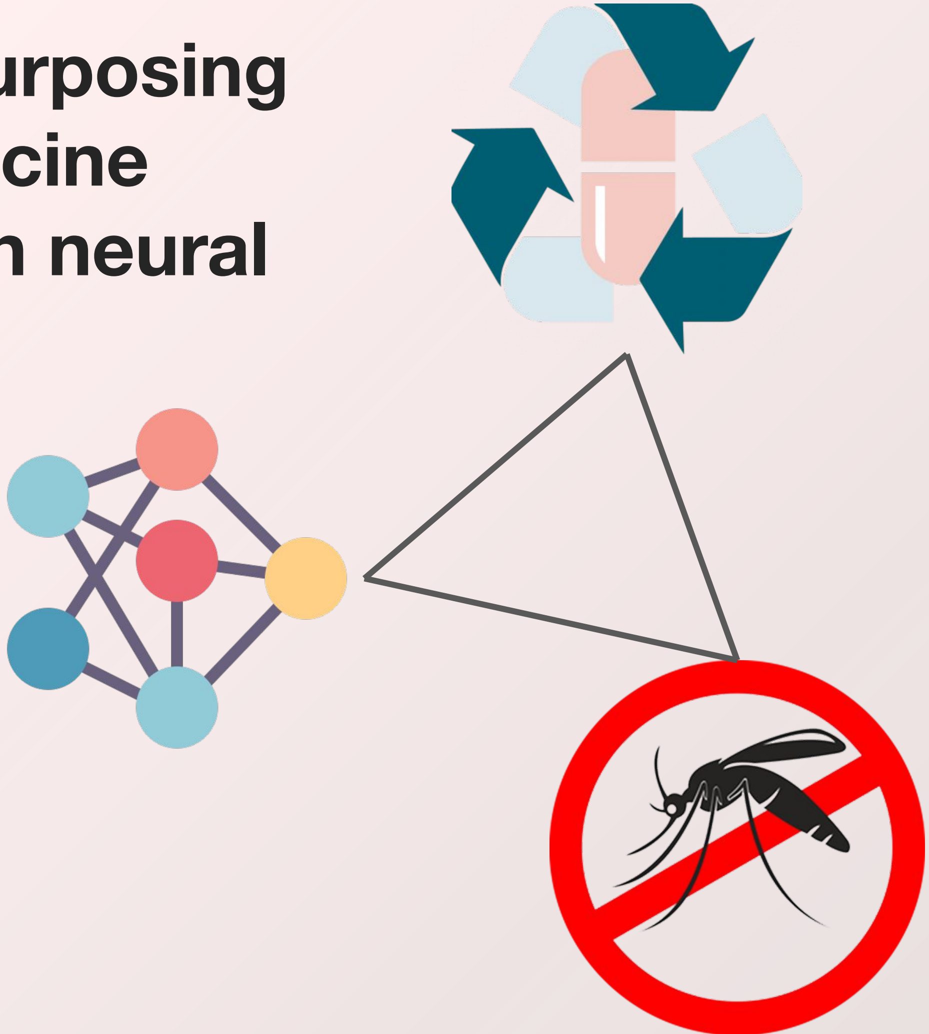
DengueDrugRep: Drug repurposing for dengue using a biomedicine knowledge graph and graph neural networks



Sebastián Ayala Ruano

Scientific Programming project

27th October 2023

Outline

Introduction

Methods

Results

Conclusions

Conference paper to get inspired

Conference on Cloud Computing, Big Data & Emerging Topics
↳ JCC-BD&ET 2023: **Cloud Computing, Big Data & Emerging Topics** pp 105–117 | Cite as

Home › Cloud Computing, Big Data & Emerging Topics › Conference paper

Drug Repurposing Using Knowledge Graph Embeddings with a Focus on Vector-Borne Diseases: A Model Comparison

Diego López Yse & Diego Torres

Conference paper | First Online: 11 August 2023

Part of the Communications in Computer and Information Science book series (CCIS,volume 1828)

