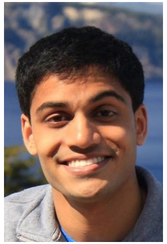
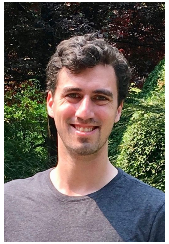
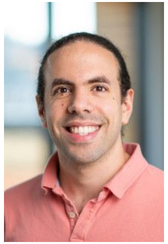
Scientific Language Models for Biomedical Knowledge Base Completion: An Empirical Study

Rahul Nadkarni[1], David Wadden[1], Iz Beltagy[2], Noah A. Smith[1,2], Hannaneh Hajishirzi[1,2], Tom Hope[1,2]

[1] Paul G. Allen School of Computer Science & Engineering, University of Washington

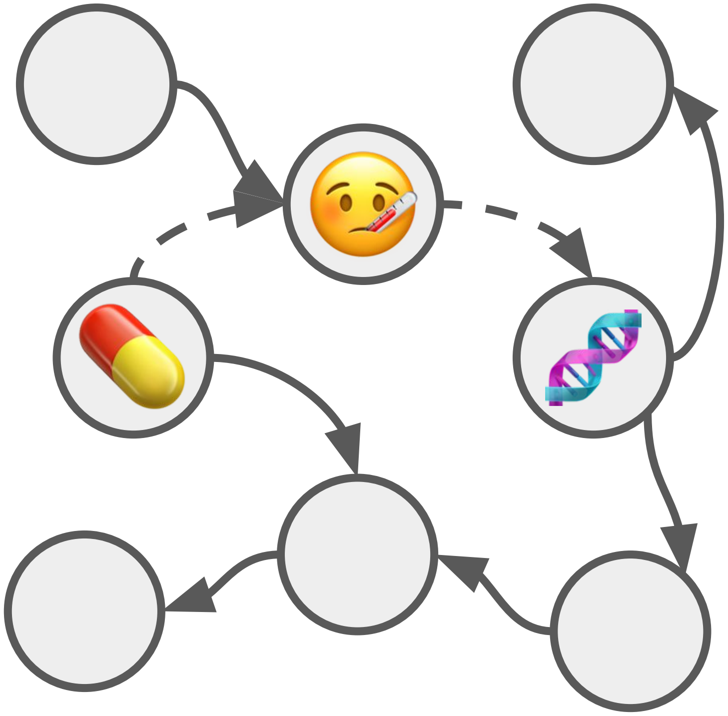
[2] Allen Institute for Artificial Intelligence (AI2)


PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

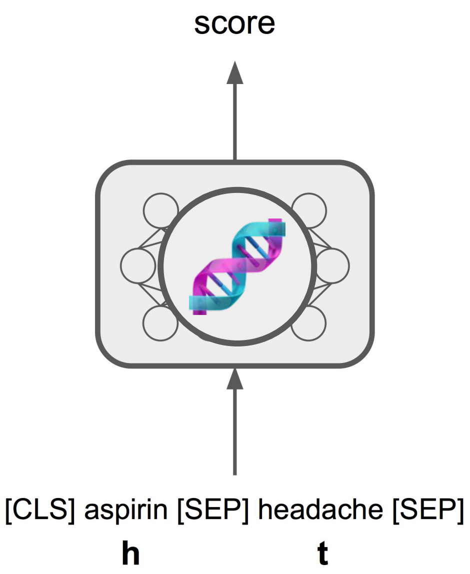
AI2

Biomedical Knowledge Graph Completion

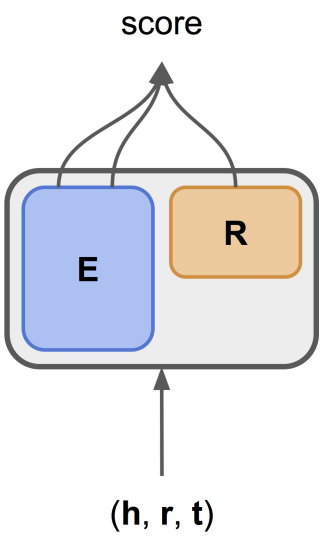


- Relations between entities
  - Repurposing drugs for diseases
  - Mapping diseases to genes
- Frame as biomedical knowledge graph completion

LMs for Biomedical Knowledge Graph Completion



scientific Transformer LM



KGE

First to systematically apply scientific LMs for KG completion and compare to KGE models

