previously-mentioned embedding models, resulting in seven different compound predictions by disease. The result of all embedding model predictions for each disease (49 in total) was a ranked list of the 97,238 DRKG entities, ordered by their predicted effectiveness in treating the target disease. Afterward, all predicted compound IDs were mapped to their original data sources to identify their names.

We used ClinicalTrials.gov as an external validation point to measure the quality of our predictions. After extracting all compounds used in ClinicalTrials.gov for the seven target diseases, compound names were normalized against those from our model's predictions. For example, in the case of malaria, "Paracetamol" (defined as a compound in ClinicalTrials.gov to treat the disease) was mapped to "Acetaminophen" (which exists in DRKG).

5.3 Results

As detailed in table 1, performance metrics results were compared across all models, with the best outcomes highlighted in light blue.

Model	AMR	Hits@1	Hits@3	Hits@5	Hits@10
TransE	0.023	0.008	0.066	0.091	0.132
TransH	0.062	0.006	0.013	0.021	0.039
TransR	0.019	0.012	0.063	0.088	0.131
DistMult	0.039	0.015	0.035	0.049	0.076
RESCAL	1.000	0.000	0.000	0.000	0.000
UM	0.044	0.000	0.034	0.049	0.074
ERMLP	0.018	0.027	0.079	0.109	0.163

Table 1. Embedding models metrics table.

Results show that ERMLP was the best performer both on AMR (scoring the lowest value of 0.018) and hits@k metrics (exhibiting the highest values on all measures) from the seven embedding models, while RESCAL was the worst on all of them (scoring 1.000 on AMR and 0.000 on all hits@k measures).

We also collected all reported compounds for treating the seven target diseases from the ClinicalTrials.gov database to test the embedding model results against ground truth. Next, we validated whether the previously mentioned metrics represented real-world performance.

To evaluate the performance of this method, we reported the position at which compounds proposed by the models matched the ones from the ClinicalTrials.gov database for a given disease. Since we used DRKG, the best possible position for a single predicted compound would be 1 (meaning that the first predicted compound by the model matched one defined in ClinicalTrials.gov for the same disease), and the worst one 97,238 (the total number of entities in DRKG). We also considered the number of compounds identified in ClinicalTrials.gov for a given disease that existed in DRKG (matching compounds). This way, if we had 5 compounds that existed both in DRKG and ClinicalTrials.gov for a given disease, the best possible model would predict those compounds from positions 1 to 5, and the worst one, from positions 97,234 to 97,238.

Table 2 details the results for one of the seven targeted diseases (filariasis), highlighting the best results in light blue.

Filariasis - matching compounds: 11			
Model	**First hit**	**Median hit**	**Last hit**
TransE	117	716	6,798
TransH	1,048	29,513	61,479
TransR	1	238	5,381
DistMult	37	474	22,054
RESCAL	8,703	44,404	64,470
UM	19	4,448	86,238
ERMLP	164	894	58,562

Table 2. Filariasis metrics table.

Following these outcomes, we observed that 11 compounds were identified in ClinicalTrials.gov to treat filariasis that also existed in DRKG. TransR was the best-performing model by reaching a first hit (”albendazole”) at position 1 of its 97,238 ranked predictions, a median hit using 238 predictions, and matching all 11 compounds with 5,381 ranked predictions.

We also present in table 3 the top 5 drug repurposing candidates for filariasis, predicted by their best-performing models as identified in table 2. We detail the compound name, its ranked position according to the model and highlight in green compounds that were learned during the model training process:

Filariasis: top 5 predictions			
TransR		**DistMult**	
Drug	**Position**	**Drug**	**Position**
Albendazole	1	Doxycycline	37
Azithromycin	12	Praziquantel	177
Praziquantel	17	Azithromycin	230
Rifampicin	37	Albendazole	245
Ivermectin	153	Ivermectin	373

Table 3. Filariasis predicted compounds.

Results show that from the 11 matching compounds between DRKG and ClinicalTrials.gov for filariasis, the TransR model identified the first 5 using 153 ranked compounds, while DistMult did it using 373. Interestingly, while

DistMult needed 37 ranked compounds to get the first hit, TransR got four matching compounds with the same number of predictions.

6 Discussion

When measuring internal model predictions using hits@k and AMR metrics, ERMLP was identified as the best-performing model. Nevertheless, when comparing model predictions against an external ground truth, results were diverse, except for the RESCAL model showing the worst performance consistently on all diseases.

Several models matched partial compounds against ground truth using less than 100 ranked predictions (e.g., TransE for dengue, TransR for malaria and filariasis). What's more, these matchings include compounds that were not learned during the model's training processes.

