Knowledge Graph Embedding for Chemical Effect Prediction

Faglunsj 21.01.20

Erik B. Myklebust

ErikBMyklebust

ebm@niva.no

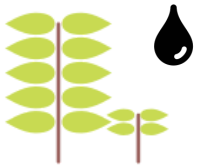
Norwegian Institute for Water Research

The Alan Turing Institute

CITY UNIVERSITY OF LONDON EST 1894

UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

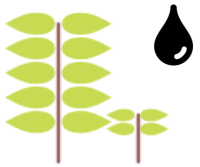
Ecological Risk Assessment



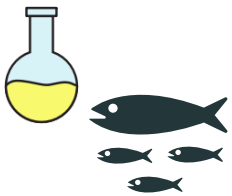
Exposure

Risk assessment is an estimation of cumulative risk on individuals, populations, communities, and ecosystems from chemical pollutants.

Ecological Risk Assessment



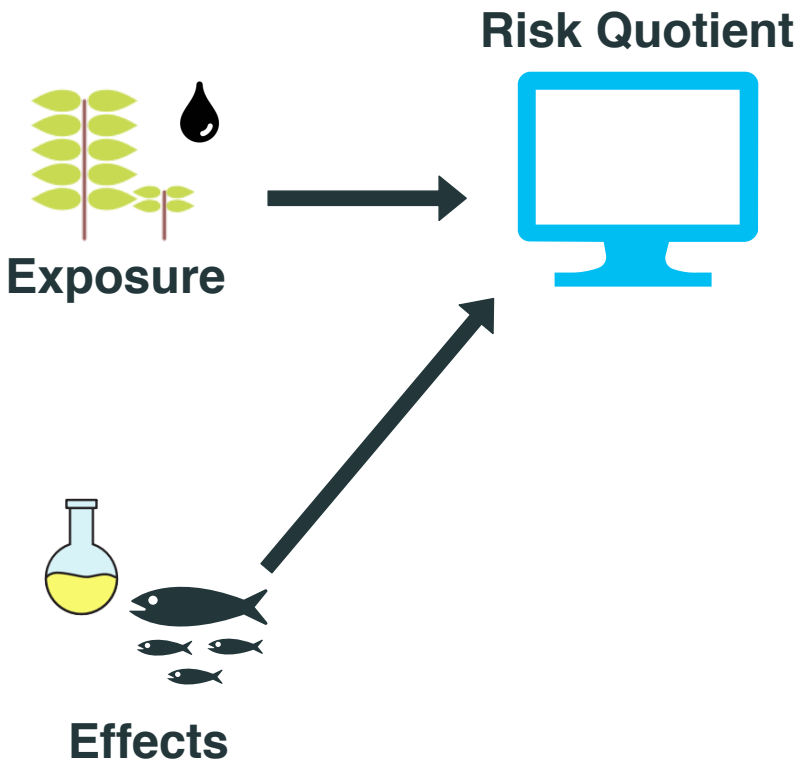
Exposure



Effects

Effect concentrations are found using organism experiments.

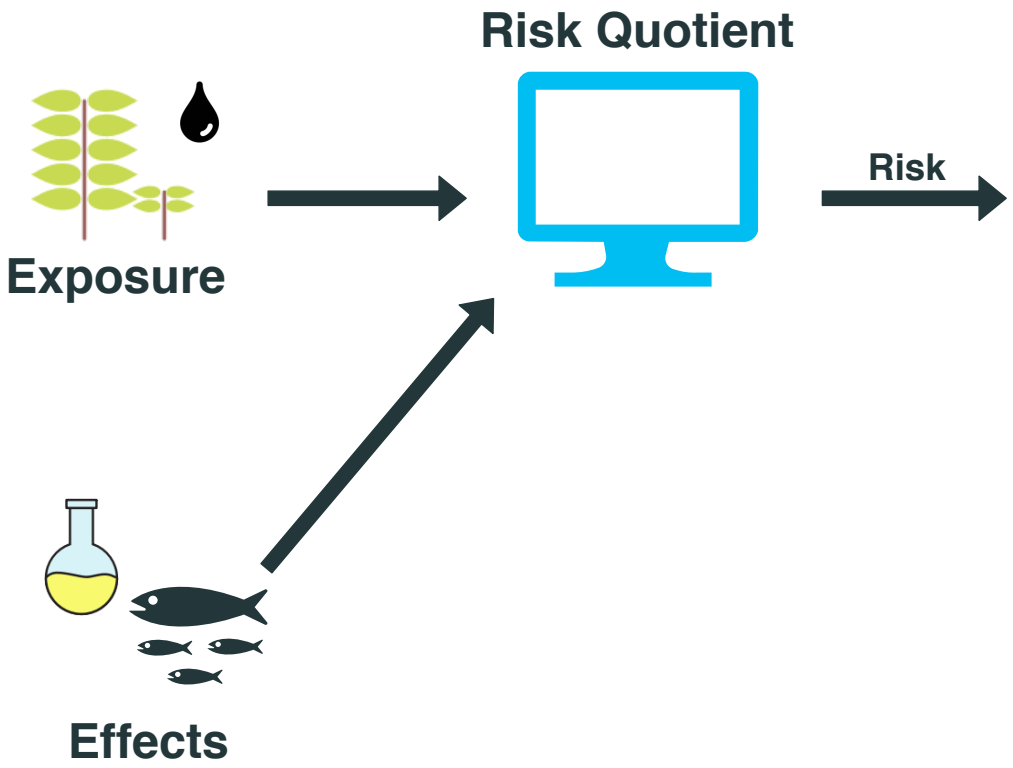
Ecological Risk Assessment



$$RQ = \frac{\text{environmental concentration}}{\text{effect concentration}}$$

RQs coverage is limited by effect concentration experiments.

Ecological Risk Assessment

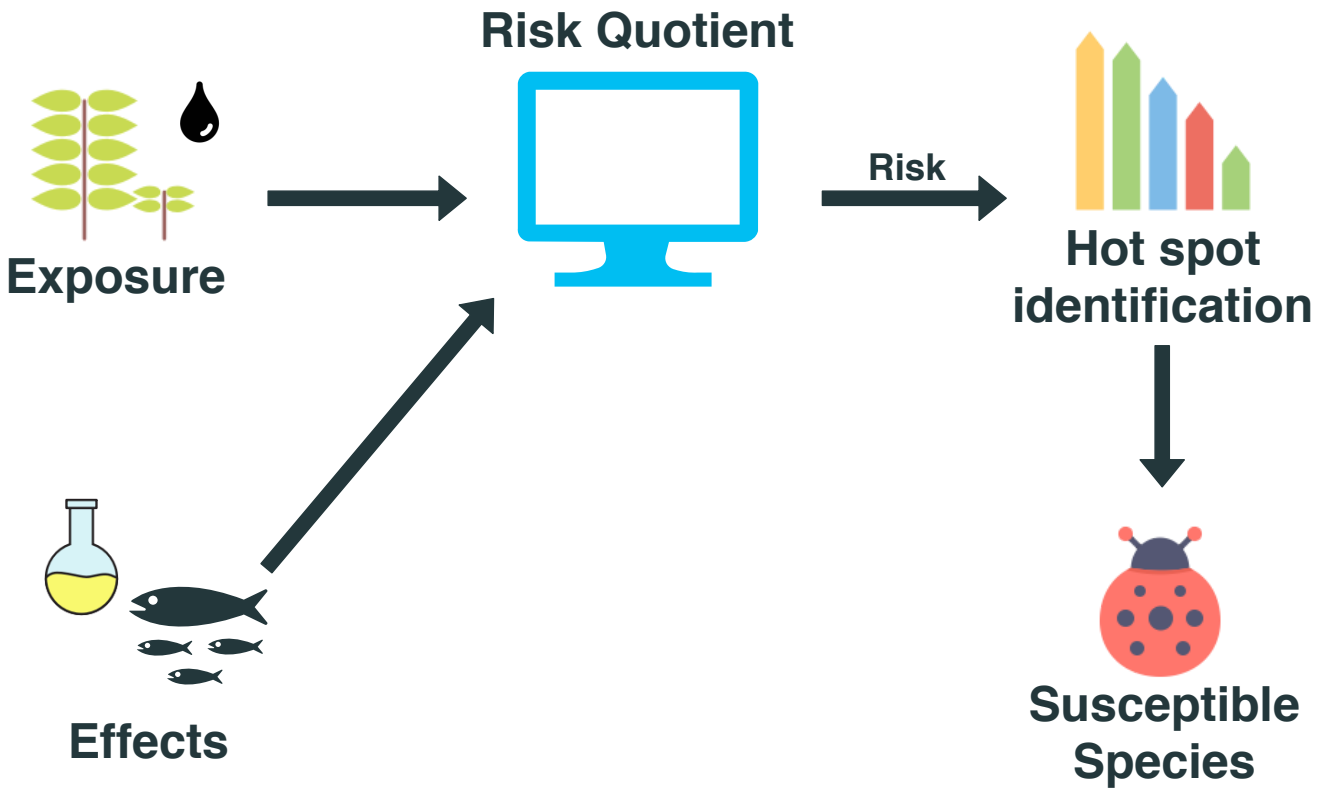


$$risk_{group} \approx \sum^{chemicals} RQ$$

Risk for a group of species.