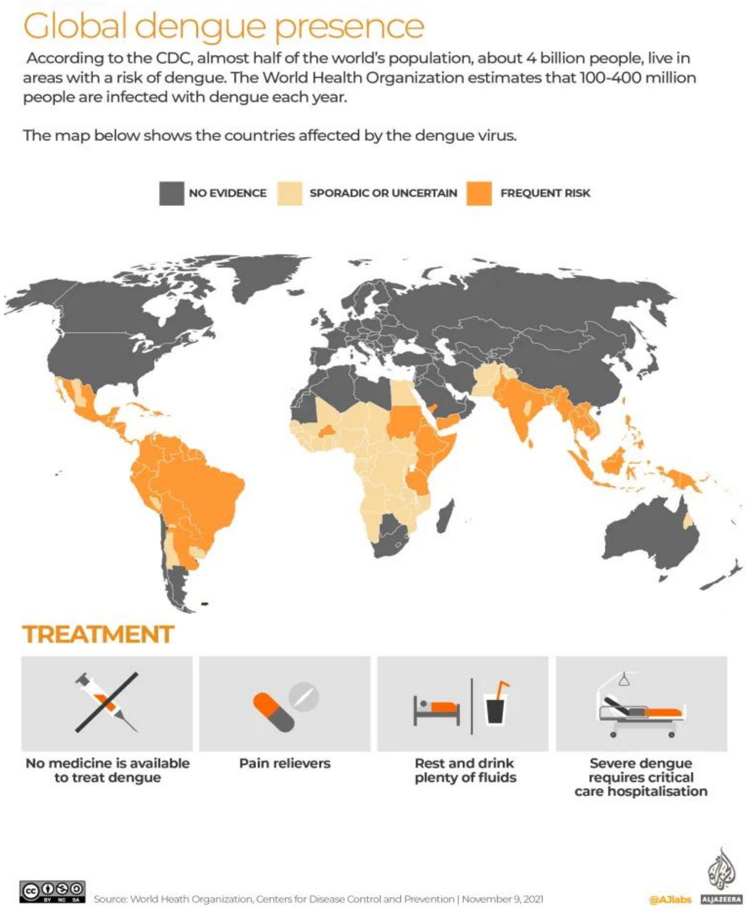
Abstract

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.

https://link.springer.com/chapter/10.1007/978-3-031-40942-4_8

Dengue disease

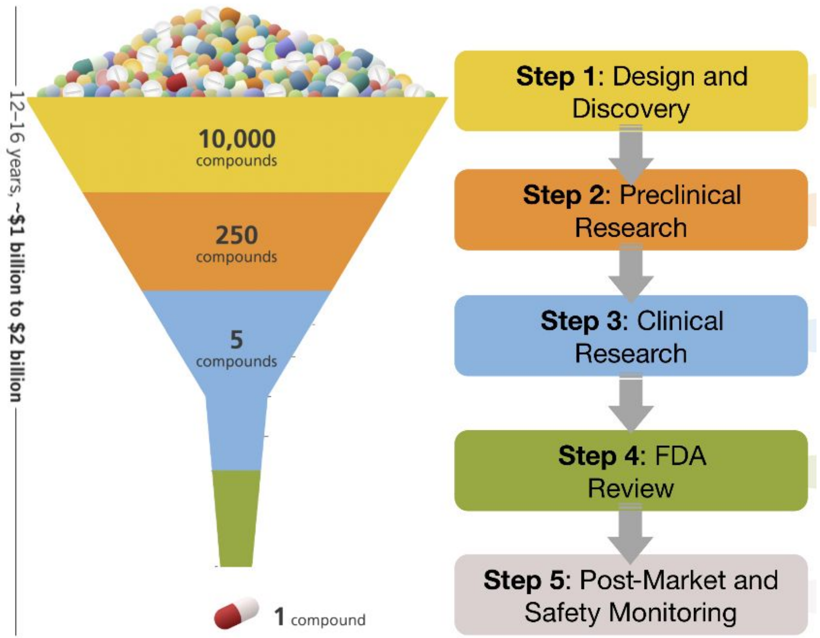
- Viral infection transmitted by the Aedes mosquitoes
- Neglected tropical disease
- Estimated 390 million dengue virus infections/year
- Fastest spreading, epidemic-prone infectious disease
- Severe dengue: life-threatening complications



Drug repurposing

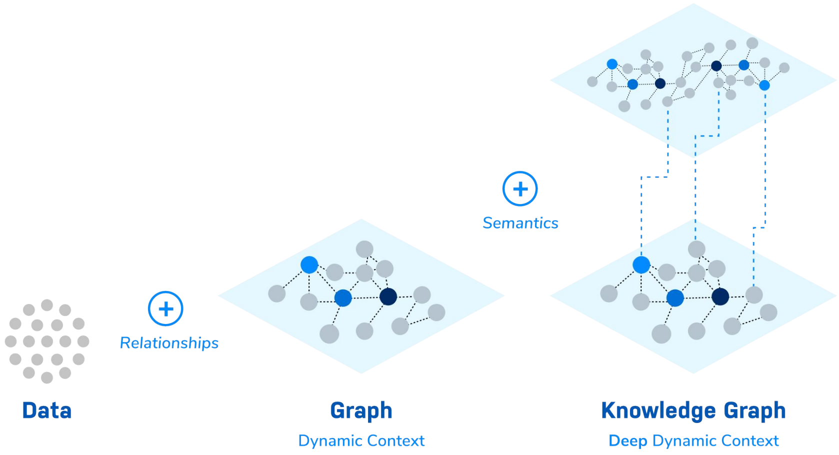
Identifying new uses for pre-approved drugs

