Demystifying Drug Repurposing Domain Comprehension with Knowledge Graph Embedding

Edoardo Ramalli*, Alberto Parravicini*, Guido W. Di Donato*, Mirko Salaris*
Céline Hudelot†, Marco D. Santambrogio*

*Politecnico di Milano, DEIB, Milan, Italy, †Université Paris-Saclay CentraleSupeléc, MICS Lab, Gif-sur-Yvette, France
*{edoardo.ramalli, alberto.parravicini, guidowalter.didonato, mirko.salaris, marco.santambrogio}@polimi.it
†celine.hudelot@centralesupelec.fr

Abstract—**Drug repurposing is more relevant than ever due to drug development's rising costs and the need to respond to emerging diseases quickly. Knowledge graph embedding enables drug repurposing using heterogeneous data sources combined with state-of-the-art machine learning models to predict new drug-disease links in the knowledge graph. As in many machine learning applications, significant work is still required to understand the predictive models' behavior. We propose a structured methodology to understand better machine learning models' results for drug repurposing, suggesting key elements of the knowledge graph to improve predictions while saving computational resources. We reduce the training set of 11.05% and the embedding space by 31.87%, with only a 2% accuracy reduction, and increase accuracy by 60% on the open ogbl-biokg graph adding only 1.53% new triples.**

Index Terms—**Drug Repurposing, Biomedical Knowledge Graph, Knowledge Graph Embedding, Link Prediction, Machine Learning**

I. INTRODUCTION

Discovering new drugs is a tricky, expensive, and slow mission. It involves different stages that often require clinical trials to move forward. Drug Repurposing (DR) discovers new therapeutic uses for existing drugs, reducing time-to-market by 30% to 80% (Figure 1) and cost (∼80%) with a lower failure risk in the trails compared to a new chemical entity [1]. In the last years, the number of available biomedical information increased to the point that it is arguably impossible to manage it manually [2]. Fortunately, automatic procedures significantly benefit from integrating heterogeneous information from different sources of data. This challenge has spurred significant research in Biomedical Informatics, intersecting disciplines such as data integration and representation learning. While results are promising, formal methodologies taking into account biomedical knowledge are essential to better understand these outcomes [3], [4]. We applied Machine Learning (ML) to DR, and, to achieve higher interpretability, we investigated the meaning behind the prediction scores with different analyses. We propose a methodology to understand how the network structure representing a biomedical domain influences the prediction accuracy in graph representation learning applications. As a result, we can leverage this information to improve the quality of the network and subsequently improve the predictions and reduce the computational resources needed to learn the representation of the biomedical domain.

A Knowledge Graph (KG) is a network of heterogeneous entities connected by specific relationships capable of representing a complex domain semantic [4]. Encoding the biomedical domain in a KG translates the task of DR into finding possible new connections between a drug and a disease [5]. Reusing a pre-existent biomedical KG is difficult because databases often restrict data redistribution without a commercial license. We represent a biomedical domain that is beneficial for DR, combining different available sources of information. We leverage different representation learning techniques to extract knowledge from structured and unstructured data from curated databases such as Uniprot [6], CTD [7], DrugBank [8], and OMIM [9], pursuing the importance of data sharing in the biomedical field [10].

Knowledge Graph Embeddings (KGEs) are representations obtained through ML techniques that project the KG to a lower-dimensional space that preserves the graph structure [11]. KGEs can predict new relationships between entities in the graph: in the context of DR, we can leverage embeddings to discover new links between drugs and diseases.

Transparency, interpretability, and explainability are long-standing issues in the application of ML to natural sciences [12]. For this reason, we studied the quality of DR predictions obtained from KGEs with different types of analysis, combining domain knowledge with the ML outcomes. Our methodology also hints at strategies to reduce the computational cost of KGEs, with insignificant accuracy detriment.

This work can be valuable to the pharmaceutical industry, having the potential to speed up the *compound identification* stage of the DR approach, which can require up to 2 years of work (Figure 1). Moreover, our methodology reduces the KGE representation size on a novel KG by 31.87% (with a comparable reduction in training time), with only a 2% accuracy reduction, and increases the accuracy by 60% on the open `ogbl-biokg` (BioKG) graph [13] with the addition of only 1.53% new triples.

Our main contributions are:

- A new biomedical KG built from free databases specifically to address DR (Section II-C).
- A methodology to analyze the efficacy of embedding models for DR while saving resources (Section II-D).
- An analysis of the relationship between the prediction quality and data structure, discovering which links and

Fig. 1: Drug Discovery and Drug Repurposing timeline.