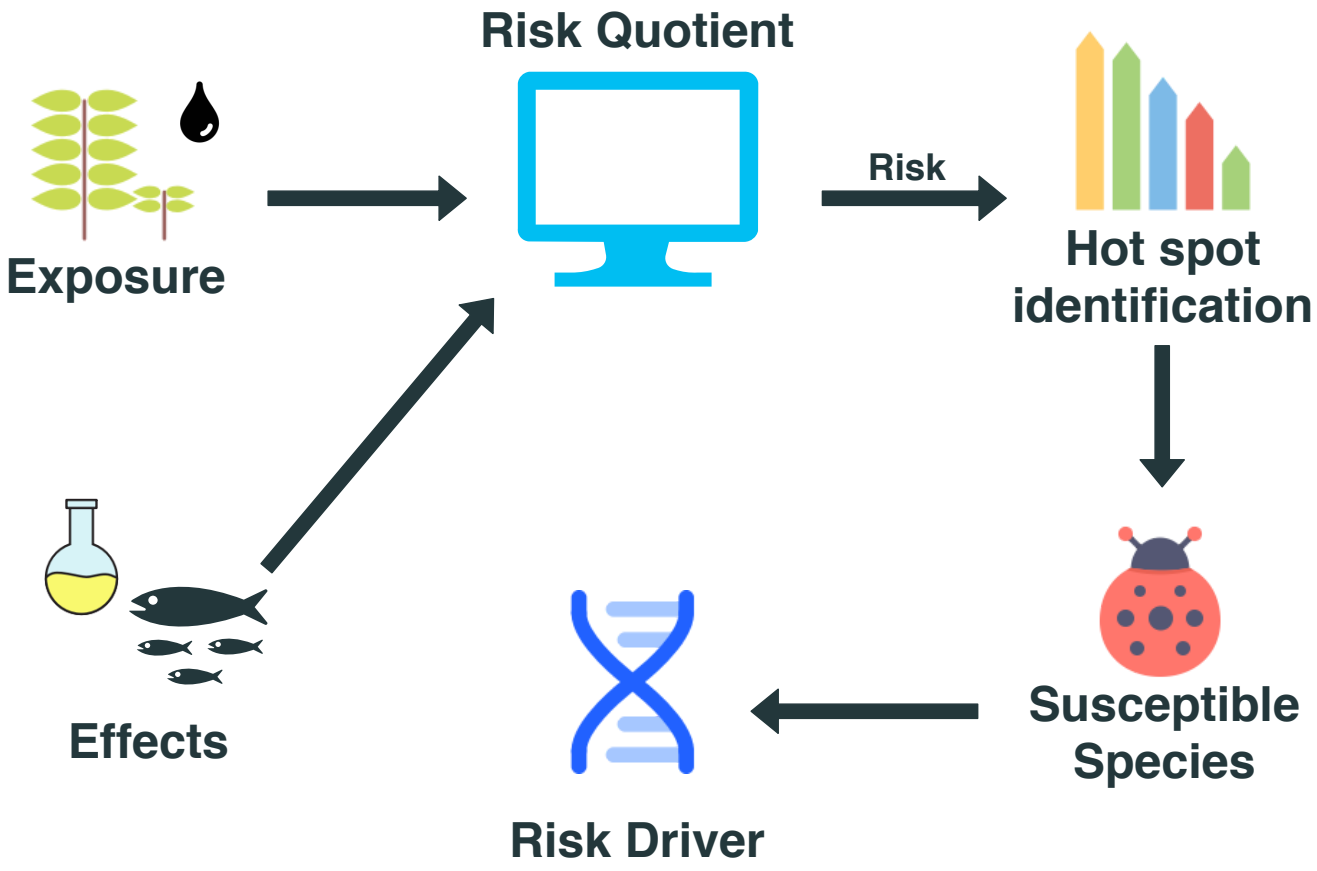
The group can contain all species in the ecosystem.

Ecological Risk Assessment



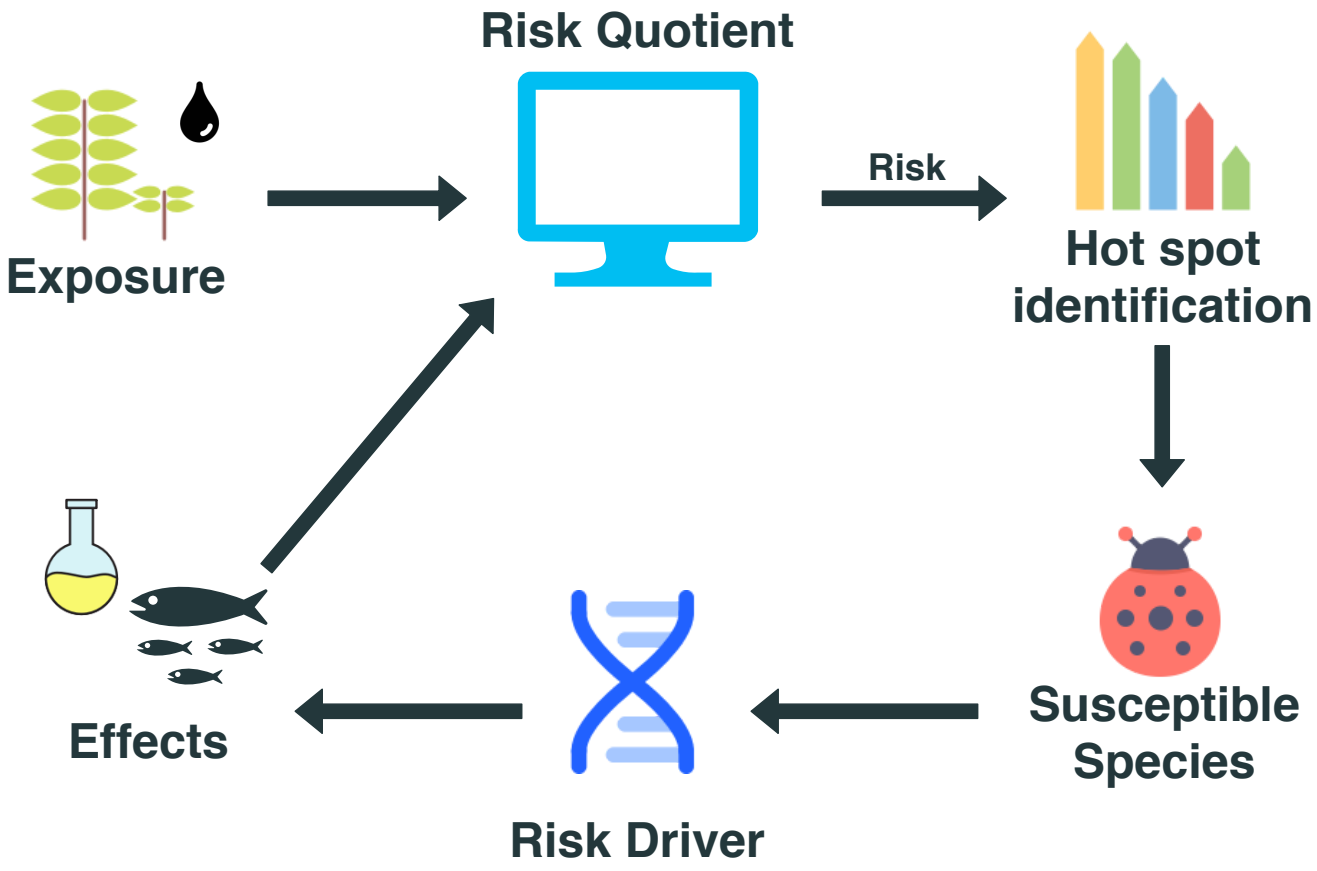
The risk is used to find further susceptible species.