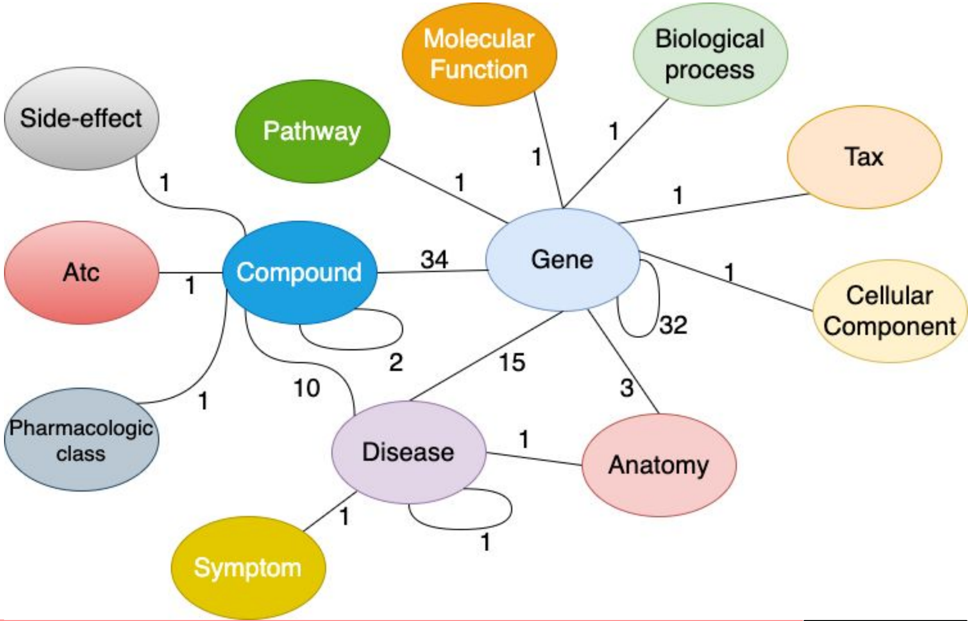
- Treatment of new diseases
- Novel druggable targets



Dataset - Drug Repurposing Knowledge Graph (DRKG)

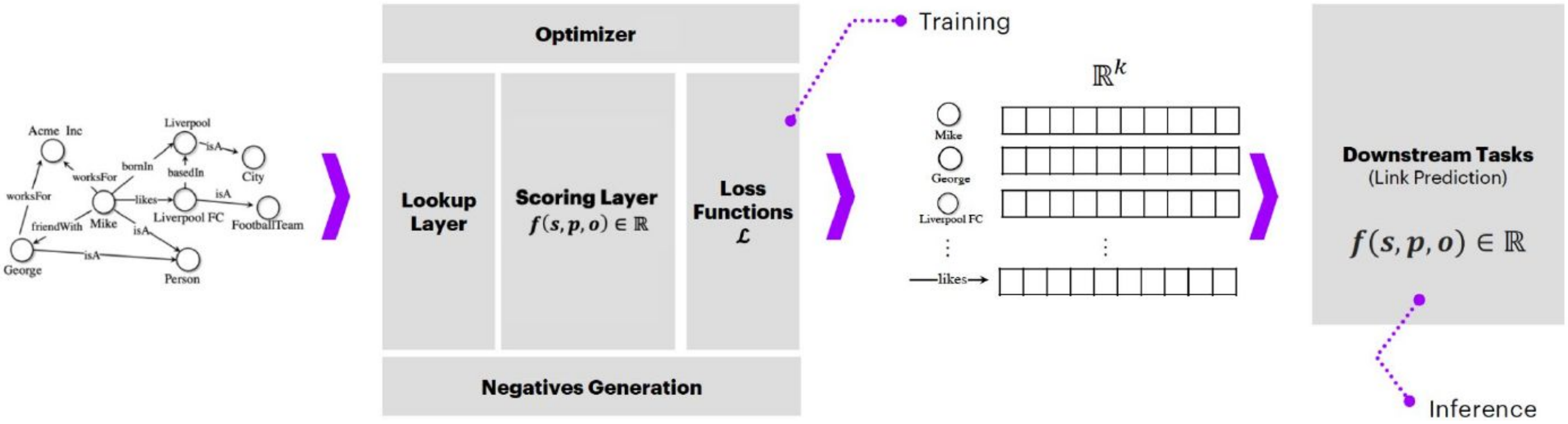
- **Knowledge graph**: heterogeneous network with different types of nodes and edges
- Includes information from **DrugBank**, **Hetionet**, **GNBR**, **String**, **IntAct** and **DGIdb**
- 97,238 entities (13 types) and 5,874,261 triplets (107 edge-types)
- 24,306 compounds from 17 databases