Fig. 2: The schema of our biomedical KG with four different types of entities and specific relationships connecting them.

entities are more incisive in DR (Section III).

II. DATA & METHODS

The literature offers several embedding models for KGs [14] and describes various biomedical databases used for different applications, in particular to apply Graph Machine Learning (GML) techniques to DR [5], [15].

This section introduces the general idea of the embedding model used to support DR and presents the used datasets.

A. Knowledge Graph and Knowledge Graph Embedding

A KG is a network that specifies the type of connection between two entities [16]: Figure 2 represents the schema of the KG built for this work. KGs are commonly represented as lists of triples. A triple is composed of three elements: two entities, called head h and tail t, and a relationship r that connects them. Due to the KG structure, it is easy to integrate the graph with heterogeneous information sources.

KGE models represent the KG in a lower-dimensional space that condenses and preserves the original information but also allows extracting hidden information. A KGE differs from another by the representation space, the scoring function, and the encoding models' additional features [14]. The embedding model uses a representation space with a mathematical structure to encode peculiar relation properties of the KG into a low-dimension representation vector. The model score function measures the embedded triple's plausibility together with additional information from the graph.

Link Prediction (LP) is the task of predicting facts in a KG to forecast the existence of a missing triple, leveraging the learned embedded representation [14]. As such, DR can be seen as LP between a drug and a disease.

B. Selecting an Embedding Model

Multiple state-of-the-art KGE techniques exist, without a single one outperforming the others in similar conditions [14], [17]. For our application, we found that TransE is the best compromise between computational complexity and prediction accuracy. TransE requires fewer data and parameters than competing embedding models to provide excellent accuracy [14]; moreover, it can also scale to more extensive databases, making it suitable for easy prototyping. TransE models relationships by interpreting them as translations operating on the entities' low-dimensional embeddings [18]. Given a training set $\mathbf{S}$ of triples $\{(h, r, t) \in \mathbf{S}\}$, TransE learns for each entity h and t, and for each relation r, a vector representation of chosen size K. The primary idea behind TransE is that the functional relation induced by the r-labeled edges corresponds to a translation of the embeddings. The vector sum of $h + r = t$ when (h, r, t) is a true triple (t should be the nearest neighbor of $h + r$), while $h + r$ should be far away from t if the triple is false, i.e. a negative triple, denoted as $(h', r, t') \in \mathbf{S}'$, with $S' \cap S = \emptyset$. TransE, to learn such embedding, minimizes a loss function $\mathcal{L}$ (Eq. 1), computed as sum of dissimilarity measure d (L1 or L2 norm) over the training set [18].

$$\mathcal{L} = \sum_{(h,r,t) \in \mathbf{S}} \sum_{(h',r,t') \in \mathbf{S}'} [d(h + r, t) - d(h' + r, t')]_+ \quad (1)$$

To predict a new link in the graph, the model replaces a triple's element with a specific subset of entities, and it ranks each new triple using the cost function against the learned embedding. The top-ranking triples are plausible triples, and for this reason, possible new connections in the KG [18].

C. Integrating Heterogeneous Sources of Data

To represent the heterogeneous information of a biomedical domain as a KG, it is necessary to start from curated databases [15]. To address DR, we propose a KG which is the result of the combination of heterogeneous sources of information. We combine free biomedical curated databases composed of unstructured information as in DrugBank [8] and UniProt [6],

and structured information as in CTD [7], and OMIM [9]. In the case of unstructured information fields, we use the biomedical `en_ner_bc5cdr_md` Named Entity Recognition (NER) system [19] to extract meaningful connections between entities from the textual content. Although the KG is easily extensible, the integration process is not immediate due to the naming complexity of the different entities. The same entity can have more commercial names, or a database can use a different naming system that made it necessary to use dictionary databases, like OMIM, to homogenize diverse names. The result is a compact, effective KG built to address the problem of DR. We use a second biomedical KG, BioKG [13]. This graph contains more entities and relations than ours, but it is not built to specifically target DR. In Table I, there are the characteristics of the two biomedical KG. In particular, we observe that our KG has fewer triples overall, but it has more drug-disease relationships that are useful for DR. This consideration holds true even if BioKG has other drugs-related information such as side effect entities.