For diseases with a low or zero percentage of ground truth compounds in DRKG available for training (e.g., yellow fever), embedding models significantly underperformed and required thousands of ranked predictions to reach first matches against ground truth.

The methodological approach described in section 4 can be easily tested for enhancements incorporating techniques like hyperparameter optimization, graph filtering, and different loss functions for model training. From a modeling perspective, additional embedding models should be explored, and different combinations of dataset splits and loss functions to identify their impact on performance.

Additionally, it is possible to restrict model performance evaluation to our relations of interest (e.g., "Compound treats Disease") to identify models more appropriate for drug repositioning during hyperparameter optimization instead of good ones at predicting all types of relations.

Finally, expanding the ground truth to other databases besides ClinicalTrials.gov might help to identify additional hits and potential drugs for repurposing.

7 Conclusions and Further Work

This article compared seven knowledge graph embedding models applied to DRKG, focusing on seven specific vector-borne diseases. We introduced an evaluation pipeline to assess the embedding models for drug repurposing tasks, measuring their performance using internal metrics and a ground truth source.

Ensembling strategies should be explored by combining multiple knowledge graphs embedding models to exploit their complementary aspects.

Alternative performance evaluation metrics like AUROC, Precision-Recall curve, F1 score, Mean Reciprocal Rank, NDCG, or Average Precision should be considered for broader model comparison and understanding.

Lastly, developing additional evaluation criteria (e.g., molecular analysis [27]) could further increase the success rate of drug-repurposed candidates in laboratory validation.

References

1. Abbas, K., Abbasi, A., Dong, S., Niu, L., Yu, L., Chen, B., Cai, S.M., Hasan, Q.: Application of network link prediction in drug discovery. BMC Bioinformatics **22**(1) (Apr 2021). https://doi.org/10.1186/s12859-021-04082-y, `https://doi.org/10.1186/s12859-021-04082-y`
2. Ali, M., Berrendorf, M., Hoyt, C.T., Vermue, L., Galkin, M., Sharifzadeh, S., Fischer, A., Tresp, V., Lehmann, J.: Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 8825–8845 (Dec 2022). https://doi.org/10.1109/tpami.2021.3124805, `https://doi.org/10.1109/tpami.2021.3124805`
3. Ali, M., Berrendorf, M., Hoyt, C.T., Vermue, L., Sharifzadeh, S., Tresp, V., Lehmann, J.: PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. Journal of Machine Learning Research (2021)
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (May 2000). https://doi.org/10.1038/75556, `https://doi.org/10.1038/75556`
5. Barrasa, J., Hodler, A.E., Webber, J.: Knowledge Graphs. O'Reilly Media (2021)
6. Barratt, M.J., Frail, D. (eds.): Drug repositioning: bringing new life to shelved assets and existing drugs. John Wiley & Sons, Hoboken, N. J (2012)
7. Berrendorf, M., Faerman, E., Vermue, L., Tresp, V.: On the ambiguity of rank-based evaluation of entity alignment or link prediction methods (2020)
8. Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Hoyt, C.T., Hamilton, W.L.: Understanding the performance of knowledge graph embeddings in drug discovery. Artificial Intelligence in the Life Sciences **2**, 100036 (Dec 2022). https://doi.org/10.1016/j.ailsci.2022.100036, `https://doi.org/10.1016/j.ailsci.2022.100036`
9. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: A semantic matching energy function for learning with multi-relational data. Machine Learning **94**(2), 233–259 (May 2013). https://doi.org/10.1007/s10994-013-5363-6, `https://doi.org/10.1007/s10994-013-5363-6`
10. Cai, H., Zheng, V.W., Chang, K.C.C.: A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications (Feb 2018), `http://arxiv.org/abs/1709.07604`, number: arXiv:1709.07604 arXiv:1709.07604 [cs]
11. Chen, X.: TTD: Therapeutic target database. Nucleic Acids Research **30**(1), 412–415 (Jan 2002). https://doi.org/10.1093/nar/30.1.412, `https://doi.org/10.1093/nar/30.1.412`
12. Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., Duan, Z.: Knowledge graph completion: A review. IEEE Access **8**, 192435–192456 (2020). https://doi.org/10.1109/access.2020.3030076, `https://doi.org/10.1109/access.2020.3030076`
13. Choi, W., Lee, H.: Inference of Biomedical Relations Among Chemicals, Genes, Diseases, and Symptoms Using Knowledge Representation Learning. IEEE Access **7**, 179373–179384 (2019). https://doi.org/10.1109/ACCESS.2019.2957812, `https://ieeexplore.ieee.org/document/8931752/`