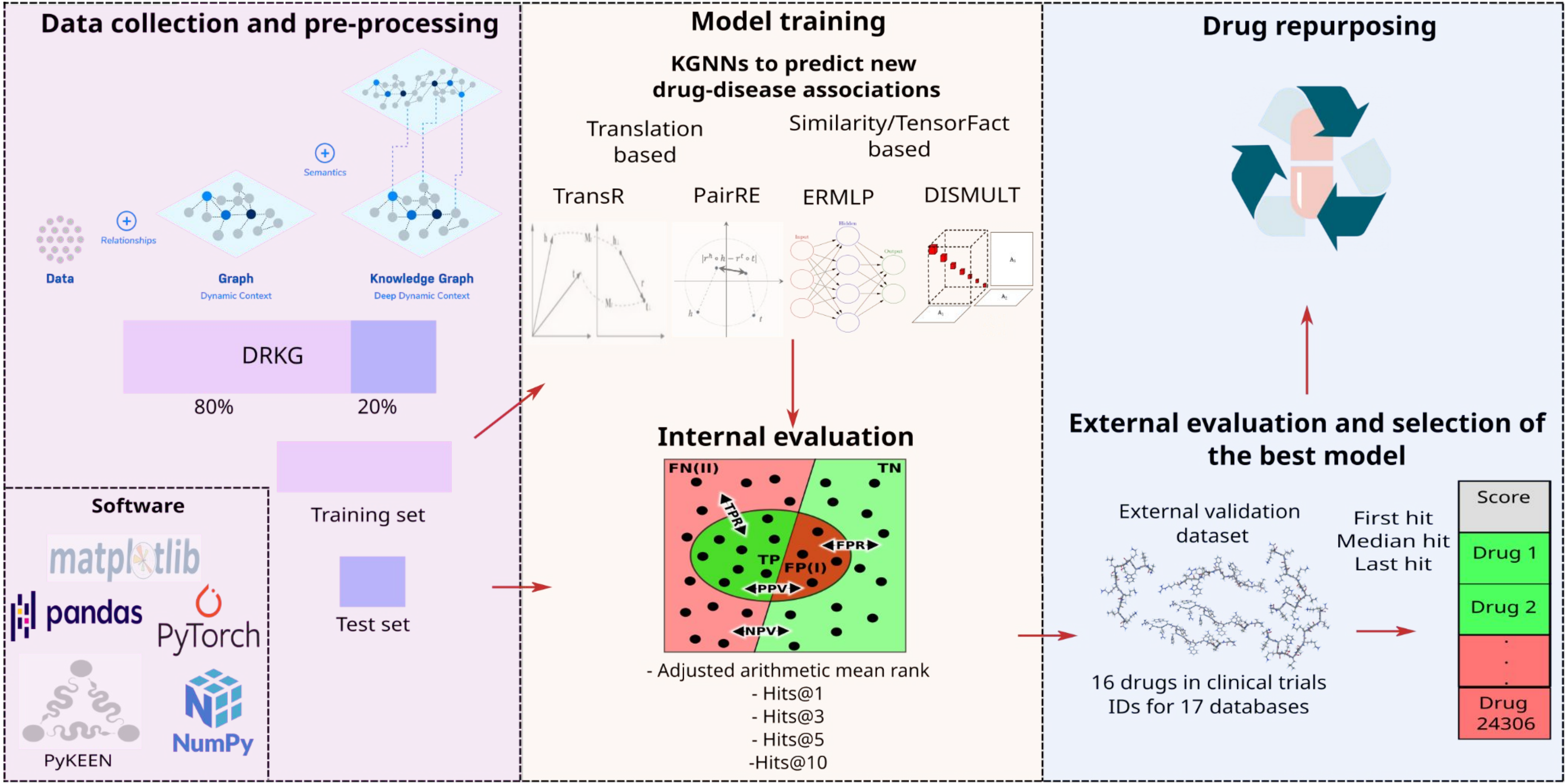


Models - Knowledge Graph Neural Networks (KGNNs)

