BioBLP: A modular framework for representation learning over biomedical knowledge graphs

Journal of Biomedical Semantics (2023)

Michael Cochez
Email: m.cochez@vu.nl

Modified from slides by Dimitrios Alivanistos

Team

Vrije Universiteit amsterdam

University of Amsterdam

Elsevier

Dimitrios Alivanistos
d.alivanistos@vu.nl

Payal Mitra
p.mitra@elsevier.com

Paul Groth
p.groth@uva.nl

DiscoveryLab

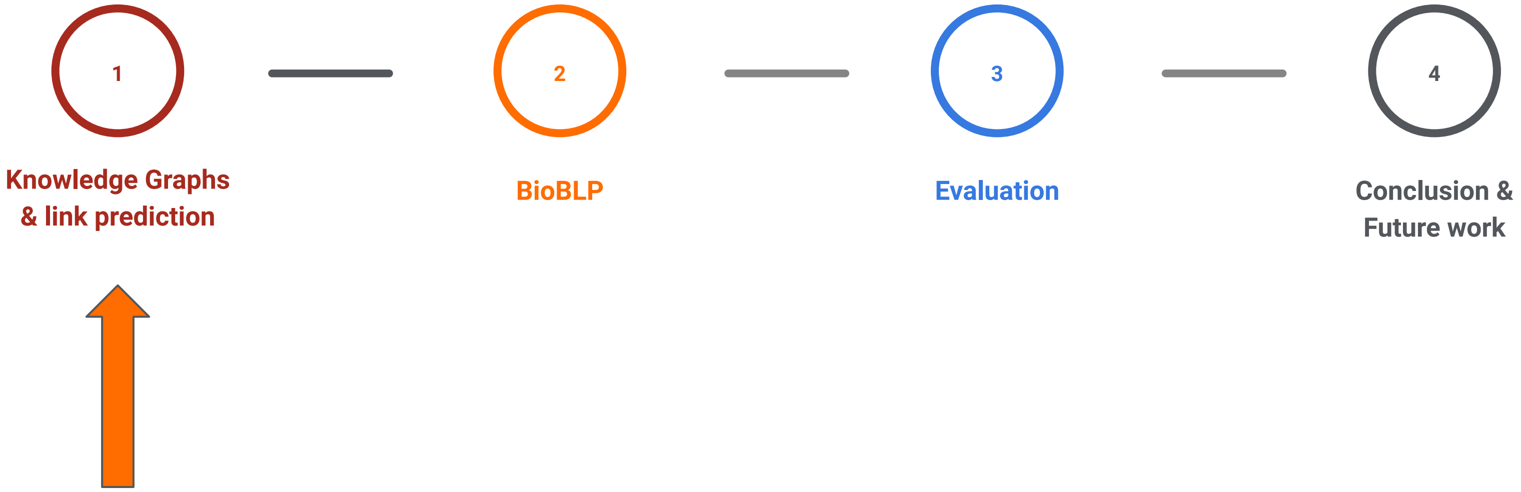
Michael Cochez
m.cochez@vu.nl

Daniel Daza
d.dazacruz@vu.nl

Thom Pijnenburg
t.pijnenburg@elsevier.com

GRAPH MASSIVIZER

Today



1 Knowledge Graphs & link prediction

2 BioBLP

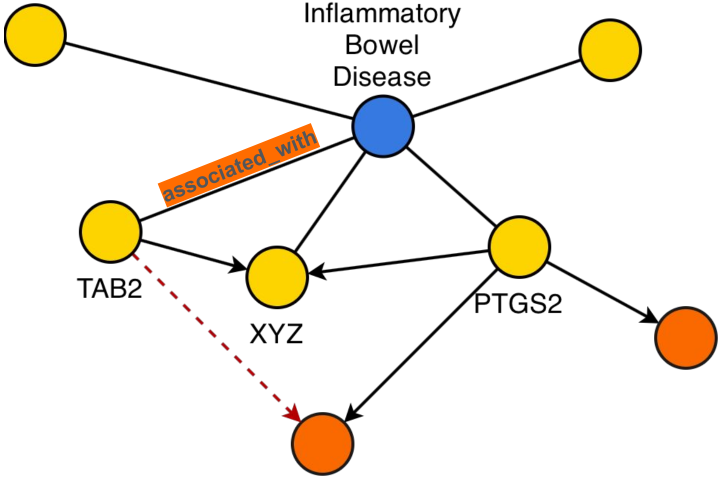
3 Evaluation

4 Conclusion & Future work

Knowledge Graphs

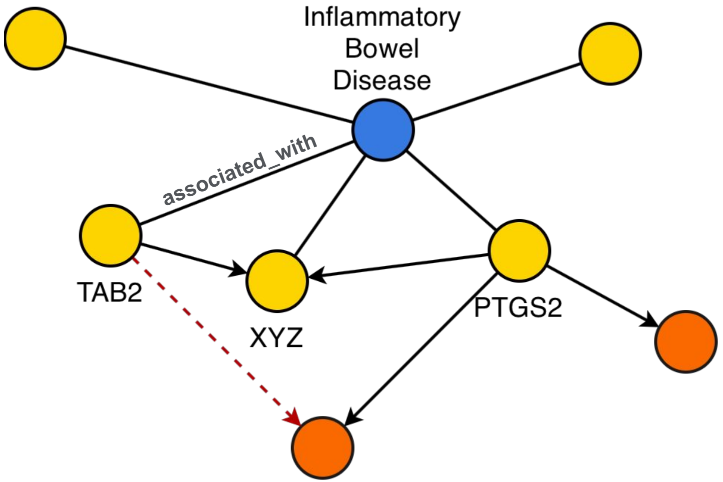
Data structures that:

- Model relational knowledge in the form of triples `<subject, predicate, object>` e.g. **`<TAB2, associated_with, IBD>`**