Datasets

RepoDB

drugs, diseases

2.7k entities, 6.7k triples

Hetionet

drugs, diseases, genes, symptoms, side effects

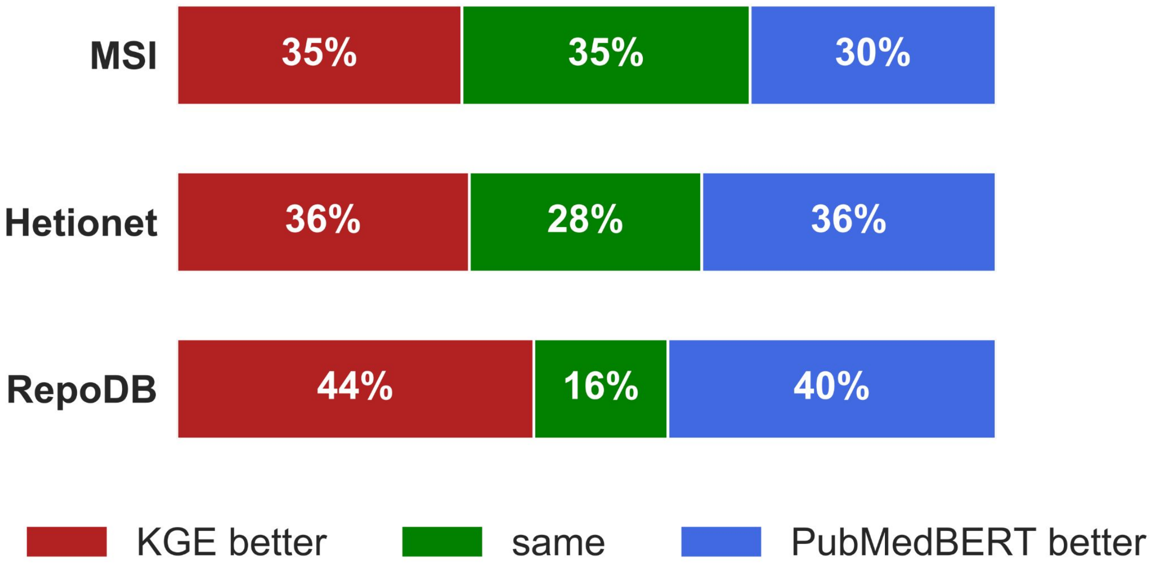
12.7k entities, 156k triples

MSI

drugs, diseases, proteins, protein functions

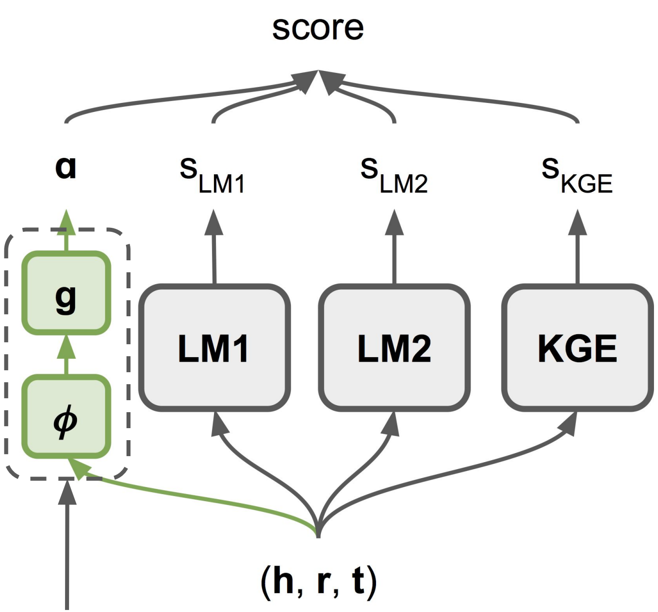
30k entities, 485k triples

Relative Performance

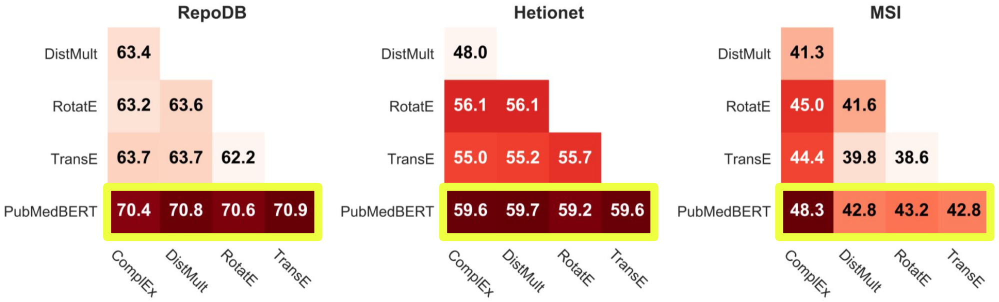


LMs and KGE models perform well on different subsets of examples