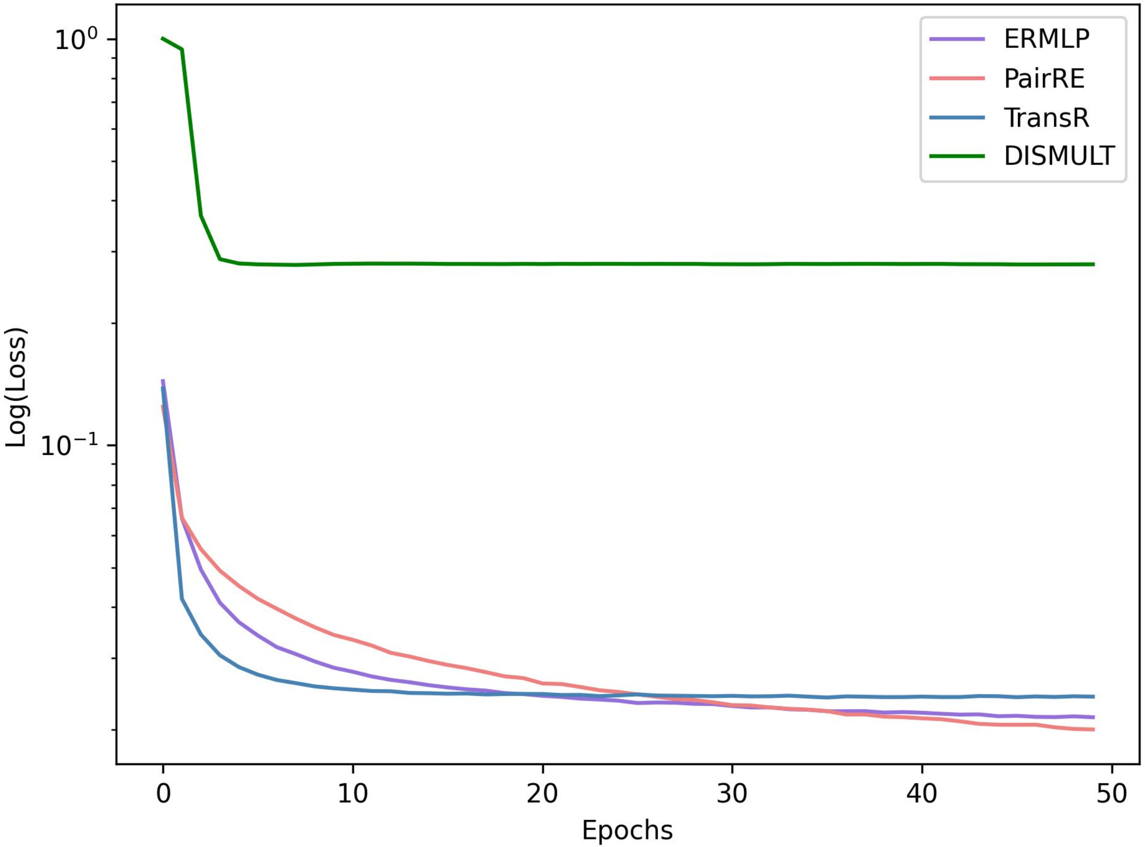
- **KGNNs**: type of neural networks that can learn from graph data
- **Graph embeddings**: function that maps nodes into a low dimensional vector
- Drug repurposing problem can be formulated as a **link prediction task** in a KG
- **Goal**: predict new **drug-disease associations**