- Widely adopted in industry + academia
- Excellent fit for biomedical data.
    - Is it?

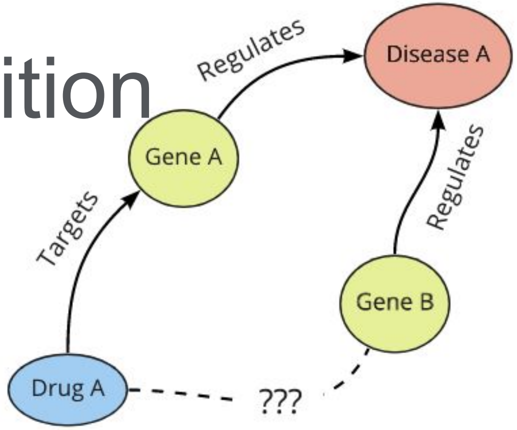


KG Characteristics

- They can contain errors…
- They can contain conflicting information
- Inherent incompleteness:
  - Knowledge evolves over time.
  - Knowledge becomes deprecated.
  - Continuous data integration and updating is tough & expensive.



Link prediction methods - intuition



- **Estimate the likelihood of links -** with machine learning
- Learn from the links which exist in in the graph
    - And from those that do not

- Goal: find a **scoring function** for edges

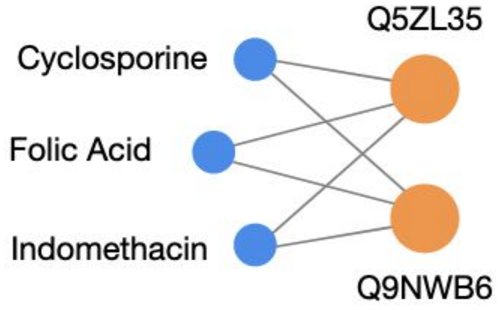
$$score(DrugA, Targets, GeneB)$$

- The arguments to this function are **embeddings**, learned vectors representing nodes and relations