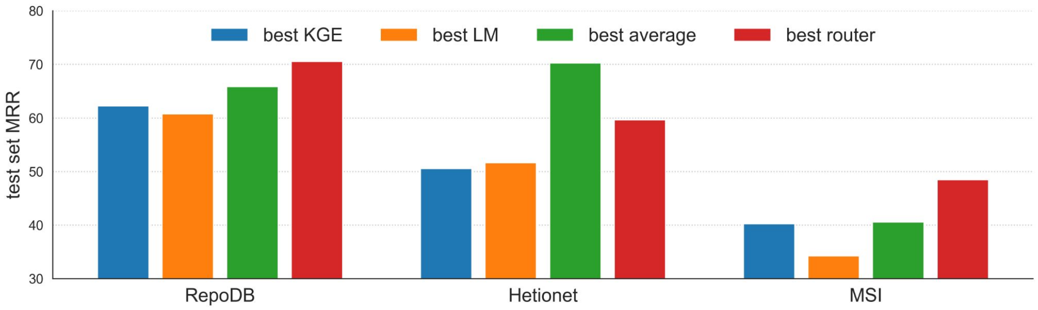
Integrating Models



- weighted average
- router classifier

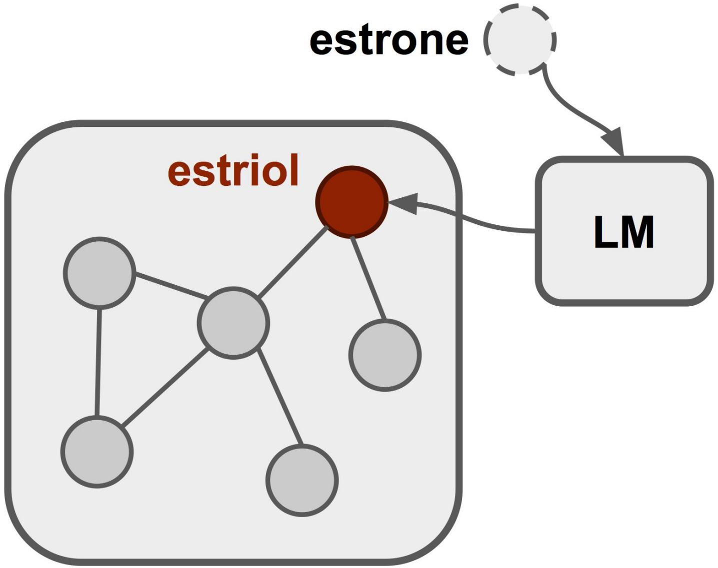


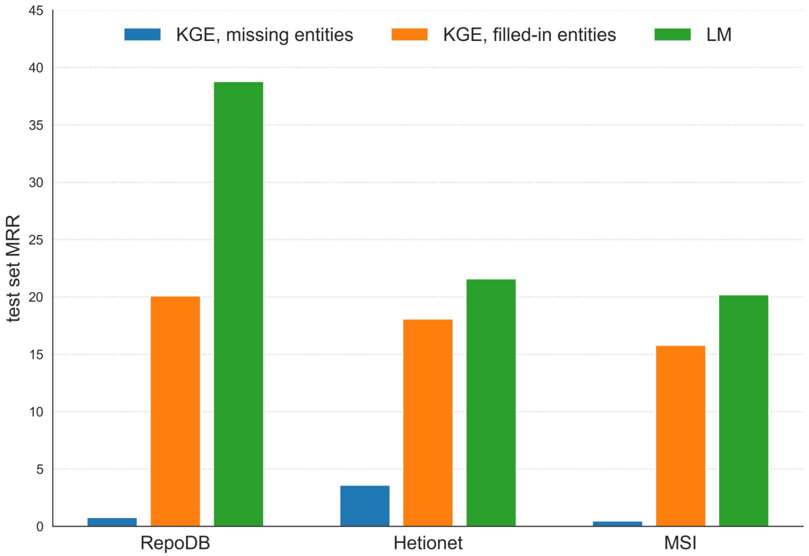
Combinations with an LM perform better



Best router can outperform best weighted avg.

Inductive Performance





LMs perform well (and can improve KGE performance) on unseen entities

For code and data, visit:

`github.com/rahuln/lm-bio-kgc`