Ecological Risk Assessment



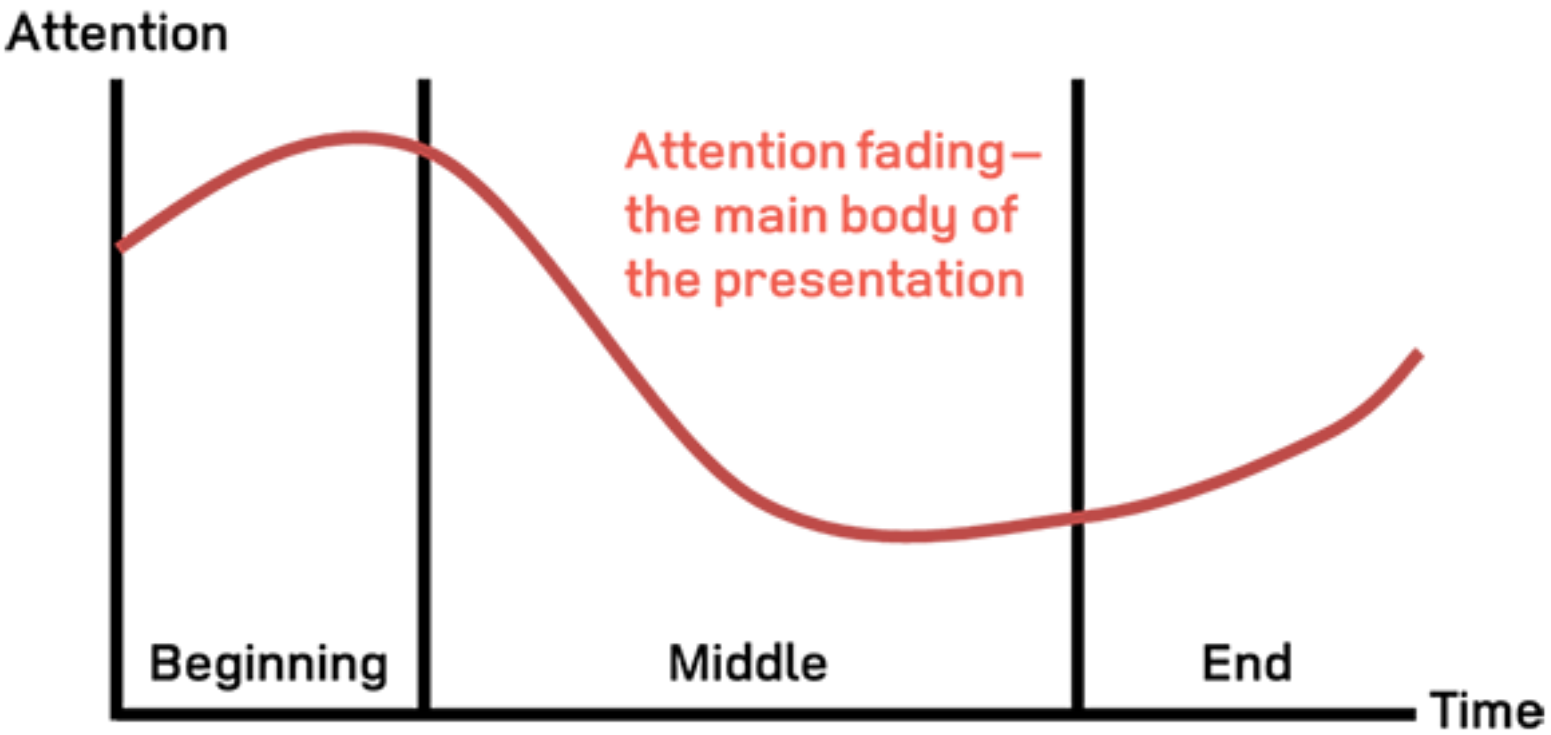
Risk driver describes *how* the chemical affects an organism.

Ecological Risk Assessment

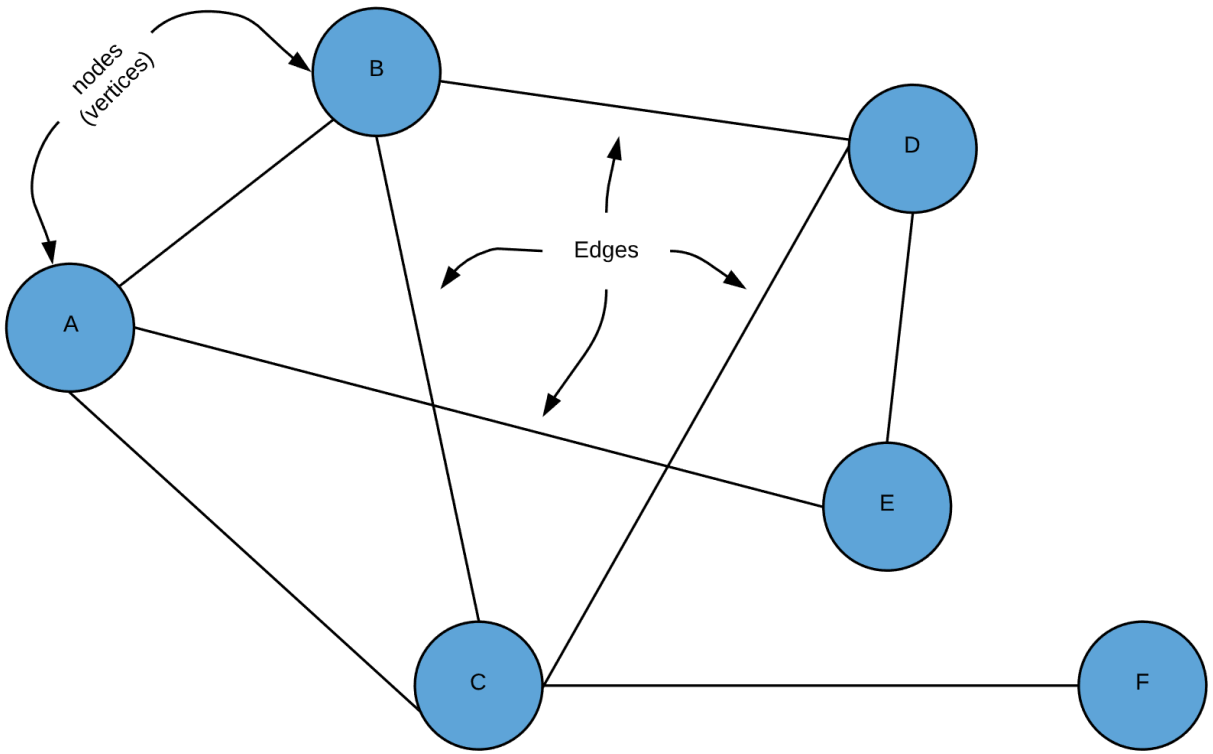


New effect hypotheses are then tested in the laboratory.