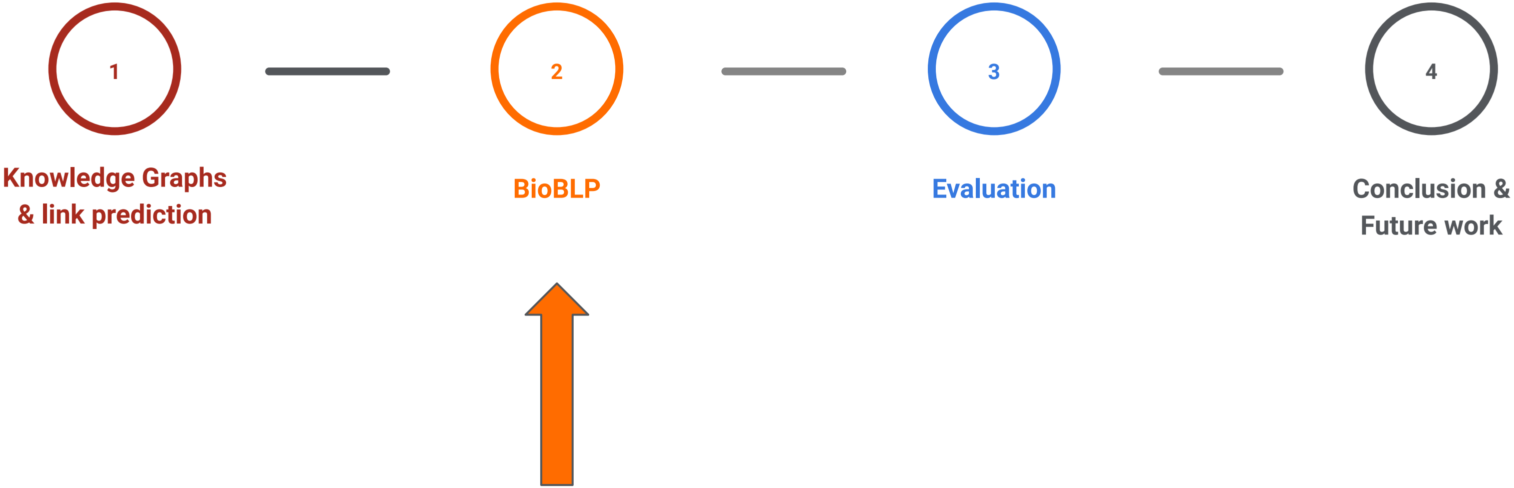
Shortcomings of Link Prediction

What if 2 entities are connected with exactly the same neighbours?

- Topologically indistinguishable…
- TransE, ComplEx, RotatE are unable to deal with this.
- Can we do more?



Today



1

Knowledge Graphs & link prediction

2

BioBLP

3

Evaluation

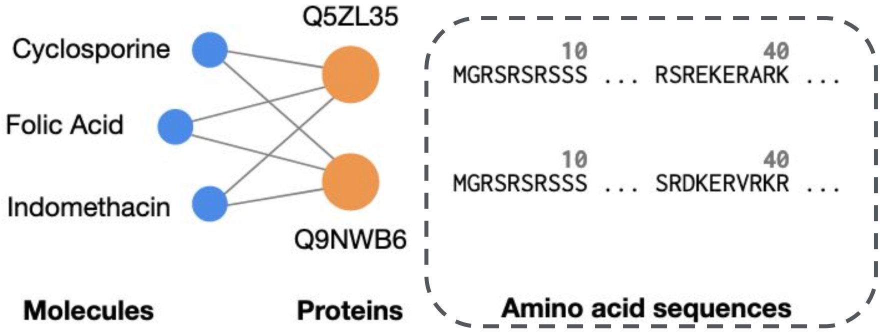
4

Conclusion & Future work

Our Goal

So… traditional LP models:

- **Only** consider the graph topology.
- Are **unable** to predict links for previously unseen entities.

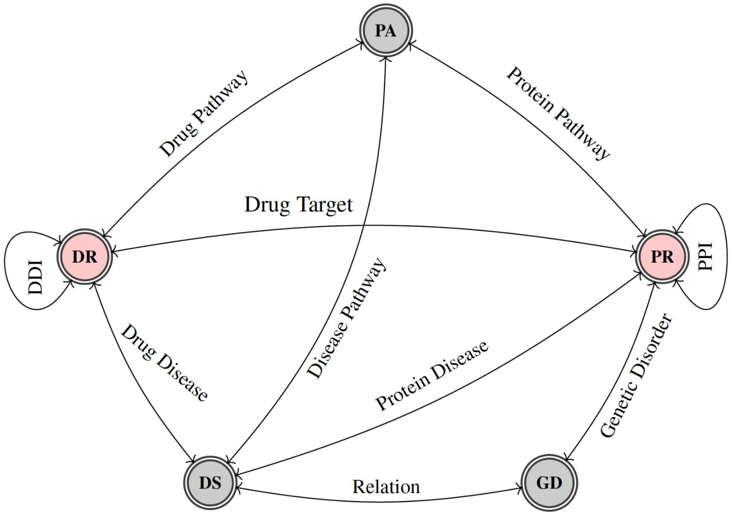


Our goal:

- Ameliorate the incompleteness problem by designing LP models that incorporate entity attributes.

The Data

- **BioKG*:** A KG for relational learning in bio data.
  - Curated data from **13 different** biomedical **DB** (MeSH, UniProt, DrugBank etc)
  - ~ 2 million triples.
    - 106,337 nodes (5 types)
    - 2,074,346 relationships (17 types).
  - Includes well-known **benchmarks.**
    - DPI-FDA, DDI, PPI

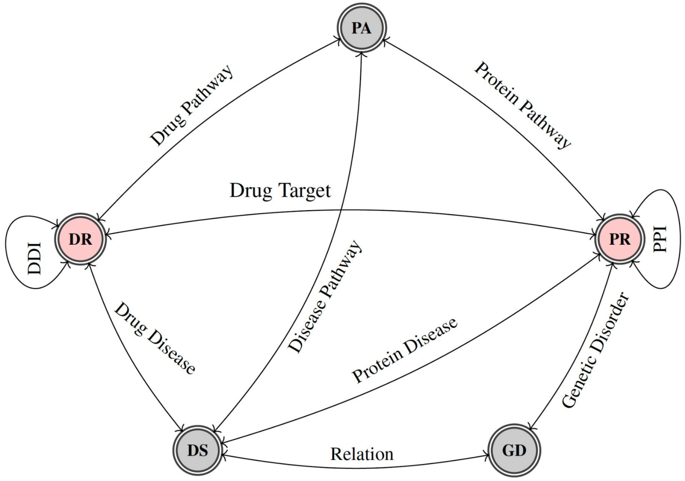


BioKG Schema. Image by Walsh et al.

