Drug Repurposing using Link Prediction on Knowledge Graphs

Otto Kißig [*1] **Martin Taraz** [*1] **Sarel Cohen** [1] **Tobias Friedrich** [1]

Abstract

The active global SARS-CoV-2 pandemic caused more than 167 million cases and 3.4 million deaths worldwide. As mentioned by Ye et al. (2021), the development of completely new drugs for such a novel disease is a challenging, time intensive process and despite researchers around the world working on this task, no effective treatments have been developed yet. This emphasizes the importance of *drug repurposing*, where treatments found among existing drugs meant for different diseases. A common approach to this is based on *knowledge graphs*, that condense relationships between entities like drugs, diseases and genes. Graph neural networks (GNNs) can then be used for the task at hand by predicting links in such knowledge graphs. Expanding on state-of-the-art GNN research, Doshi & Chepuri (2020) originally presented the model DR-COVID. We further extend their work using additional output interpretation strategies. The best aggregation strategy derives a top-100 ranking of candidate drugs, 32 of which currently being in COVID-19-related clinical trials. Moreover, we present an alternative application for the model, the generation of additional candidates based on a given pre-selection of drug candidates using collaborative filtering. In addition, we improved the implementation of the model by Doshi & Chepuri (2020) by significantly shortening the inference and pre-processing time by exploiting data-parallelism.

1. Introduction

With the novel coronavirus, a global pandemic with serious socio-economic implications for most parts of our daily lives is active (Nicola et al., 2020). The limited ability to take precautions for an unsuspected event like this and the rapid spread make finding an effective treatment as necessary as difficult, since the disease-specific knowledge is limited at the beginning and human lives are lost every day. Known and approved drugs happen to be well-studied, thus, they pose a good starting point for swift development of treatments, and an emerging tactic in fighting the pandemic (Shah et al., 2020). DrugBank, an extensive database compiling information about drugs approved by the US Food and Drug Administration as well as experimental drugs, contained more than 2 300 approved drugs and over 4 500 experimental drugs as of 2018; both with a strong upward trend (Wishart et al., 2018). This emphasizes the need for computer aided development of treatments.

Drug repurposing with knowledge graphs, as first described by Ashburn & Thor (2004), is the current state-of-the-art approach for finding possible treatments for novel diseases among known drugs using machine learning. Applying drug repurposing allows for a better way to maneuver through the pandemic. It can lead to better treatments for patients infected with one of the COVID-19 strains and a better understanding of the characteristics of the individual strains. Today, we approach the problem of drug repurposing using machine learning, focusing on deep learning methods. The idea of predicting unknown links between entities in a knowledge graph is traditionally known as *Collaborative Filtering*, as described by Sarwar et al. (2001). In this work we expand on the concept of *graph embeddings*, which map a fixed-size feature vectors to graph nodes and relations. A state-of-the-art technique for the creation of such embeddings based on deep neural networks (DNNs) is TRANSE (Bordes et al., 2013).

Knowledge graph embeddings are already utilized to solve different tasks related to drug discovery, e.g., they are used to predict potential drug targets for diseases to reduce cost and increase speed in the drug development process in general (Ye et al., 2021). Regarding the specific application of drug repurposing relying on edge prediction in a knowledge graph of biomedical data (see Section 2), Gysi et al. (2020) present a novel classification approach to this problem by implementing and merging various different ideas and techniques into one ensemble classifier. At its core, they deploy a DNN with an encoder-decoder structure. The encoder mechanism of it, which is based on the *Decagon* graph neural network by Zitnik et al. (2018), was initially proposed for the prediction of side effects of concurrent drug use.

[1]Hasso Plattner Institute, Potsdam. Correspondence to: Otto Kißig <otto.kissig@student.hpi.de>.

The 2021 ICML Workshop on Computational Biology.

Our Contribution. In this paper we extend the work of Doshi & Chepuri (2020) and Kißig et al. (2021). We offer three contributions to the deep learning and the bioinformatics community.

1. We improve the post prediction step of Doshi & Chepuri (2020) by using a clustering of similar diseases and increasing by more than 50% the number of predicted drugs in the top-100 that were or are in clinical trials.