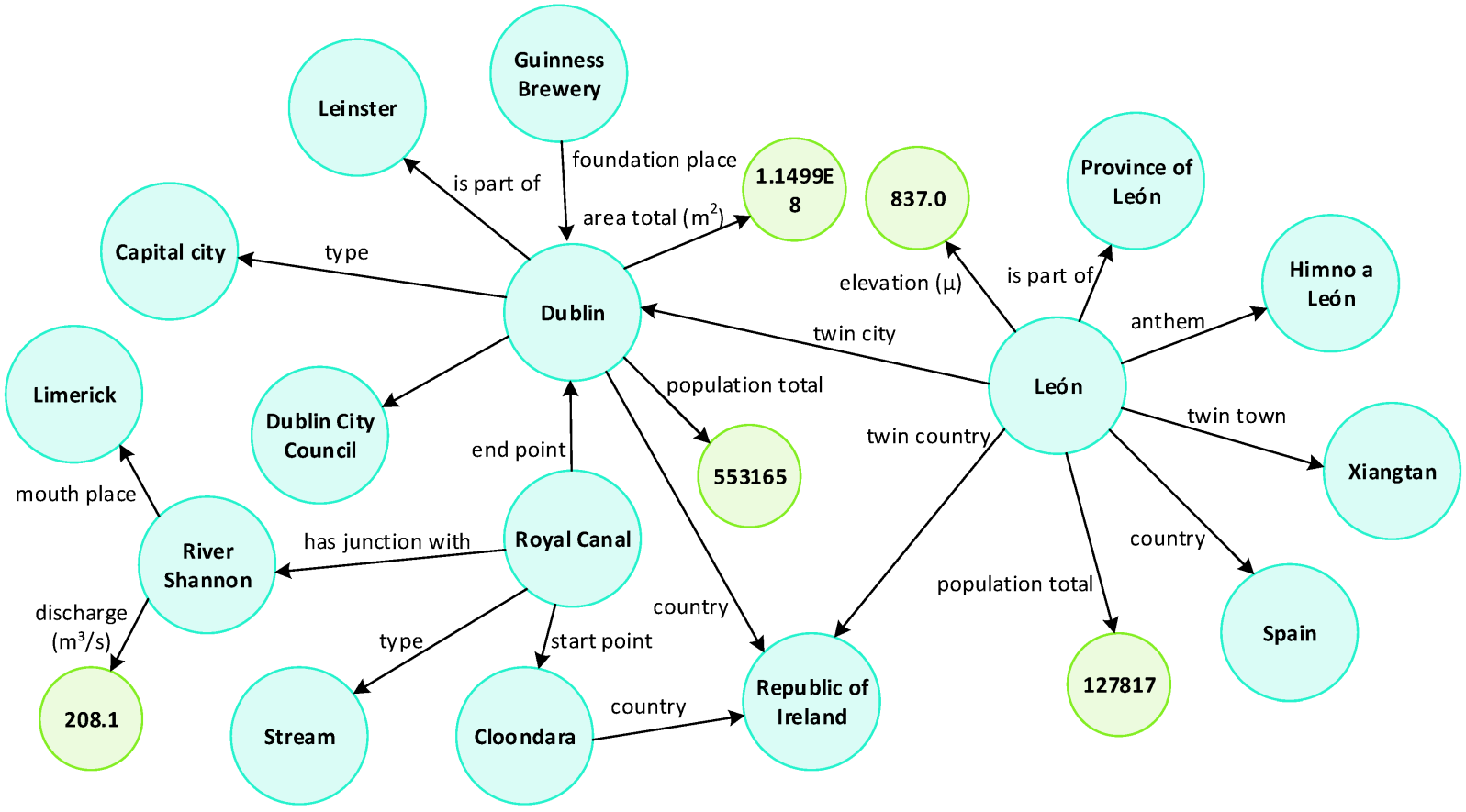
What is a graph?



What is a graph?



What is a knowledge graph?

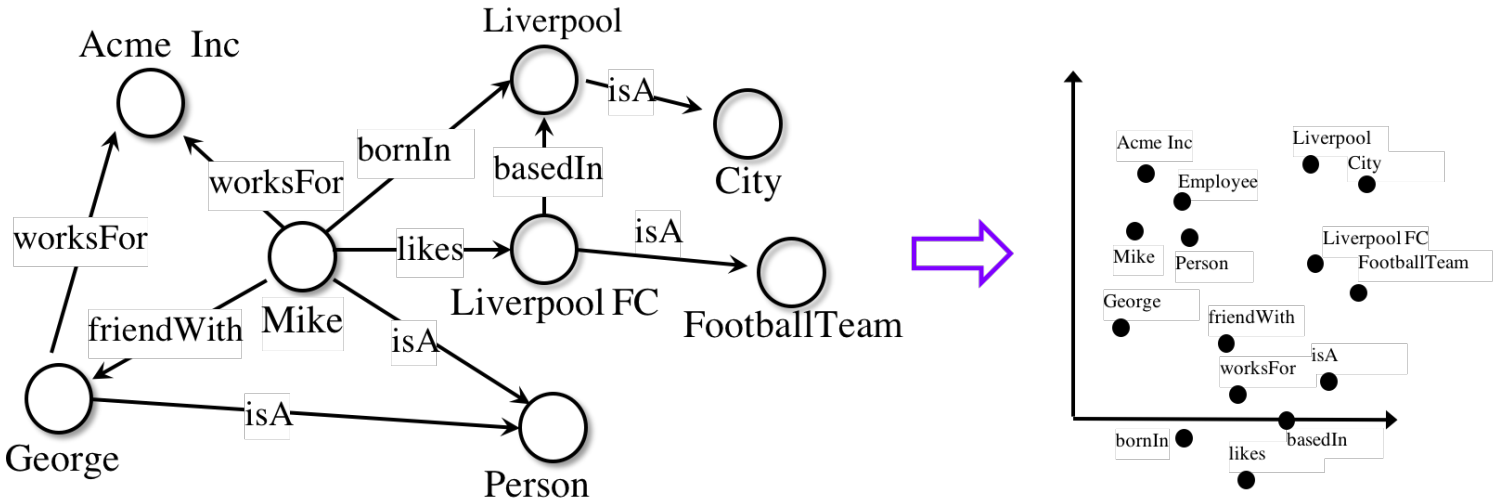


A knowledge graph is a set of triples (facts) on the form

$$(e_s, r, e_o) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$$

$\mathcal{E}$: *entities* $\mathcal{R}$: *relations*

Knowledge graph embedding



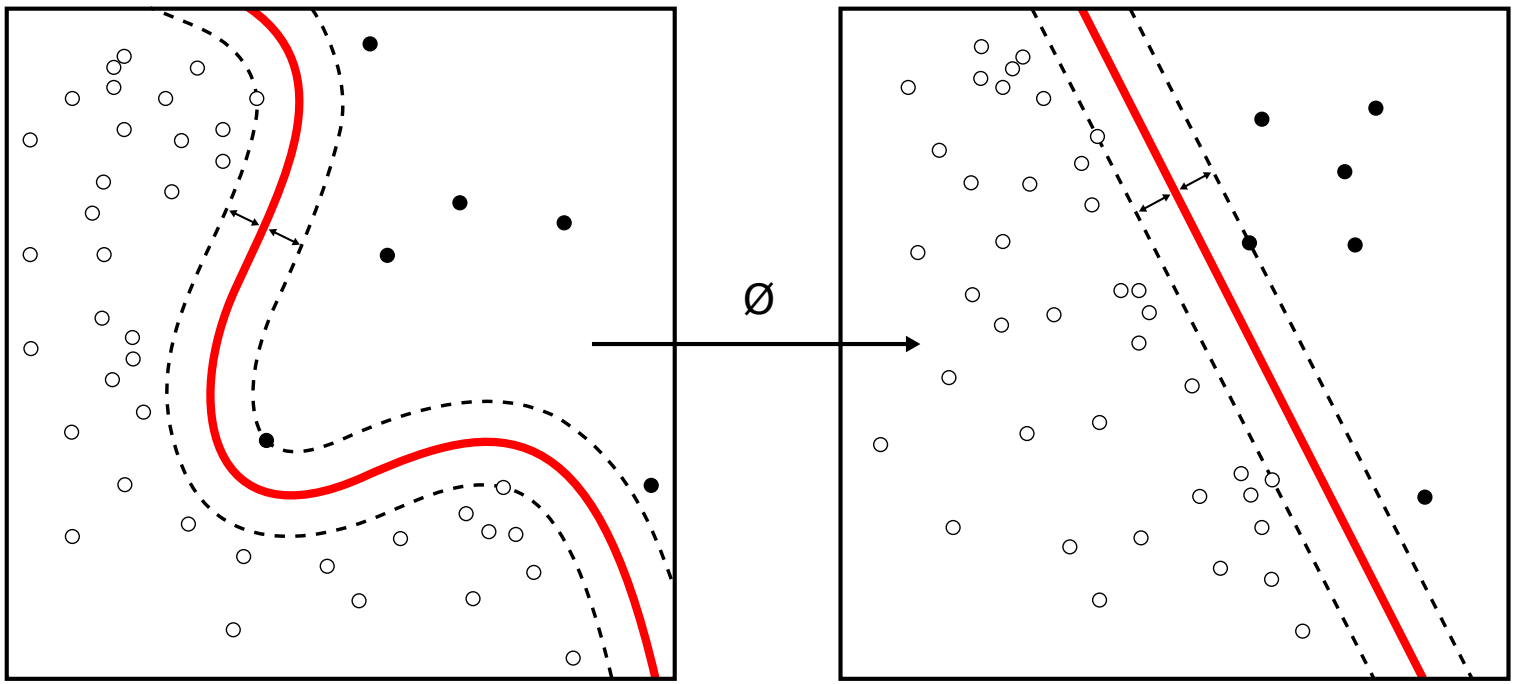
$$S = S(\mathbf{v_s}, \mathbf{v_r}, \mathbf{v_o}), \quad P((e_s, r, e_o)|KG) = \sigma(S)$$

$$S = \|\mathbf{v_s} + \mathbf{v_r} - \mathbf{v_o}\|, \quad \text{(TransE)}$$

$$S = \langle \mathbf{v_s}, \mathbf{v_r}, \mathbf{v_o} \rangle, \quad \text{(DistMult)}$$