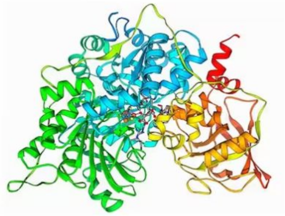
* Walsh, Brian, Sameh K. Mohamed, and Vít Nováček. "Biokg: A knowledge graph for relational learning on biological data." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020.

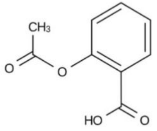
The Attributes

- **BioKG** includes entity *identifiers.*
  - Allows for attribute retrieval and incorporation from external sources.
    - Amino-acid sequences (UniProt)
    - Molecular structures - SMILES (Drugbank)
    - Disease textual descriptors (MeSH)

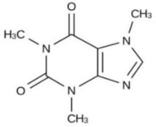


BioKG Schema. Image by Walsh et al.

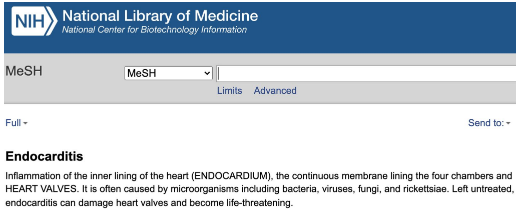




Acetylsalicylic Acid (Aspirin)
CC(=O)Oc1ccccc1C(=O)O



Epinephrine
CNC[C@H](O)c1ccc(O)c(O)c1

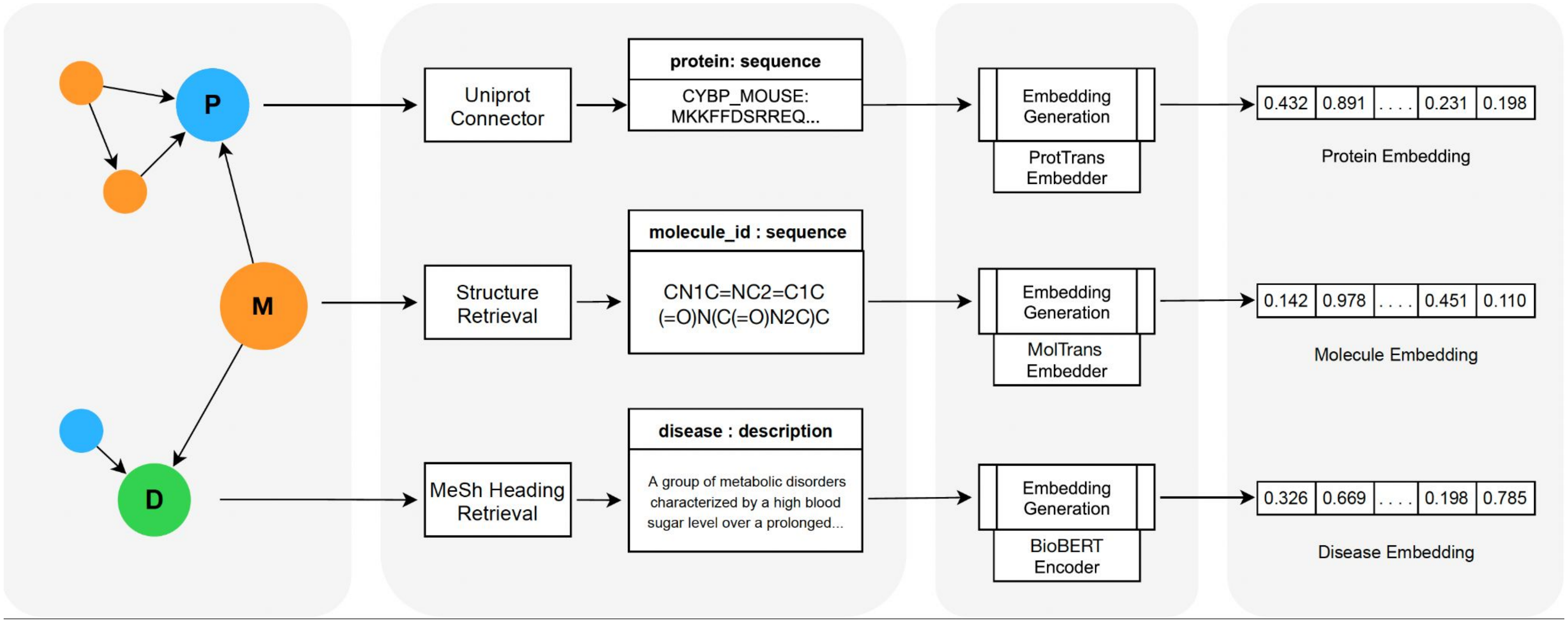


NIH
National Library of Medicine
National Center for Biotechnology Information
MeSH
MeSH
Limits
Advanced
Full
Send to:

Endocarditis

Inflammation of the inner lining of the heart (ENDOCARDIUM), the continuous membrane lining the four chambers and HEART VALVES. It is often caused by microorganisms including bacteria, viruses, fungi, and rickettsiae. Left untreated, endocarditis can damage heart valves and become life-threatening.