2. We explore the additional application of finding drug candidates similar to a manually pre-selected candidate using collaborative filtering on the same model output. We show that many drugs that are in clinical trial can be found by detecting the drugs that are the most similar (*e.g.* using cosine-distance on the embedding of the drugs) to a given known drug (or a subset of drugs) which is or was in clinical trials.

3. We re-implement[1] the model described by Doshi & Chepuri (2020) and improve it by allowing flexible neighborhood capture sizes. We also improve the implementation by Kißig et al. (2021) by improving training speed, inference time, readability and by reducing pre-processing time from 30 minutes to 2 minutes by leveraging matrix operations. We further extend the implementation to support Self-Label-Enhancement.

2. Dataset

Our work relies on the Drug Repurposing Knowledge Graph (DRKG) by Ioannidis et al. (2020), which compiles data from different biomedical databases. It contains 97 238 entities belonging to 13 entity types and 5 874 261 triplets belonging to 107 edge types. We restrict ourselves to 98 edge types between 4 entity types, namely gene, compound, anatomy and disease, which leaves us with a knowledge graph with 69 036 entities and 4 885 854 edges. In particular, it contains drugs and substances as *compound* entities, as well as different COVID-19 variants as *disease* entities. The edge types include e.g. *compound-treats-disease* edges, which is the kind of edge our model predicts.

One part of DRKG are the precomputed TRANSE embeddings trained using `dgl-ke` by Zheng et al. (2020). To train our model to predict whether a given edge in some *compound-treats-disease* relation exists, we have to create suitable training data. To provide our model with both positive and negative samples for training, for each positive edge we sample 30 non-edges in the dataset, which results in a ratio similar to DR-COVID. This process tries to account for the imbalance of edges and non-edges in the ground truth. The set of edges included in the dataset is not complete, however, it is quite certain to be correct. Consequently, the positive edges are given a higher weight in the loss calculation, and the higher number of negative edges (which are not certain to be truly negative) are given a lower weight. To prevent too much imbalance in the individual minibatches, we use a weighted random batch sampler that over-samples the positive samples yielding an expected ratio of $1 : 1.5$ of positive to negative samples in each batch.

[1]Our implementation of the experiments and the model can be found here: https://drive.google.com/file/d/1j4RF7bKquz1W1i9ZY4TX0CWd9rz6NE0l/view?usp=sharing.

3. Model Architecture

The architecture of our model is illustrated in Figure 1. It consists of a SIGN (Frasca et al., 2020) architecture encoder, which provides an embedding $y \in \mathbb{R}^{250}$ for each node. We apply $tanh$ to the encoder output and forward it into our decoder. Given two nodes u, v, the decoder takes their encodings y_u, y_v and assigns a score $s_{u,v} \in [0, 1]$, which measures the probability for an edge between nodes u and v to exist. The decoder consists of two linear layers $\ell_1(u)$ and $\ell_2(v)$ that process the encodings y_u and y_v via a sigmoid function, that is, $\sigma(y_v \cdot \ell_1(y_u) + y_u \cdot \ell_2(y_v))$. The loss of the model is computed using a binary cross entropy loss with logits with weights set as described in Section 2.

Implementation. The dataset presents itself as a list of triples, each posing source, relation-type and sink of an edge. This is accompanied by precomputed knowledge graph embeddings. For the preprocessing we first filter out the edges belonging to the part of the knowledge graph we restrict ourselves to. We then construct a graph with the help of DGL (Wang et al., 2019). To compute the neighborhood embeddings we feed into the model, we first derive an adjacency matrix $A \in \{0, 1\}^{n \times n}$ from the reduced graph, from which the edges we try to predict, i.e., *compound-treats-disease* edges, have been removed. We then derive the normalized graph Laplacian $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ where $D_{i,i}$ is the degree of node i. Suppose $X \in \mathbb{R}^{n \times 400}$ is the matrix of graph embeddings for the n nodes, then the kth neighborhood is defined as $\tilde{A}^k X$.