Methods workflow



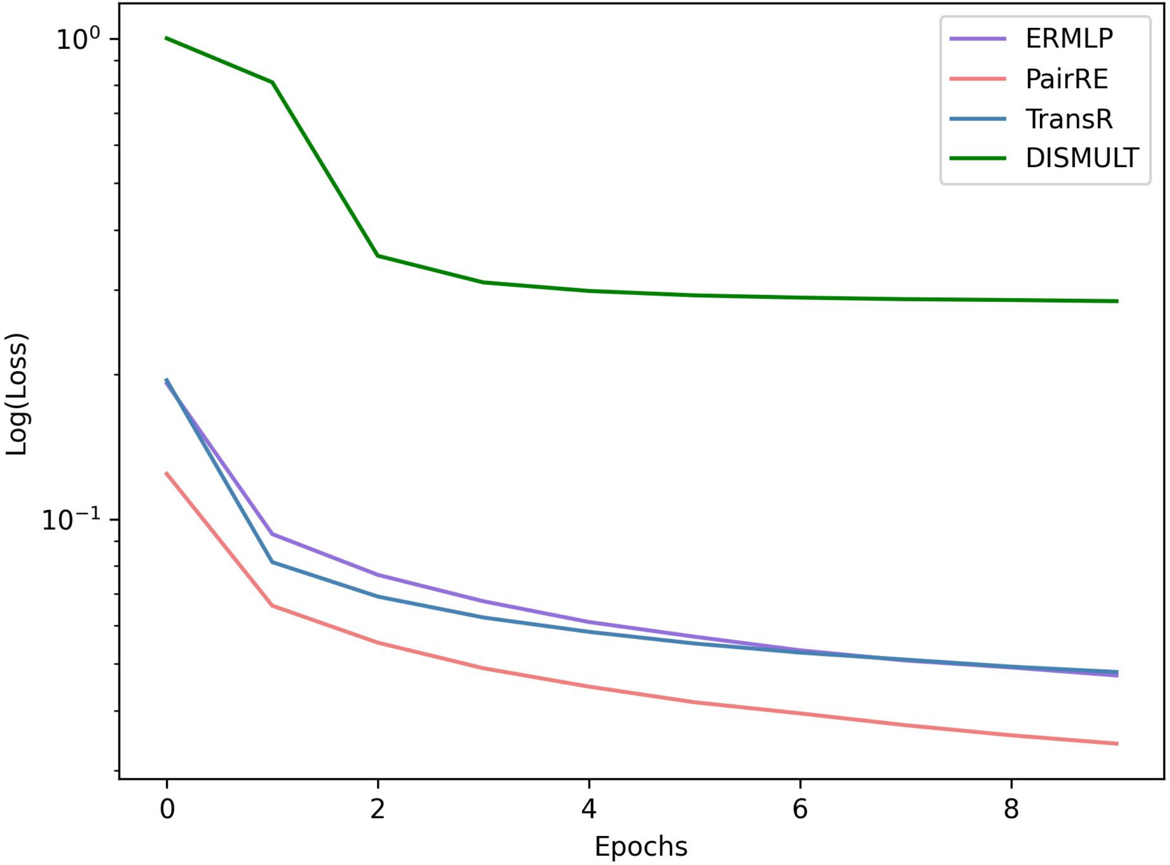
Internal evaluation results (all triplets)



Internal evaluation results (all triplets)

Model	Adjusted arithmetic mean rank	Hits@1	Hits@3	Hits@5	Hits@10
PairRE	0.0179	0.0211	0.114	0.16	0.221
ERMLP	0.0188	0.0276	0.079	0.110	0.164
TransR	0.0193	0.0121	0.064	0.088	0.132
DISMULT	0.040	0.016	0.036	0.050	0.076

Internal evaluation results (drug repurposing triplets)



Internal evaluation results (drug repurposing triplets)

Model	Adjusted arithmetic mean rank	Hits@1	Hits@3	Hits@5	Hits@10
DISMULT	0.0293	0.00781	0.0188	0.0270	0.0424
ERMLP	0.03107	0.00857	0.02091	0.02866	0.04616
PairRE	0.03337	0.00409	0.01190	0.0178	0.03149
TransR	0.0392	0.001260	0.00422	0.00731	0.01398

External evaluation results

All triplets

Model	First hit	Median hit	Last hit
ERMLP	26	1432	11354
DistMult	31	1407	10689
PairE	219	1397	17149
TransR	440	2297	16402

Drug repurposing triplets

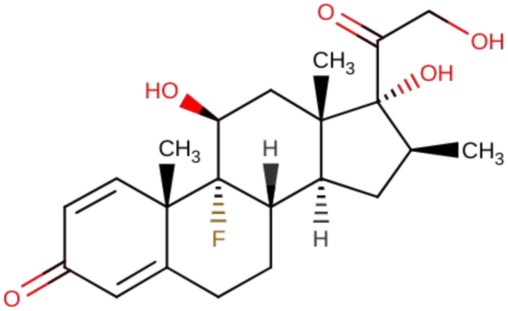
Model	First hit	Median hit	Last hit
ERMLP	9	2849	15369
DistMult	17	2655	18005
PairE	82	1352	5619
TransR	323	4321	18393

Prediction examples with the ERMLP model

1st: Betamethasone

Long-acting corticosteroid with immunosuppressive and anti-inflammatory properties