Attribute Encoders



Uniprot Connector

protein: sequence

CYBP_MOUSE: MKKFFDSRREQ...

Embedding Generation

ProtTrans Embedder

0.432	0.891	. . . .	0.231	0.198

Protein Embedding

Structure Retrieval

molecule_id : sequence

CN1C=NC2=C1C (=O)N(C(=O)N2C)C

Embedding Generation

MolTrans Embedder

0.142	0.978	. . . .	0.451	0.110

Molecule Embedding

MeSh Heading Retrieval

disease : description

A group of metabolic disorders characterized by a high blood sugar level over a prolonged...

Embedding Generation

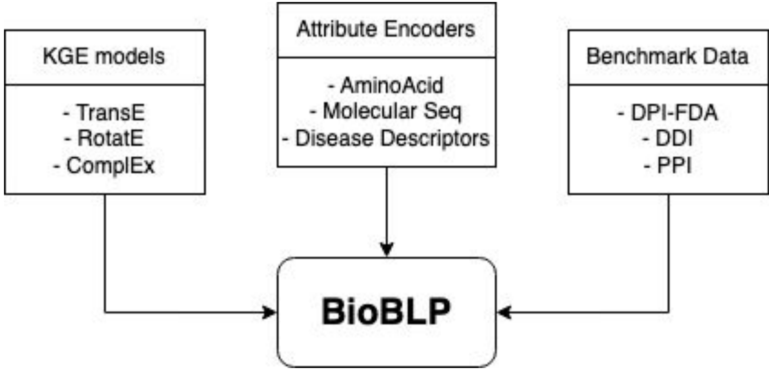
BioBERT Encoder

0.326	0.669	. . . .	0.198	0.785

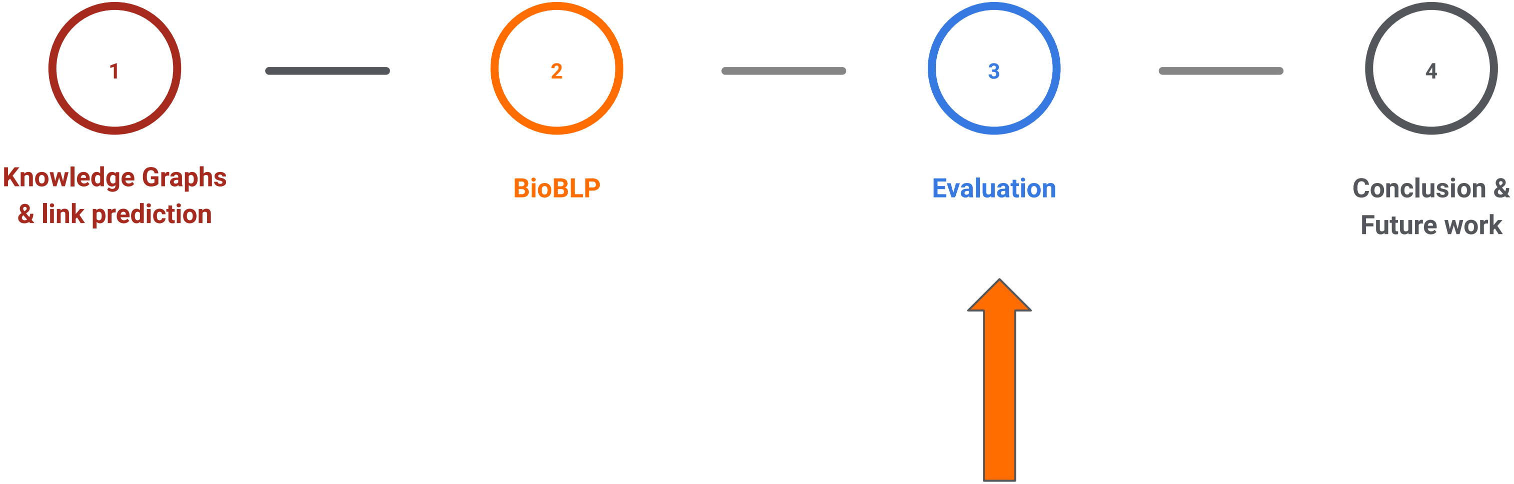
Disease Embedding

Bringing it all together: BioBLP

- A modular framework for learning from multimodal biomedical KGs. Allows for:
  - Training of out-of-the-box KGE models (ComplEx, RotatE, TransE)
  - Design attribute-specific "pluggable" encoders:
    - This work: *ProtTrans, MolTrans, BioBERT*
  - Efficient model (pre)training.