$$S = \mathbf{v_r}^T(\mathbf{v_s} \star \mathbf{v_o}), \quad \text{(HolE)}$$

Machine Learning



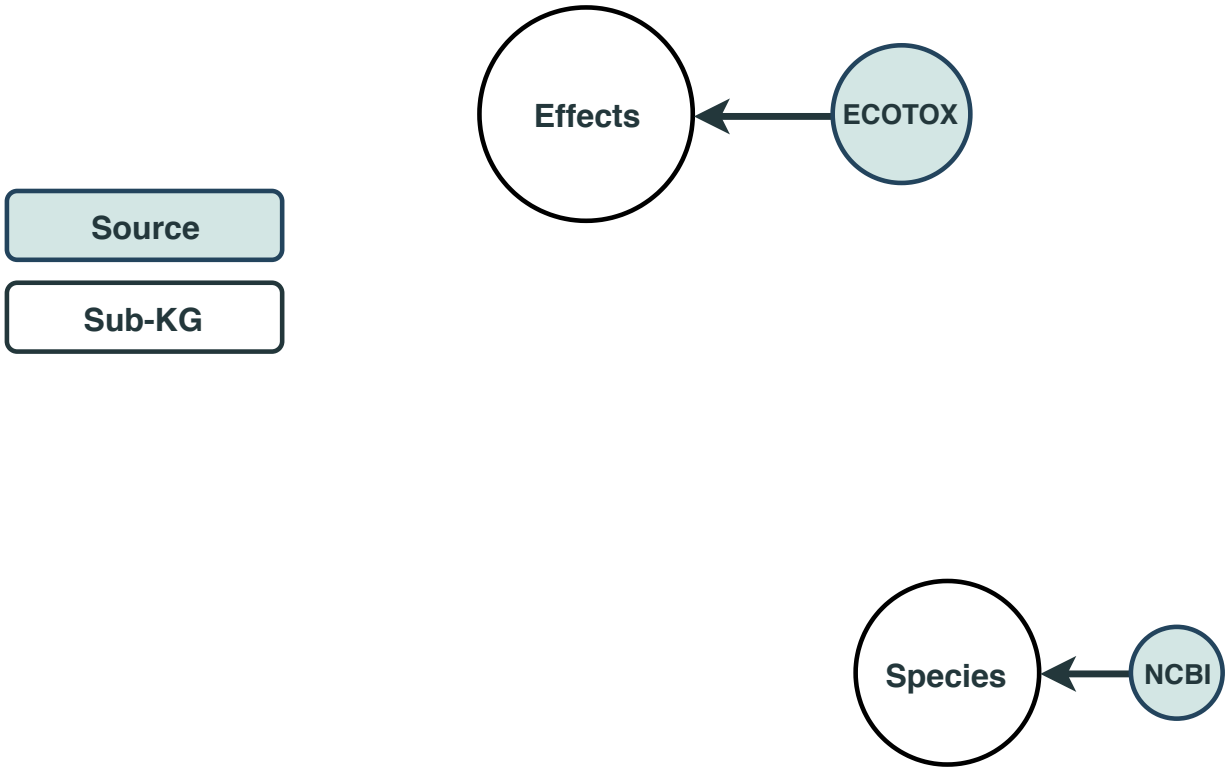
The TERA Knowledge Graph

The Toxicological and Risk Assessment (TERA) knowledge graph integrates data sources varying in format.

The TERA Knowledge Graph

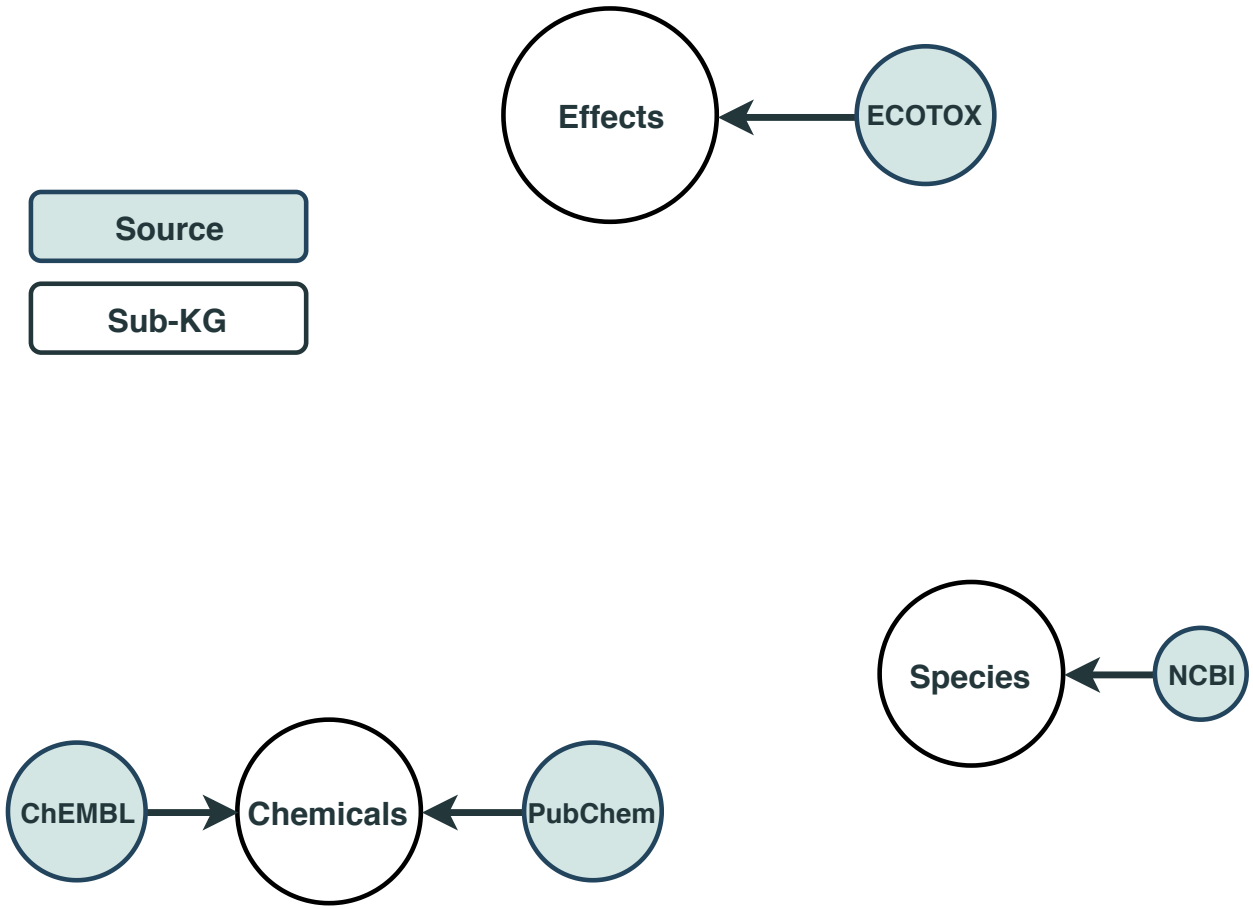
ECOTOX is the largest (public) source of effect data.