4. Output Interpretation

In this section we present different strategies for interpreting the scores that the model outputs for the application of predicting the top-k most promising compound nodes for a given set of disease nodes D. Note that this is important as there are multiple COVID-19 diseases. Let n be the total amount of compound nodes. Predicting all $n \cdot |D|$ edge combinations, our model yields a matrix of scores $S \in \mathbb{R}^{|D| \times n}$. For each of the following strategies we first perform a standardization of the scores per disease using $\hat{s}_{dc} = \frac{s_{dc} - \mu(s_{d*})}{\sigma(s_{d*})}$, where d is the index of a disease in D, c

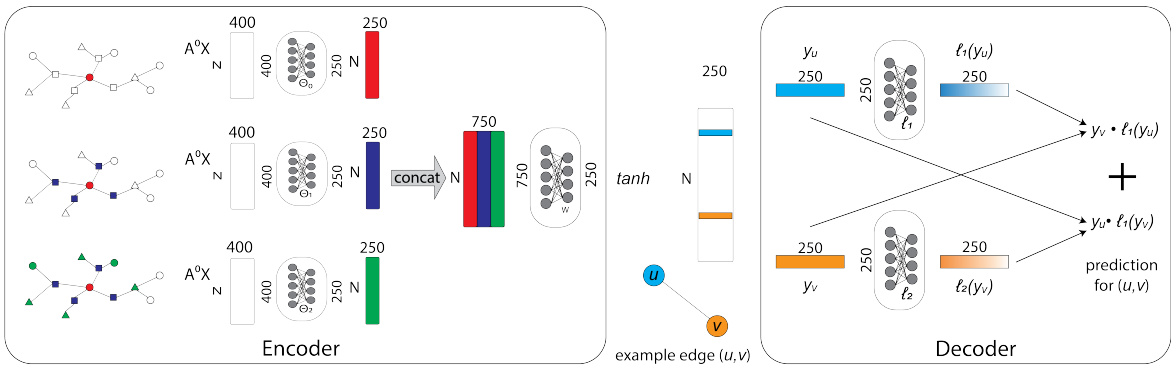

Figure 1. The architecture of our model as described in Section 3.

being the index of the compound, $\mu(s_{d*})$ and $\sigma(s_{d*})$ denote the mean and standard deviation over all diseases.

Certain "mild" diseases may be affected by plenty of compounds resulting in those being linked more likely. The standardization helps to achieve a better comparability across different diseases, allowing us to identify the suited compounds for every disease individually and compare those. However, this could also give good scores to some compounds in the case of diseases with no "good" scores in the first place, potentially yielding some less useful proposals.

An aggregation strategy takes our matrix of standardized scores $(\hat{s}_{dc})$ and derives a list of compounds from it, the top-k of which are our result. We propose the following aggregation strategies. For **global score mean**, we calculate the means of $(\hat{s}_{dc})$ along axis 0, that is, over all diseases per compound; then we sort the compounds by their respective scores and select the top-k. For **global score maximum**, we find the maxima of $(\hat{s}_{dc})$ along axis 0; then again we sort the compounds and select the top-k. For **union over disease rankings**, we calculate top-x compounds per disease with x as small as possible such that we get at least k unique compounds in the union. We then concatenate all those top-x lists together to get a top-k compound list.

We also propose **greedy max-min fairness**. Inspired by a game-theoretic approach from auction theory, where we think of the COVID-strains as players and the compounds as items from which we can only pick a small set, we try to heuristically find a set of compounds that will maximize the COVID strain whose total score is the minimum. Note that Global Score Mean can be considered as allocating the drugs to the COVID strains in a way that obtains the maximum social welfare. In contrast, in the Greedy Max Min Fairness we allocate the candidate drugs among the COVID-strains in a way favoring fairness over social welfare. More precisely, we rank the drugs by iteratively selecting the drug that benefits the disease with the lowest sum of scores over all already selected drugs. From this ordering we then pick the top-k drugs. Because our standardized model outputs $\hat{s}_{dc}$ can be negative, we normalize these by additively shifting them into the positive numbers. This bias however does not interfere with the resulting order because it increases uniformly on all parts of the sum.