Today



1 Knowledge Graphs & link prediction

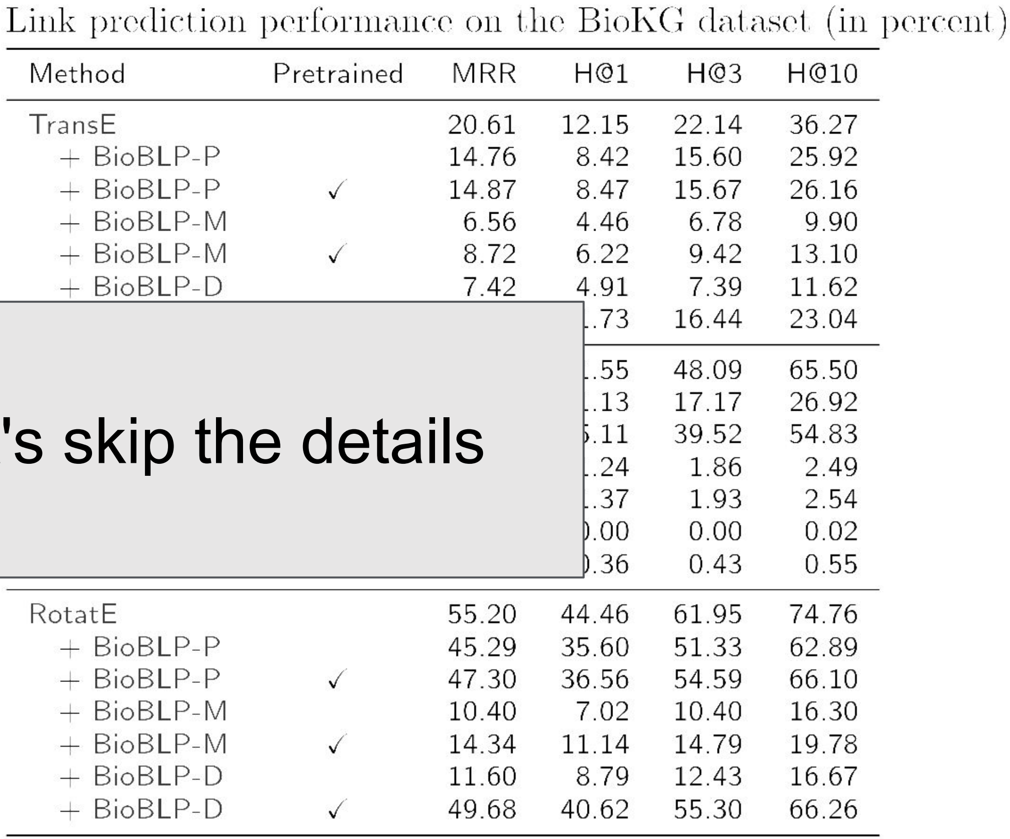
2 BioBLP

3 Evaluation

4 Conclusion & Future work

Results

- Not the performance increase we hoped
- Scoring function matters
  - TransE is nearly always
  - ComplEx cannot work w molecules and dis
  - RotatE usually the best
- Pre-training helps



Link prediction performance on the BioKG dataset (in percent)

Method	Pretrained	MRR	H@1	H@3	H@10
TransE		20.61	12.15	22.14	36.27
+ BioBLP-P		14.76	8.42	15.60	25.92
+ BioBLP-P	✓	14.87	8.47	15.67	26.16
+ BioBLP-M		6.56	4.46	6.78	9.90
+ BioBLP-M	✓	8.72	6.22	9.42	13.10
+ BioBLP-D		7.42	4.91	7.39	11.62
[illegible]	[illegible]	[illegible]	.73	16.44	23.04
[illegible]	[illegible]	[illegible]	.55	48.09	65.50
[illegible]	[illegible]	[illegible]	.13	17.17	26.92
[illegible]	[illegible]	[illegible]	.11	39.52	54.83
[illegible]	[illegible]	[illegible]	.24	1.86	2.49
[illegible]	[illegible]	[illegible]	.37	1.93	2.54
[illegible]	[illegible]	[illegible]	).00	0.00	0.02
[illegible]	[illegible]	[illegible]	).36	0.43	0.55
RotatE		55.20	44.46	61.95	74.76
+ BioBLP-P		45.29	35.60	51.33	62.89
+ BioBLP-P	✓	47.30	36.56	54.59	66.10
+ BioBLP-M		10.40	7.02	10.40	16.30
+ BioBLP-M	✓	14.34	11.14	14.79	19.78
+ BioBLP-D		11.60	8.79	12.43	16.67
+ BioBLP-D	✓	49.68	40.62	55.30	66.26