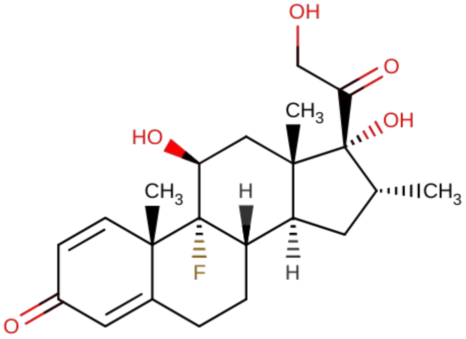
Diseases: dermatologic disorders, gastrointestinal diseases, and hematological disorders



2nd: Dexamethasone

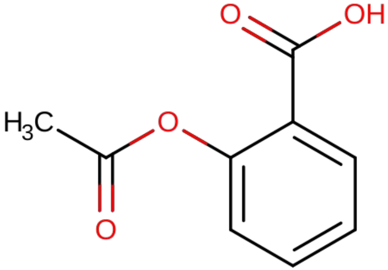
Glucocorticoid that is used for the treatment of various inflammatory conditions

Diseases: bronchial asthma, as well as endocrine and rheumatic disorders.



3rd: Aspirin

Salicylate used to treat pain, fever, inflammation, migraines, and reducing the risk of major adverse cardiovascular events



Conclusions and next steps

- Main goal accomplished - familiarize with KGs, KGNNs, PyKEEN, and how to apply these concepts and tools for drug repurposing

- Active and recent research field - lack of unified standards

- The ERMLP model had the highest performance based on the internal and external performance metrics - confirmed findings of conferance paper

- Potential next steps
    - Creation of a web app with the best model - accessibility
    - Increase training epochs for models evaluated with drug repurposing triplets
    - Try hyperparameter optimization and graph filtering methods
    - Extend list of ground-truth compounds (more clinical trials/databases)

Do you want to know more?

Slides: https://doi.org/10.5281/zenodo.10045649

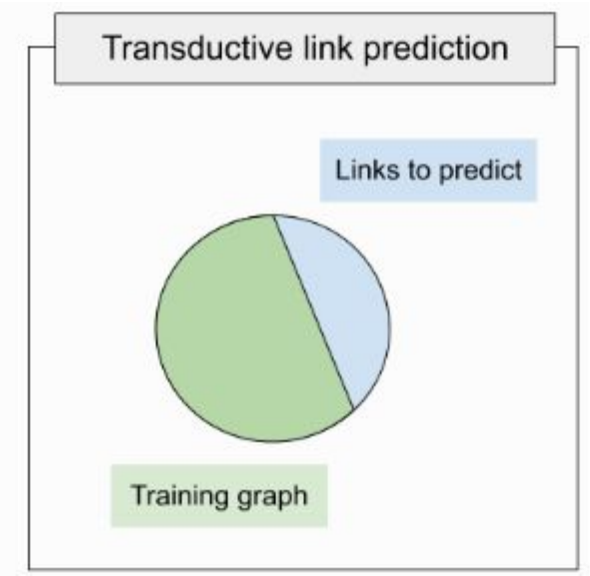
Code: https://github.com/sayalaruano/DengueDrugRep

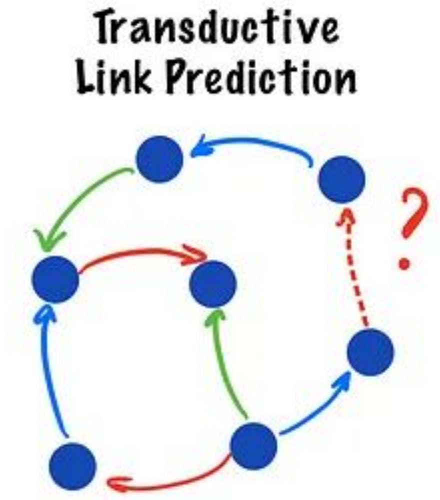
Pre-trained models: https://zenodo.org/doi/10.5281/zenodo.10010151