Furthermore, in **cluster score maximum**, grouping similar disease types can be used to enhance the accuracy of our top-k predictions. We perform such a grouping using the k-means clustering algorithm. For each cluster, which now represents a group of similar diseases, we use a mean reduction to calculate the score of a compound and then reduce to the maximum across these clusters. A sensible number of clusters to create can be chosen by performing a principal component analysis (PCA) (Pearson, F.R.S., 1901) on the standardized scores. Lastly, for **union over cluster rankings**, we perform the top-x selection on clusters calculated with the clustering method described above. This not only allows us to use a greater x because we have fewer lists to pick from, but also to get more consistent top picks because of the internal averages that we apply inside each cluster.

5. Evaluation

To test our compound ranking methods, we apply each to retrieve a top-100 list of proposed candidates. We then compute the number of intersections with the compounds that are part of a clinical trial related to COVID-19 according to the U.S. National Library of Medicine (World Health Organization, 2021). For this we use a compiled Kaggle dataset (Pandey, 2021) containing compound names.

We implement the model using PyTorch (Paszke et al., 2019). We train it using the AdamW optimizer (Kingma & Ba, 2015). We use 90% of the data for training and the rest for validation. The training is performed on Google Colab utilizing a Nvidia Tesla T4 and it takes ∼2 minutes to prepare the graph dataset. We train our model using 25 epochs with a starting learning rate of 10^{-5} and a weight decay of 10^{-2}. Each training epoch took us 30 seconds, which is a significant improvement over the 610 seconds of the implementation by Doshi & Chepuri (2020) and can be

Aggregation strategy	# hits
Single Disease (median)	20
Global Score Maximum	22
Global Score Mean	30
Greedy Max Min Fairness	23
Cluster Score Maximum with KMeans(k=8)	18
Cluster Score Maximum with KMeans(k=3)	20
Union over Disease Rankings (DR-COVID, (Doshi & Chepuri, 2020))	21
Union over Cluster Rankings with KMeans(k=8)	24
Union over Cluster Rankings with KMeans(k=3) (Kißig et al., 2021)	**32**

Table 1. Hits of proposed candidates in actual clinical trials.

attributed to the exploitation of data parallelism we added.

We compare the top-100 results of predicting compounds for `SARS-CoV2 E` of our obtained model to those predicted utilizing the weights of (Doshi & Chepuri, 2020). While their model's top-100 predictions include 22 compounds showing up in clinical trials, we only reach 15. We suspect the hand-made adjustments to the dataset utilizing undisclosed data sources are responsible for this discrepancy, as this is the sole missing part in our reimplementation. Consequently, we used their published weights for the evaluation of the post classification methods presented in Section 4.

The results of the the different post classification procedures can be found in Table 1. We see that our Union over Cluster Rankings with KMeans(k=3) outperforms the other approaches, yielding 32 hits. A PCA on the prediction scores indicates that there are 3 clusters among the COVID strains, making a choice of $k = 3$ sensible. In contrast, DR-COVID's aggregation method, Union over Disease Rankings, reaches just 21 hits in our evaluation process. We observe that the hits are not evenly distributed along the rankings of the aggregation strategies, with more hits towards the places 60 and higher, allowing to weigh up prediction validity against the number of predictions.

6. Collaborative Filtering

Suppose we already have pre-selected some candidates for clinical trials. Now we would like to identify similar candidates that could be interesting. This new application can be approached using collaborative filtering on our model output. We measure the similarity[2] along the model's edge predictions per compound.

We test this application by ranking the remaining compounds of our dataset by the cosine similarity to pre-selected candidates. Our pre-selections are sampled randomly from the clinical trial dataset. In the case of one single pre-selected candidate, for selecting the top-100 drugs ranked by similarity to the pre-selected candidate we get a mean of 18 (min. 0, max. 32) hits. Conducting the experiment with 15 pre-selected candidates and selecting drugs corresponding to the top-100 of a global ranking of all similarities yields on average 18 (min. 0, max. 37) hits.

[2]To precisely define the cosine similarity between two given drugs i, j, let $\hat{s}_{*i}, \hat{s}_{*j}$ be their prediction scores along the disease dimension. Then their similarity is defined as $\hat{s}_{*i} \cdot \hat{s}_{*j}$.