The TERA Knowledge Graph



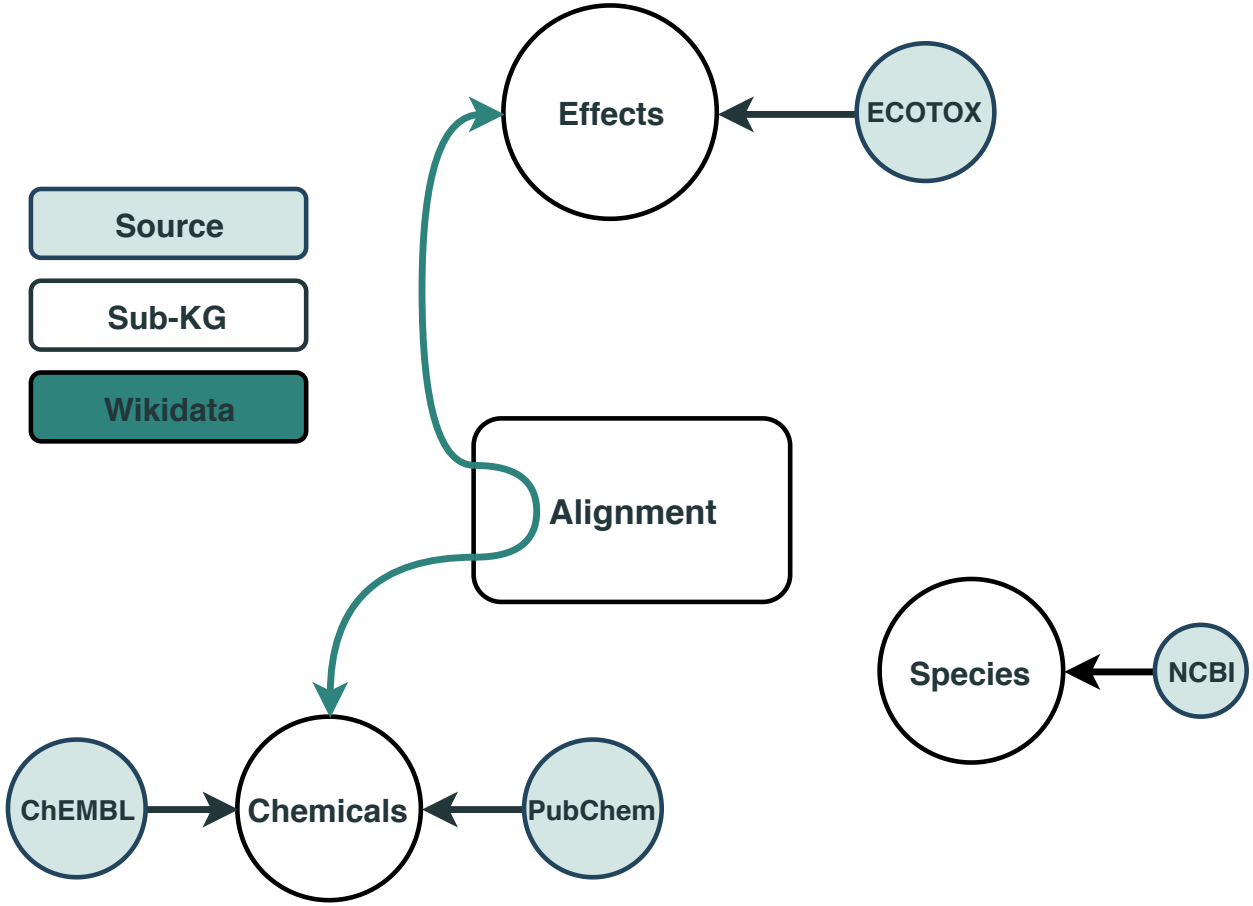
NCBI's tabular taxonomy is converted to a hierarchy.

The TERA Knowledge Graph



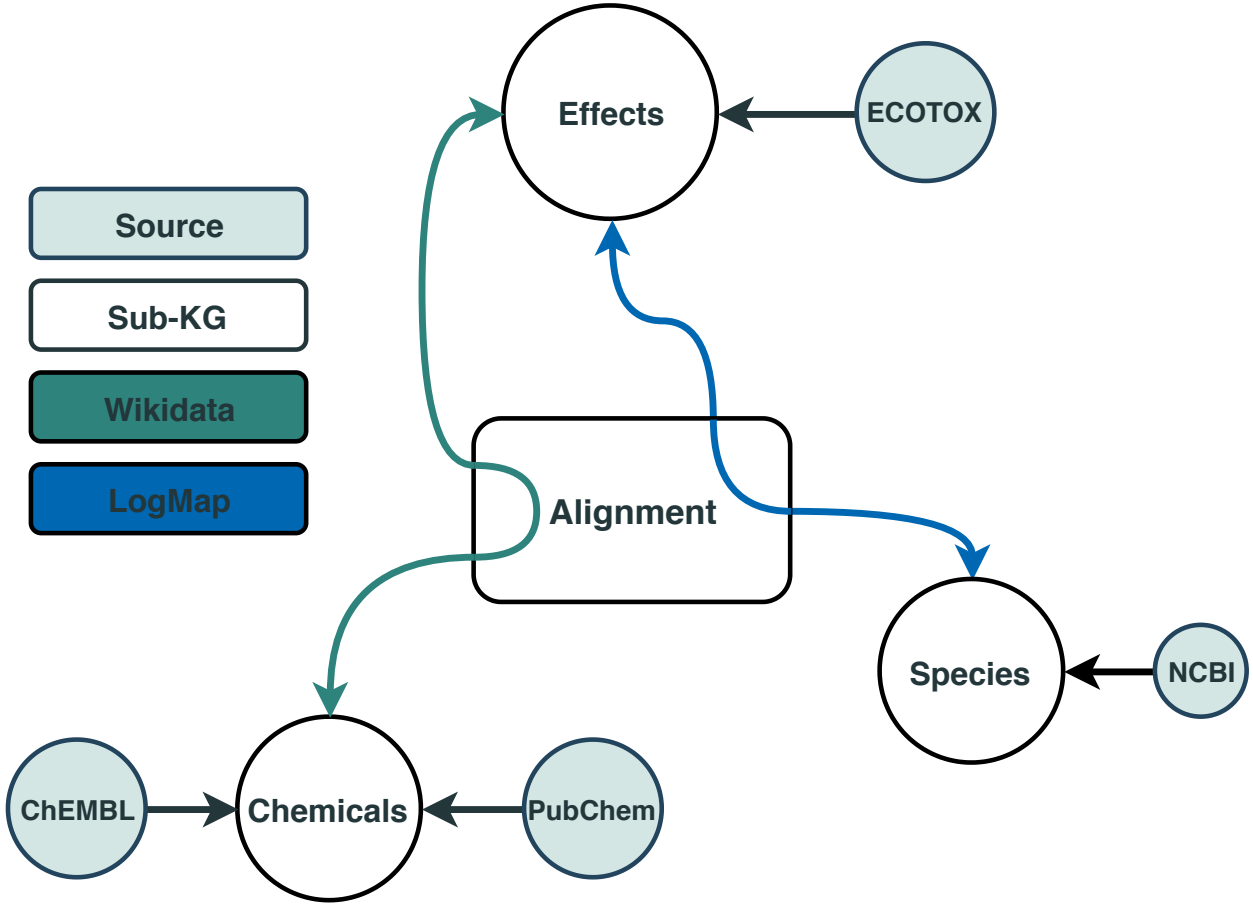
Importing the ChEMBL and PubChem knowledge graph.

The TERA Knowledge Graph



Aligning proprietary chemical identifiers in ECOTOX to open identifiers in PubChem.

The TERA Knowledge Graph



Aligning taxonomies using ontology alignment tool LogMap.