D. Proposed analysis methodology

We propose the following procedure to analyze the embedding applied on KG as a result of the experimental results in Section III: ① Set the embedding model with Hyper Parameters (HPs) coming from a KG with similar size and domain. This step helps to reduce the HPs research space. ② Optimization of HPs, in such a way that the fine-tuning of the embedding model produces the best prediction performances. ③ Carry out feature ablation or extension analyses that determine if the model is learning and not memorizing. This analysis aims to highlight the KG semantic and structural strengths and criticalities. Leveraging this information, we can understand if there are not such helpful parts of the KG but that have a non-negligible impact on the resources used for the training procedure.

III. EXPERIMENTAL EVALUATION

This section presents the study results on the model hyperparameters, then shows the prediction accuracy for DR for both datasets, and finally investigates these results with feature ablation and graph extension techniques. We randomly split drug-disease triples into training, validation, and test set with a probability of 60%, 20%, 20%, respectively. Non-drug-disease triples are also added to the training set. Focusing on the DR task, we measure the embedding model's prediction accuracy using Hits@N (H@N) only against the drug-disease triples present in the validation set. H@N is the proportion of the original correct triples to the top N predictions of the model (Eq. 2). Other public results instead provide general results on all possible link predictions [13]. To compute this metric, we follow the procedure described in [13], [18]. The process consists of corrupting the validation set's triples Q, by replacing one entity of the triple with another entity from a random subset of the same type. The embedding model ranks

(a) Embedding Dimension

(b) Negative Sampling

Fig. 3: Importance of the embedding dimension and negative sampling for prediction score and computational resources.

all the corrupted triples and the valid triples against each other with H@N.

$$Hits@N = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \begin{cases} 1 & \text{if } rank_{(h,r,t)_i} \leq N \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

We average H@N scores over 10 independent training/validation cycles, with negligible variance.

A. Hyper Parameters

The embedding model requires HPs tuning to be trained effectively on a specific dataset. The two most important HPs in a KGE, according to [14], are the embedding dimension and the optimizer (with its learning rate). Other parameters, like the negative sampling, can affect the time to train a model, but they are less significant for the final accuracy. In particular, in Figure 3b, negative sampling is difficult to manage since it is hard to have a negative set (a set of false triples) available. Perturbing the triples randomly is challenging as there is no certainty that this is not a possible repurposed drug, and inserting it in the negative set would indicate to the model to penalize an actually correct representation of the triple. For this reason, we choose a low value for the negative sampling that reduces the probability of this event and saves computational resources. The result shows that a low embedding dimension yields the worst accuracy. Instead, an exaggerated embedding

TABLE I: Number of entities/relations (and percentage of total) in each KGs in our evaluation. Highest values in bold.

	Diseases	Drugs	Genes	Proteins	Side Effects	Functions	Total Entities	Drug-Disease Triples	Drug-Drug Triples	Disease-Disease Triples	Total Triples
Our KG	4596 (7.32%)	**13945 (22.21%)**	**20017 (31.87%)**	**24242 (38.6%)**	0	0	62800	**72976 (39.9%)**	0	**21913 (11.97%)**	183000
ogb-biokg	**10687 (11.20%)**	10533 (11.23%)	0	17499 (18.66%)	**9969 (10.63%)**	**45085 (48.08%)**	**93773**	5147 ($<$ 1%)	**1133686 (23.82%)**	0	**4760000**

(a) Optimizer

(b) Learning Rate

Fig. 4: Importance of the optimizer and learning rate for prediction score and computational resources.

Fig. 5: Prediction H@N score for TransE applied to our KG and to BioKG, compared to a random predictor.

Fig. 6: Comparing features ablation for H@10 accuracy.

dimension does not bring benefits but only faster overfitting and a higher computational cost. For the KG proposed in this work, the best embedding size is 128 (Figure 3a) since it is the best compromise between accuracy and model complexity. The best optimizer proved to be ADAM with learning rate $\lambda = 10^{-4}$, coherently to [17] (Figures 4a and 4b).

B. Drug Repurposing Prediction Score

TransE applied to our biomedical KG achieves a H@N score slightly above 52%. In other words, the model proposes a correct repurposing in the first ten predictions 52% of the times. This result shows that our model has significant learning capabilities: a random baseline (10 random drugs chosen as repurposing candidates) always has H@N close to 0% due to the enormous number of possible combinations.

The same embedding algorithm applied to BioKG gives H@N accuracy below 30%, when predicting the same drug-disease relations. BioKG contains more triples than our KG but less useful information for DR. From these promising results (Figure 5), we investigate how the KGE structure relates to such different accuracies, in the next section.