14. Cohen, S., Hershcovitch, M., Taraz, M., Kißig, O., Issac, D., Wood, A., Waddington, D., Chin, P., Friedrich, T.: Improved and optimized drug repurposing for the sars-cov-2 pandemic (2022). https://doi.org/10.1101/2022.03.24.485618
15. Cotto, K.C., Wagner, A.H., Feng, Y.Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies, N.C., Griffith, O.L., Griffith, M.: DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Research **46**(D1), D1068–D1073 (Nov 2017). https://doi.org/10.1093/nar/gkx1143, https://doi.org/10.1093/nar/gkx1143
16. Dai, Y., Wang, S., Xiong, N.N., Guo, W.: A survey on knowledge graph embedding: Approaches, applications and benchmarks. Electronics **9**(5) (2020). https://doi.org/10.3390/electronics9050750, https://www.mdpi.com/2079-9292/9/5/750
17. Doshi, S., Chepuri, S.P.: A computational approach to drug repurposing using graph neural networks. Computers in Biology and Medicine **150**, 105992 (Nov 2022). https://doi.org/10.1016/j.compbiomed.2022.105992, https://doi.org/10.1016/j.compbiomed.2022.105992
18. Florin Ratajczak, et al.: Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. BMC Bioinformatics **23**(84) (2022). https://doi.org/https://doi.org/10.1186/s12859-022-04608-y, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04608-y
19. Gao, Z., Ding, P., Xu, R.: KG-predict: A knowledge graph computational framework for drug repurposing. Journal of Biomedical Informatics **132**, 104133 (Aug 2022). https://doi.org/10.1016/j.jbi.2022.104133, https://doi.org/10.1016/j.jbi.2022.104133
20. Hamosh, A.: Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research **33**(Database issue), D514–D517 (Dec 2004). https://doi.org/10.1093/nar/gki033, https://doi.org/10.1093/nar/gki033
21. Hermjakob, H.: IntAct: an open source molecular interaction database. Nucleic Acids Research **32**(90001), 452D–455 (Jan 2004). https://doi.org/10.1093/nar/gkh052, https://doi.org/10.1093/nar/gkh052
22. Hewett, M.: PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Research **30**(1), 163–165 (Jan 2002). https://doi.org/10.1093/nar/30.1.163, https://doi.org/10.1093/nar/30.1.163
23. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife **6**, e26726 (Sep 2017). https://doi.org/10.7554/eLife.26726, https://elifesciences.org/articles/26726
24. Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.C.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A.: Knowledge Graphs. ACM Computing Surveys **54**(4), 1–37 (May 2022). https://doi.org/10.1145/3447772, https://dl.acm.org/doi/10.1145/3447772
25. Hwang, S., Kim, C.Y., Yang, S., Kim, E., Hart, T., Marcotte, E.M., Lee, I.: HumanNet v2: human gene networks for disease research. Nucleic Acids Research **47**(D1), D573–D580 (Nov 2018). https://doi.org/10.1093/nar/gky1126, https://doi.org/10.1093/nar/gky1126