Effect Prediction Problem Definition

Effect Prediction Problem Definition

C_1

C_2

C_3

Chemicals

Effect Prediction Problem Definition

C_1

S_1

C_2

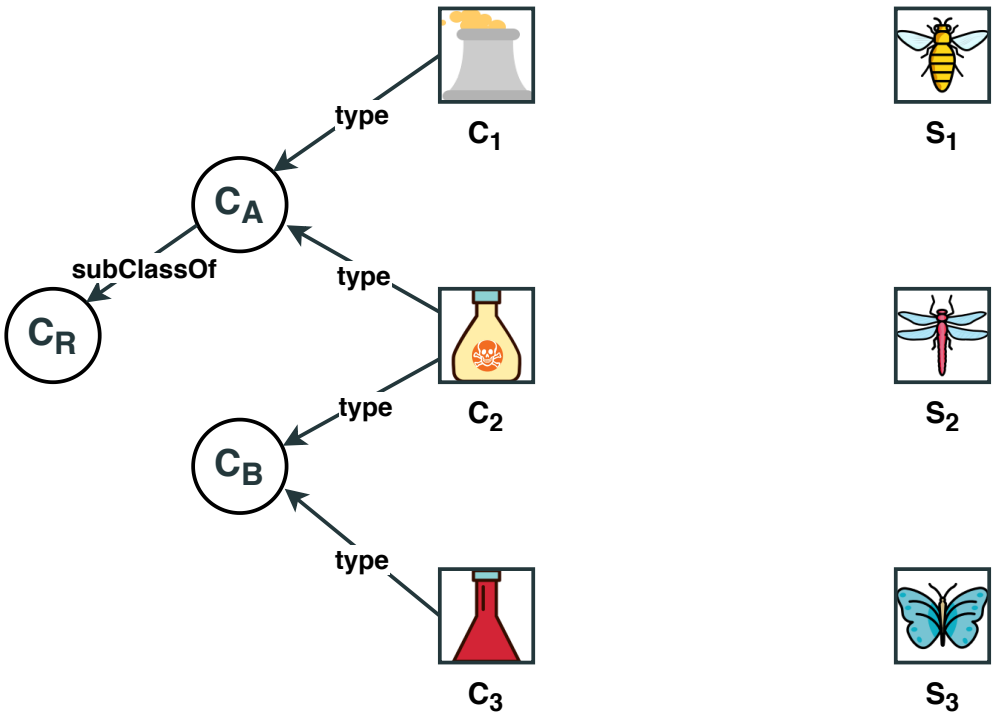
S_2

C_3

S_3

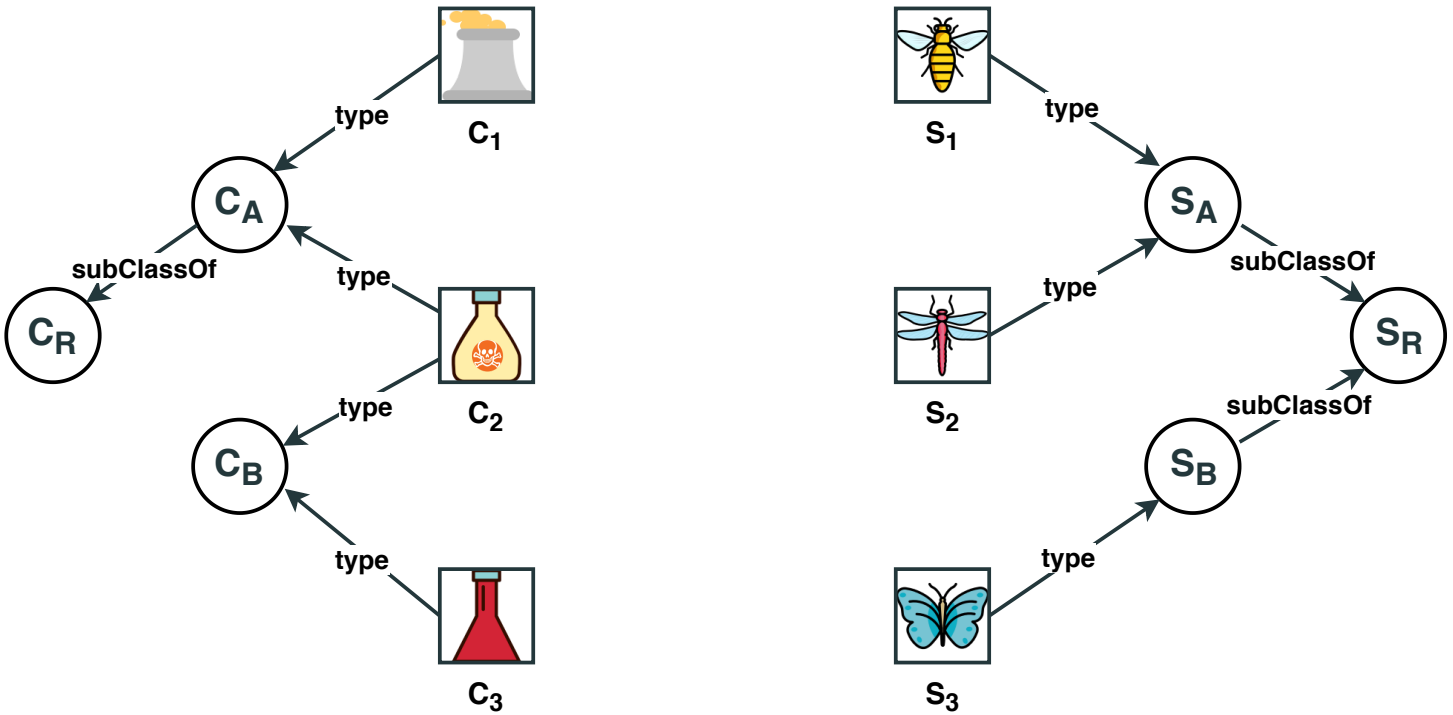
Species

Effect Prediction Problem Definition



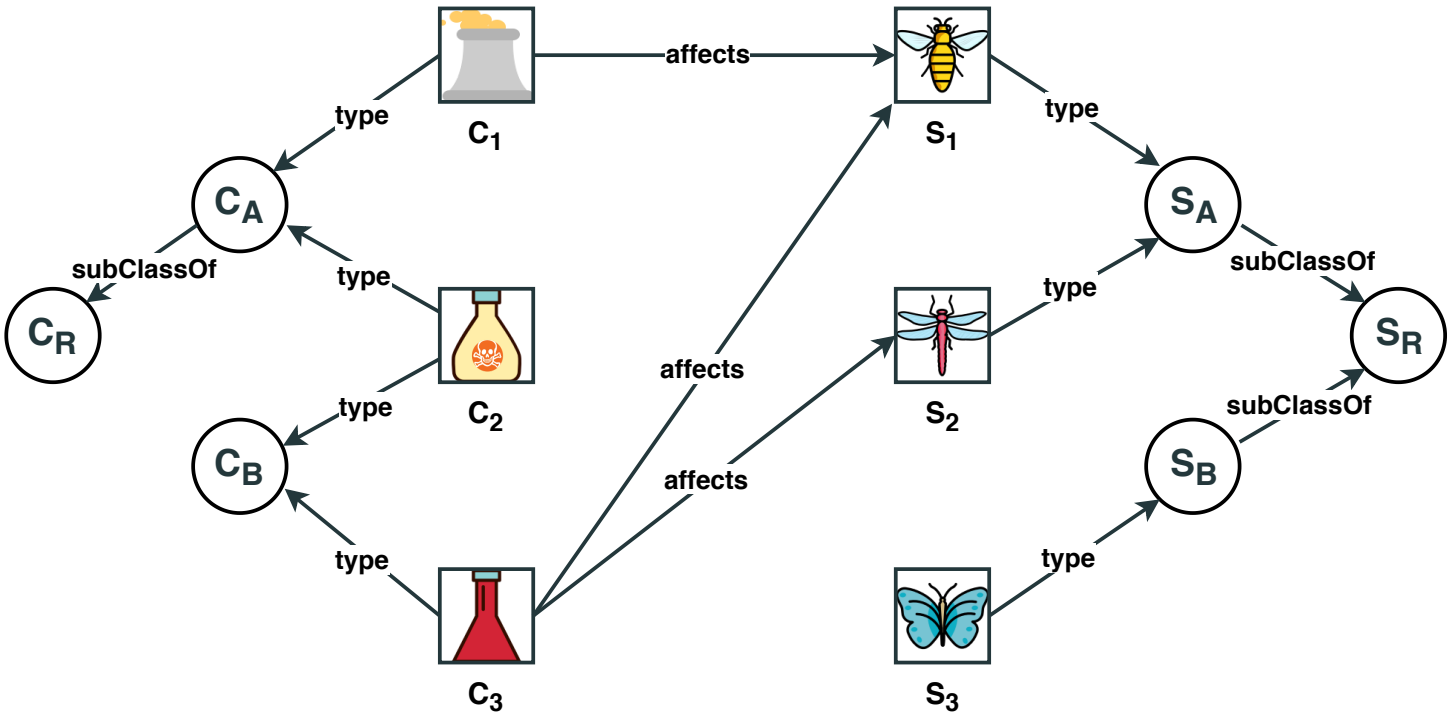