C. Feature Ablation

To understand which parts of the input data are more critical in the training procedure, we systematically apply *feature ablation* to our KG, reducing its size and entity types. This methodology highlights which part of the KG is critical to the DR task and sheds a light on how the KGE model relates to the graph structure, the model outcomes, and the domain knowledge. A summary of the results is in Figure 6.

1) Only Drug-Disease: If the dataset contains only drug-disease relationships, the accuracy score drops below 16%.

This striking accuracy loss shows that the other removed triples are essential for good predictions.

2) No gene: Removing gene entities from the KG results in a $\sim 2\%$ H@N loss. Although gene entities represent more than 30% of the entities in the KG, they do not appear to be essential for DR. This result suggests an important consequence: bigger KG, with a higher computational cost for training and inferences, do not provide a significant improvement of DR accuracy if they do not contain meaningful triples for the problem at hand.

3) No Parents: Another test is to remove from the dataset the relation that connects a disease to another one. This kind of relation expresses a hierarchy between diseases: a disease can be classified based on its specification, but it belongs to a more generic family group. This information can be helpful because it is very likely that a drug that treats a disease could benefit a similar drug that belongs to the same family. This idea is confirmed by the results of the H@10 score on the validation set. If the disease-disease relations, representing 10% of the dataset, are removed from the training set, accuracy decreases by 5-10%.

4) 50% Triples Removal: Randomly removing 50% of the triples has a clear impact on the accuracy result. The results show that increasing the number of information in the KG improves the model's accuracy significantly even if the score is not proportional to the graph's dimension.

D. Extension of `ogbl-biokg` *(BioKG)*

Extending BioKG with other drug-disease relations used in our KG, with the support of dictionary databases used to translate the entity references, improves the accuracy of $\sim 60\%$ as shown in Figure 5. This result indicates that the more useful data is available for a specific task, the more accurate the model will be in the prediction.

IV. CONCLUSIONS

The results presented in this work allow us to conclude that, in the case of KGEs, what most influences a prediction task is the graph's structure. Our methodology also significantly reduces the computational resources necessary to train the model and produce excellent results in the specific DR task. The embedding model helps to understand which parts of the graph are essential for a specific task and suggests which parts improve. Possible future works concern the extension of the KG with other types of entities and the complete automation of the analysis procedure.

REFERENCES

[1] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nature reviews Drug discovery*, 2019.

[2] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomedical Informatics Insights*, 2016.

[3] D. Han, S. Wang, C. Jiang, X. Jiang, H.-E. Kim, J. Sun, and L. Ohno-Machado, " Trends in biomedical informatics: automated topic analysis of JAMIA articles ," *Journal of the American Medical Informatics Association*, 2015.

[4] I. Tiddi, F. Lécué, and P. Hitzler, *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, ser. Studies on the Semantic Web. IOS Press, 2020.

[5] T. Gaudelet, B. Day, A. R. Jamasb, J. Soman, C. Regep, G. Liu, J. B. R. Hayter, R. Vickers, C. Roberts, J. Tang, D. Roblin, T. L. Blundell, M. M. Bronstein, and J. P. Taylor-King, "Utilising graph machine learning within drug discovery and development," 2021.

[6] T. U. Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research*, 2018.

[7] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "Comparative Toxicogenomics Database (CTD): update 2021," *Nucleic Acids Research*, 2020.

[8] D. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, 2006.

[9] V. McKusick, "Mendelian inheritance in man and its online version, omim," *American journal of human genetics*, 2007.

[10] L. Federer, Y.-L. Lu, D. Joubert, J. Welsh, and B. Brandys, "Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff," *PloS one*, 2015.

[11] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, 2016.

[12] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, 2020.

[13] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," 2020.

[14] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo, "Knowledge graph embedding for link prediction: A comparative analysis," *ACM Trans. Knowl. Discov. Data*, 2021.

[15] D. N. Nicholson and C. S. Greene, "Constructing knowledge graphs and their biomedical applications," *Computational and Structural Biotechnology Journal*, 2020.

[16] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," 2016.

[17] L. Costabello, S. Pai, C. L. Van, R. McGrath, N. McCarthy, and P. Tabacof, "Ampligraph: a library for representation learning on knowledge graphs," 2019.

[18] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Neural Information Processing Systems (NIPS)*, 2013.

[19] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019.