Extra slides

Understanding the training of KGGNs better

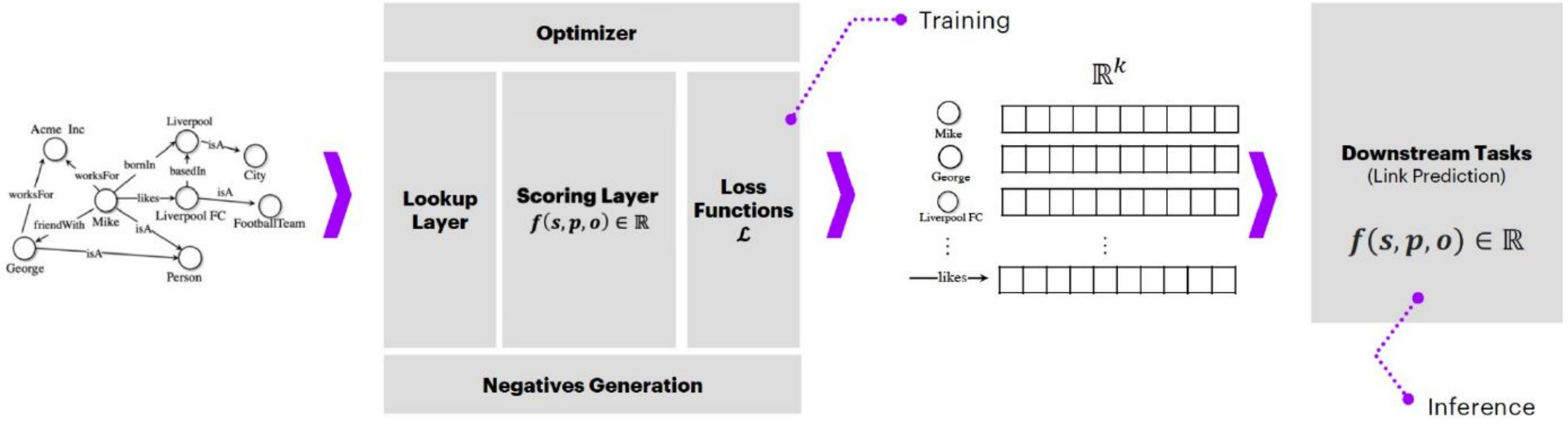
- Training graph includes all entities for inference (validation, test, or custom predictions).
- Missing links to be predicted connect already seen entities within the train graph





Understanding evaluation of KGGNs

- Evaluation set of triples $\mathcal{T}_{eval} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ and each triple in this set $(h, r, t) \in \mathcal{T}_{eval}$
  - Right-side prediction task: a pair of head entity and relation are given and aim to predict the tail, i.e.(h, r, ?). Score each of the possible choices (h, r, e) for e∈E.
  - Left-side prediction task: a pair of tail entity and relation are provided and aim to predict the head, i.e.(?, r, t). Score each of the possible choices (e, r, t) for e∈E.



Understanding evaluation of KGGNs

- Internal evaluation
    - Adjusted Mean Rank: the ratio of the Mean Rank to the Expected Mean Rank, assessing a model's performance independently of the underlying set size.
    - Hits@k: the fraction of times when the correct or "true" entity appears under the top-k entities in the ranked list.
- External evaluation: analyze the predicted ranked compound list against the actual treatment drugs defined in ground truth
    - First hit
    - Median hit
    - Last hit

Understanding evaluation of KGGNs

On the Ambiguity of Rank-Based Evaluation of Entity Alignment or Link Prediction Methods

Max Berrendorf[1 [0000-0001-9724-4009]], Evgeniy Faerman[1], Laurent Vermue[2], and Volker Tresp[1,3]

[1] Ludwig-Maximilians-Universität München, Munich, Germany
{berrendorf,faerman}@dbs.ifi.lmu.de
[2] Technical University of Denmark, Kongens Lyngby, Denmark
lauve@dtu.dk
[3] Siemens AG, Munich, Germany
volker.tresp@siemens.com

Abstract. In this work, we take a closer look at the evaluation of two families of methods for enriching information from knowledge graphs: Link Prediction and Entity Alignment. In the current experimental setting, multiple different scores are employed to assess different aspects of model performance. We analyze the informativeness of these evaluation measures and identify several shortcomings. In particular, we demonstrate that all existing scores can hardly be used to compare results across different datasets. Moreover, we demonstrate that varying size of the test size automatically has impact on the performance of the same model based on commonly used metrics for the Entity Alignment task. We show that this leads to various problems in the interpretation of results, which may support misleading conclusions. Therefore, we propose adjustments to the evaluation and demonstrate empirically how this supports a fair, comparable, and interpretable assessment of model performance. Our code is available at https://github.com/mberr/rank-based-evaluation.

A Unified Framework for Rank-based Evaluation Metrics for Link Prediction in Knowledge Graphs

Charles Tapley Hoyt*
Harvard Medical School
Boston, MA, USA
cthoyt@gmail.com

Max Berrendorf*
LMU Munich
München, Germany
max.berrendorf@gmail.com

Mikhail Galkin
Mila & McGill University
Montreal, QC, Canada
mikhail.galkin@mila.quebec

Volker Tresp
LMU Munich & Siemens AG
München, Germany
volker.tresp@siemens.com

Benjamin M. Gyori
Harvard Medical School
Boston, MA, USA
benjamin_gyori@hms.harvard.edu