7. Conclusion and Future Work

Deep learning can help the development of drugs in the face of a global pandemic. Rather than looking for promising candidates by hand, one relies on graph neural networks. These build on top of compiled knowledge graphs connecting chemical compounds, diseases and individual genes and can help with this task without actually understanding the semantic meaning of individual relationship types. Being mainly good at detecting similar nodes in a graph makes them useful across many fields and in contexts beyond bioinformatics. As already shown by Ioannidis et al. (2020), predicting candidates for COVID-19 treatments using deep learning is a promising technique. We have been able to clarify the evaluation part of DR-COVID by Doshi & Chepuri (2020) and proposed an aggregation technique yielding better results. Our own implementation improves both training speed as well as readability.

Regarding the collaborative filtering on the model output we find that an informed pre-selection of some drug candidates by an expert opens the ability of the model to derive other possible candidates based on the similarity of the predicted treatment-features to the specific disease. This poses an advanced strategy in comparison to just searching for drugs similar in general. In our case we have made the pre-selection at random. A pre-selection that deliberately selects different types of drugs, possibly targeting different aspects of the disease at hand, could yield a better result set.

During our experiments we restricted ourselves to the SIGN architecture (Frasca et al., 2020) for the encoder part, since we built on top of the work by Doshi & Chepuri (2020). The recently introduced new GNN architecture SAGN (Sun & Wu, 2021) proposes a Self-Label-Enhancement mechanism that can improve model performance. We have already built but not yet fully compared this to our other approaches.

It remains open work to measure the effects of supplying the model with varying neighborhood sizes, which our reimplementation specifically allows. Moreover, a thorough analysis of the types, stages and amount of clinical trials as well the role a drug plays for a study (e.g., main treatment, mitigation of side effects) remains to be conducted.

Acknowledgements. We would like to thank the anonymous reviewers for their helpful feedback.

References

Ashburn, T. T. and Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 2004.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NeurIPS)*, pp. 2787–2795, 2013.

Doshi, S. and Chepuri, S. P. Dr-COVID: Graph neural networks for SARS-CoV-2 drug repurposing. *CoRR*, 2020.

Frasca, F., Rossi, E., Eynard, D., Chamberlain, B., Bronstein, M., and Monti, F. SIGN: Scalable inception graph neural networks. *CoRR*, 2020.

Gysi, D. M., Valle, Í. D., Zitnik, M., Ameli, A., Gan, X., Varol, O., Sanchez, H., Baron, R. M., Ghiassian, D., Loscalzo, J., and Barabási, A. Network medicine framework for identifying drug repurposing opportunities for COVID-19. *CoRR*, 2020.

Ioannidis, V. N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., and Karypis, G. DRKG - drug repurposing knowledge graph for covid-19. https://github.com/gnn4dr/DRKG/, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kißig, O., Taraz, M., Cohen, S., Doskoč, V., and Friedrich, T. Drug repurposing for multiple covid strains using collaborative filtering. In *ICLR Workshop on Machine Learning for Preventing and Combating Pandemics (MLPCP@ICLR)*, 2021.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., and Agha, R. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery*, 2020.

Pandey, P. Covid19 clinical trials dataset. https://www.kaggle.com/parulpandey/covid19-clinical-trials-dataset, 2021. Retrieved February 19th, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pearson, F.R.S., K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901.

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. Item-based collaborative filtering recommendation algorithms. In *International Conference on World Wide Web*, 2001.

Shah, B., Modi, P., and Sagar, S. R. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. *Life Sciences*, 2020.

Sun, C. and Wu, G. Scalable and adaptive graph neural networks with self-label-enhanced training, 2021.

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., and Zhang, Z. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

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., and Wilson, M. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.*, 2018.

World Health Organization. International clinical trials registry platform (ictrp). https://www.who.int/clinical-trials-registry-platform, 2021. Online; accessed February 19th, 2021.

Ye, C., Swiers, R., Bonner, S., and Barrett, I. Predicting potential drug targets using tensor factorisation and knowledge graph embeddings, 2021.

Zheng, D., Song, X., Ma, C., Tan, Z., Ye, Z., Dong, J., Xiong, H., Zhang, Z., and Karypis, G. DGL-KE: training knowledge graph embeddings at scale. In *SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinform.*, 2018.