Let's skip the details

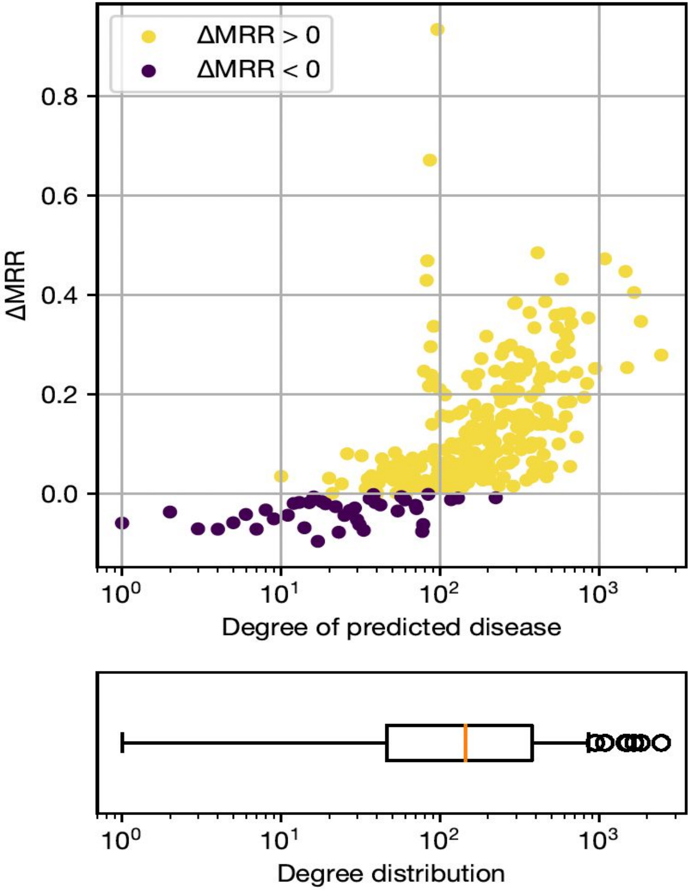
Observations

- **Performance with the additional encoders is often worse that without :-(**
  - Conclusion: prediction does not benefit from extra information
    - Or does it?

Sparse regions

- Performance metrics dominated by **densely connected** entities
  - There the **baselines** do well

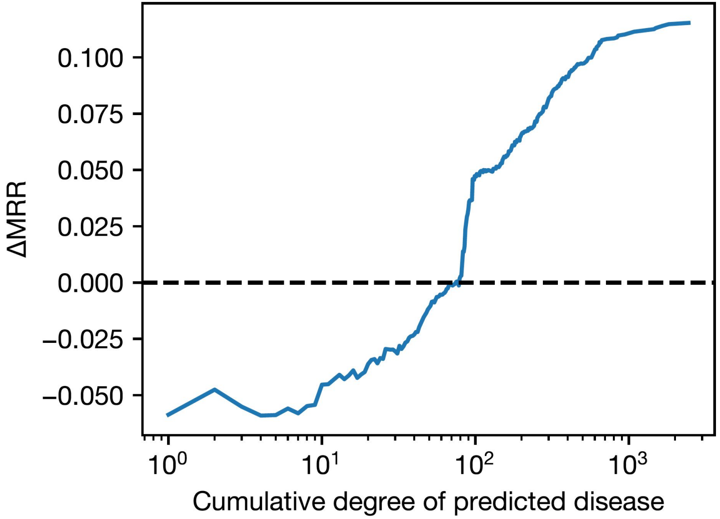
- However, half of the entities is in the low degree region (sparsely connected)
  - There we see improvement.