26. Ioannidis, V.N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., Karypis, G.: DRKG - Drug Repurposing Knowledge Graph for Covid-19 (2020)
27. Islam, M.K., Amaya-Ramirez, D., Maigret, B., Devignes, M.D., Aridhi, S., Smaïl-Tabbone, M.: Molecular-evaluated and explainable drug repurposing for COVID-19 using ensemble knowledge graph embedding. Scientific Reports **13**(1) (Mar 2023). https://doi.org/10.1038/s41598-023-30095-z, `https://doi.org/10.1038/s41598-023-30095-z`
28. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Research **44**(D1), D1075–D1079 (Oct 2015). https://doi.org/10.1093/nar/gkv1075, `https://doi.org/10.1093/nar/gkv1075`
29. Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E.L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S.R., Miller, R., Coort, S.L., Cirillo, E., Smeets, B., Evelo, C.T., Pico, A.R.: WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Research **44**(D1), D488–D494 (Oct 2015). https://doi.org/10.1093/nar/gkv1024, `https://doi.org/10.1093/nar/gkv1024`
30. Lipscomb, C.E.: Medical subject headings (MeSH). Bull. Med. Libr. Assoc. **88**(3), 265–266 (Jul 2000)
31. Ma, C., Liu, H., Zhou, Z., Koslicki, D.: Predicting drug repurposing candidates and their mechanisms from a biomedical knowledge graph. bioRxiv (2022). https://doi.org/10.1101/2022.11.29.518441, `https://www.biorxiv.org/content/early/2022/12/02/2022.11.29.518441`
32. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez gene: gene-centered information at NCBI. Nucleic Acids Research **39**(Database), D52–D57 (Nov 2010). https://doi.org/10.1093/nar/gkq1237, `https://doi.org/10.1093/nar/gkq1237`
33. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A.: Uberon, an integrative multi-species anatomy ontology. Genome Biology **13**(1), R5 (2012). https://doi.org/10.1186/gb-2012-13-1-r5, `https://doi.org/10.1186/gb-2012-13-1-r5`
34. Nicholson, D.N., Greene, C.S.: Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal **18**, 1414–1428 (2020). https://doi.org/10.1016/j.csbj.2020.05.017, `https://linkinghub.elsevier.com/retrieve/pii/S2001037020302804`
35. [Online]: What is a Knowledge Graph? — IBM — ibm.com. `https://www.ibm.com/topics/knowledge-graph`, [Accessed 14-Mar-2023]
36. Percha, B., Altman, R.B.: A global network of biomedical relationships derived from text. Bioinformatics **34**(15), 2614–2624 (Feb 2018). https://doi.org/10.1093/bioinformatics/bty114, `https://doi.org/10.1093/bioinformatics/bty114`
37. Rivas-Barragan, D., Domingo-Fernández, D., Gadiya, Y., Healey, D.: Ensembles of knowledge graph embedding models improve predictions for drug discovery. Briefings in Bioinformatics **23**(6) (Nov 2022). https://doi.org/10.1093/bib/bbac481, `https://doi.org/10.1093/bib/bbac481`
38. Ryan, S.J., Carlson, C.J., Mordecai, E.A., Johnson, L.R.: Global expansion and redistribution of Aedes-borne virus transmission risk with climate change. PLOS Neglected Tropical Diseases **13**(3), e0007213 (Mar 2019). https://doi.org/10.1371/journal.pntd.0007213, `https://dx.plos.org/10.1371/journal.pntd.0007213`
39. Sang, S., Yang, Z., Liu, X., Wang, L., Lin, H., Wang, J., Dumontier, M.: GrEDeL: A Knowledge Graph Embedding Based Method for Drug

Discovery From Biomedical Literatures. IEEE Access **7**, 8404–8415 (2019). https://doi.org/10.1109/ACCESS.2018.2886311, https://ieeexplore.ieee.org/document/8574025/

40. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research **40**(D1), D940–D946 (Nov 2011). https://doi.org/10.1093/nar/gkr972, https://doi.org/10.1093/nar/gkr972
41. Semenza, J.C., Paz, S.: Climate change and infectious disease in Europe: Impact, projection and adaptation. The Lancet Regional Health - Europe **9**, 100230 (Oct 2021). https://doi.org/10.1016/j.lanepe.2021.100230, https://linkinghub.elsevier.com/retrieve/pii/S2666776221002167
42. Song, H.J., Kim, A.Y., Park, S.B.: Learning translation-based knowledge graph embeddings by n-pair translation loss. Applied Sciences **10**(11) (2020). https://doi.org/10.3390/app10113964, https://www.mdpi.com/2076-3417/10/11/3964
43. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., von Mering, C.: STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research **47**(D1), D607–D613 (Nov 2018). https://doi.org/10.1093/nar/gky1131, https://doi.org/10.1093/nar/gky1131
44. Ursu, O., Holmes, J., Knockel, J., Bologa, C.G., Yang, J.J., Mathias, S.L., Nelson, S.J., Oprea, T.I.: DrugCentral: online drug compendium. Nucleic Acids Research **45**(D1), D932–D939 (Oct 2016). https://doi.org/10.1093/nar/gkw993, https://doi.org/10.1093/nar/gkw993
45. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research **46**(D1), D1074–D1082 (Nov 2017). https://doi.org/10.1093/nar/gkx1037, https://doi.org/10.1093/nar/gkx1037
46. Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M., Ide, N.C.: The ClinicalTrials.gov results database — update and key issues. New England Journal of Medicine **364**(9), 852–860 (Mar 2011). https://doi.org/10.1056/nejmsa1012065, https://doi.org/10.1056/nejmsa1012065
47. Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., Zhang, Z., Li, K., Karypis, G., Cheng, F.: Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. Journal of Proteome Research **19**(11), 4624–4636 (Nov 2020). https://doi.org/10.1021/acs.jproteome.0c00316, https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00316
48. Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., Kilicoglu, H.: Drug repurposing for COVID-19 via knowledge graph completion. Journal of Biomedical Informatics **115**, 103696 (Mar 2021). https://doi.org/10.1016/j.jbi.2021.103696, https://www.sciencedirect.com/science/article/pii/S1532046421000253
49. Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E.F., Yang, Y., Niu, Z.: PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in Bioinformatics **22**(4) (12 2020). https://doi.org/10.1093/bib/bbaa344, https://doi.org/10.1093/bib/bbaa344, bbaa344