Use-case: Rare Diseases

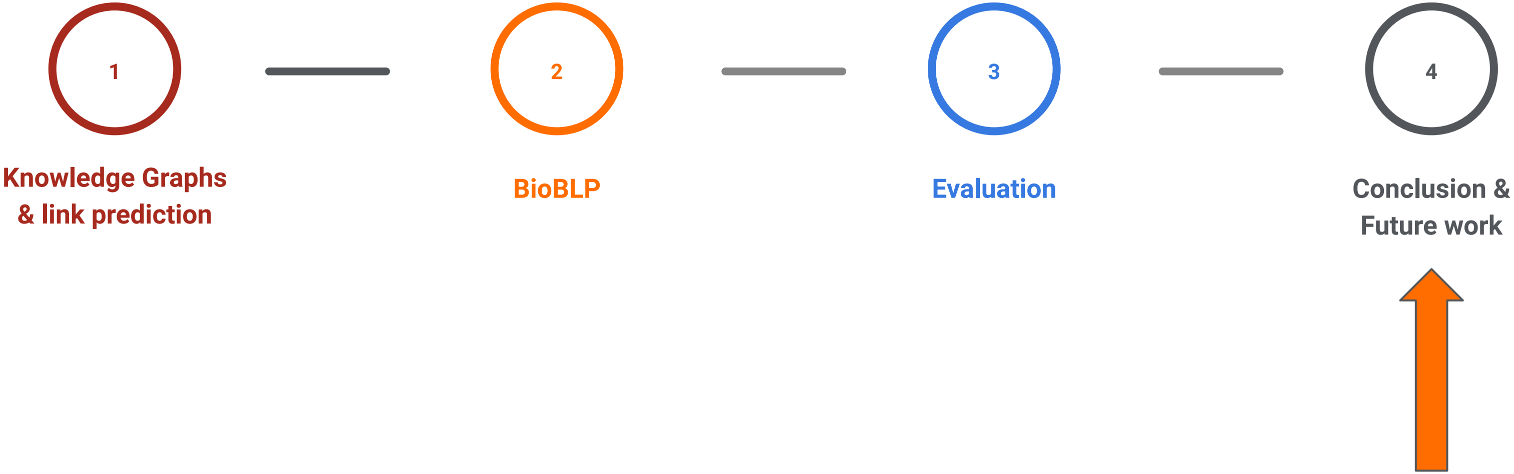
- Smaller degree indicates **understudied diseases** (e.g. Retinitis pigmentosa)

- A lot of potential to assist in treatment via **drug-repurposing**



Difference in *macro* MRR between RotatE and BioBLP-D when a disease is predicted, as we consider increasing values of node degree. We observe that when considering nodes with a degree of 74 or less, BioBLP-D results in improved link prediction performance (shown as negative values of ΔMRR)

Today



1 Knowledge Graphs & link prediction

2 BioBLP

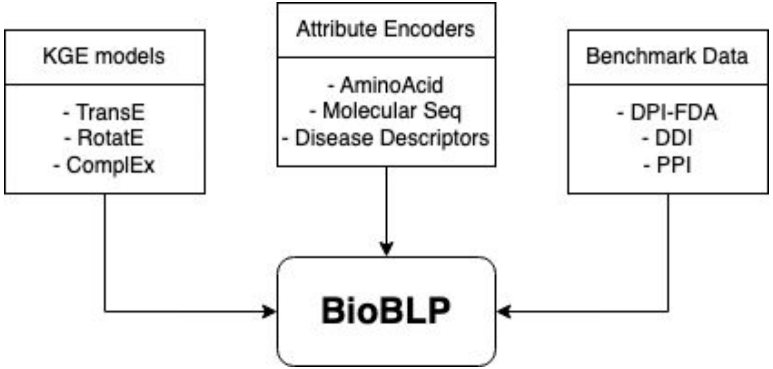
3 Evaluation

4 Conclusion & Future work

Lessons Learned

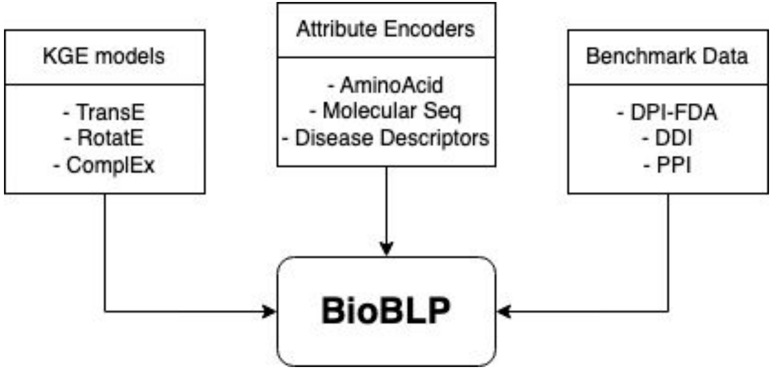
- Biodata == A lot & multimodal data
- KGEs and pretrained attribute encoders fit well together, but a challenge to optimize
    - Allows predictions with previously unseen entities
    - Additional signals from attributes can help, especially for low degree entities

Pre-training helps optimization



Future work

- More specialized encoders for entity attributes:
  - 3D structures (e.g AlphaFold)
- Add multiple modalities per entity type (SMILES + 3D structure)
- **Attention**-based mechanisms for multiple attribute representations
- Explore other biomedical evaluation tasks
- Explore inductive LP performance
  - Full use of attributes for representation



Contributions

- **A biomedical knowledge graph with multiple modalities**
- **A framework to train KGE with multiple encoders for their attributes**
- **The observation that the usual evaluation does not tell the whole story**
  - A model might not give a better score, but still work better for interesting cases

Michael Cochez m.cochez@vu.nl

Full paper: https://rdcu.be/dEufd

Code and data:
https://github.com/elsevier-AI-Lab/BioBLP