Chemical classification

Effect Prediction Problem Definition



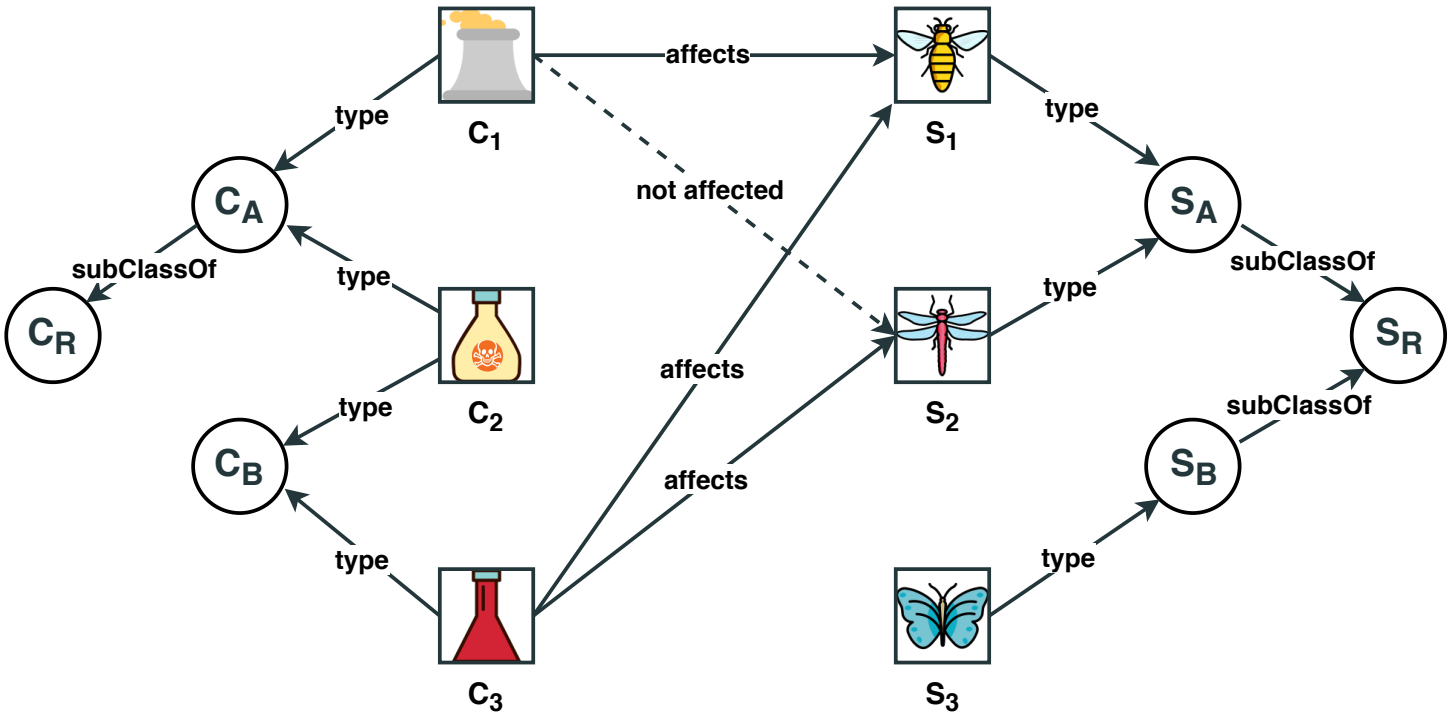
Taxonomy

Effect Prediction Problem Definition



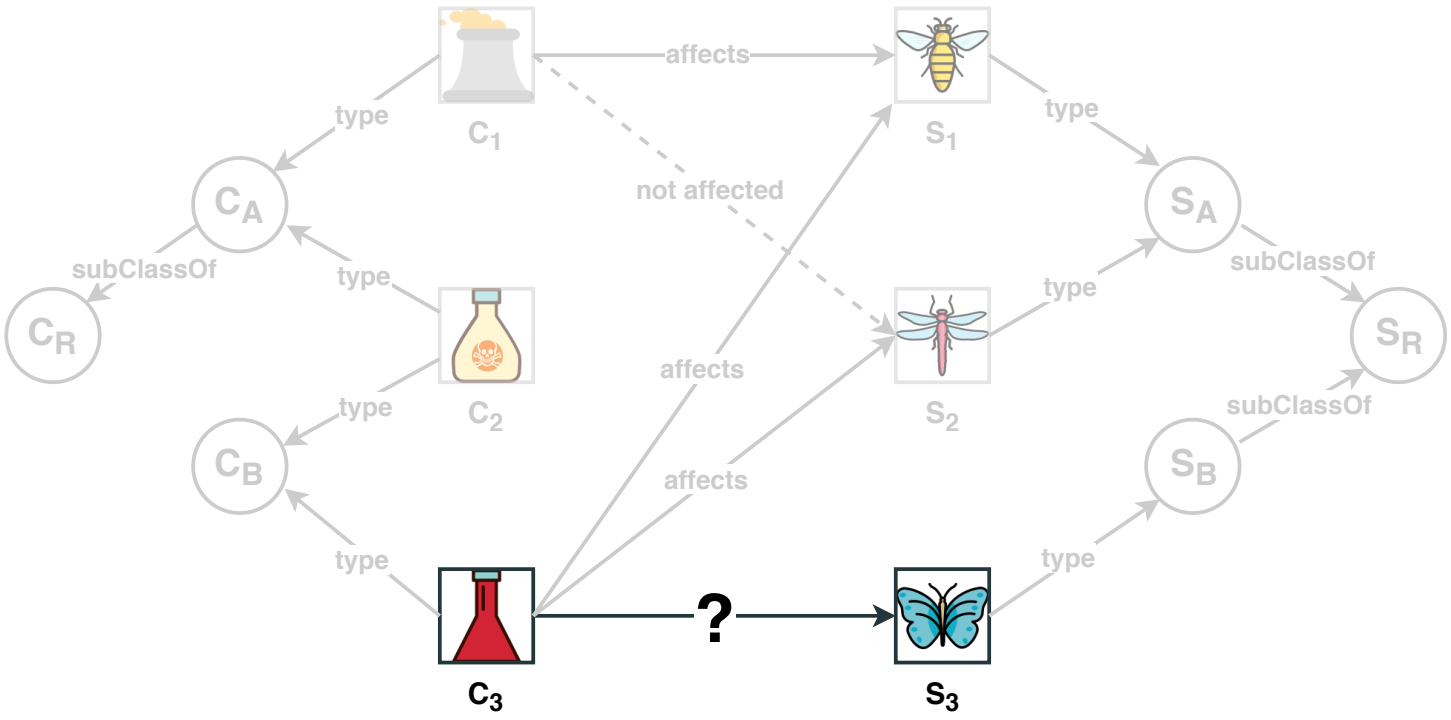
Positive samples

Effect Prediction Problem Definition



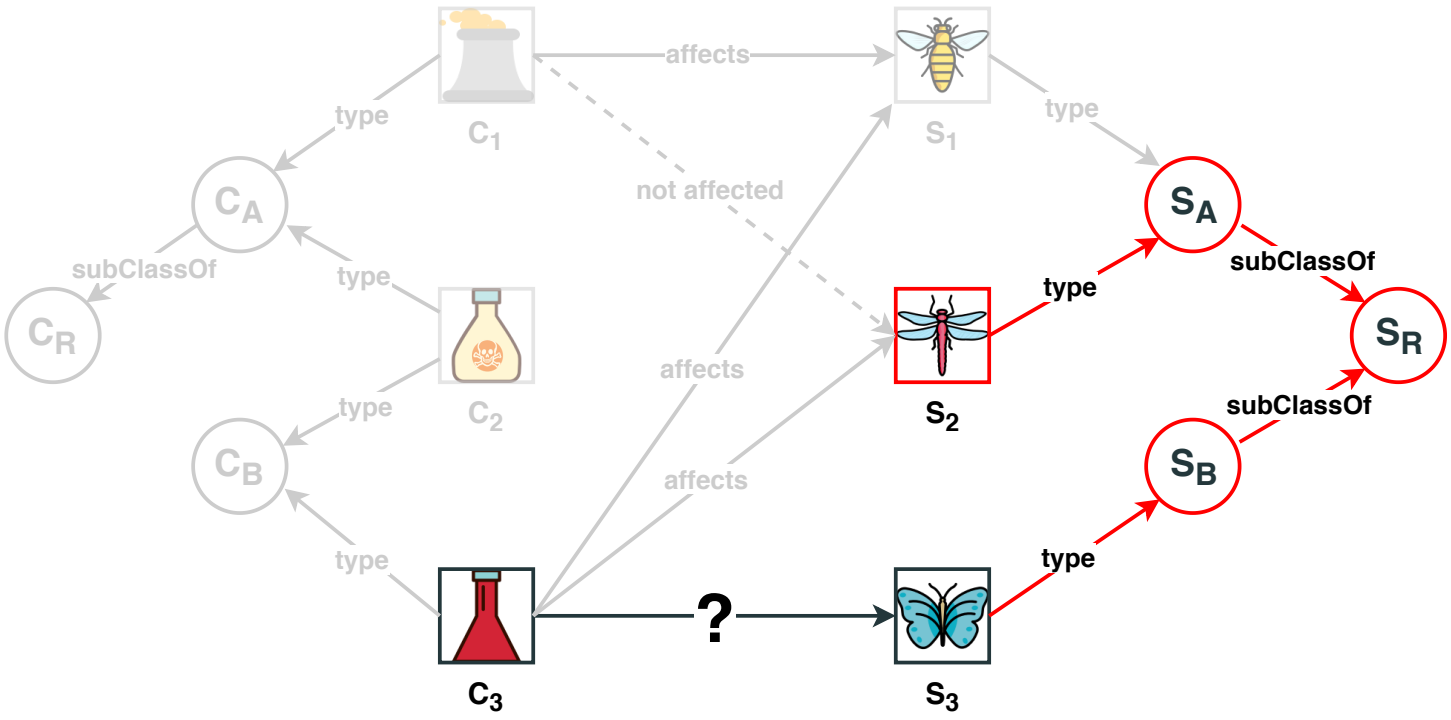
Negative samples

Taxonomic Distance Model - Baseline (BL)



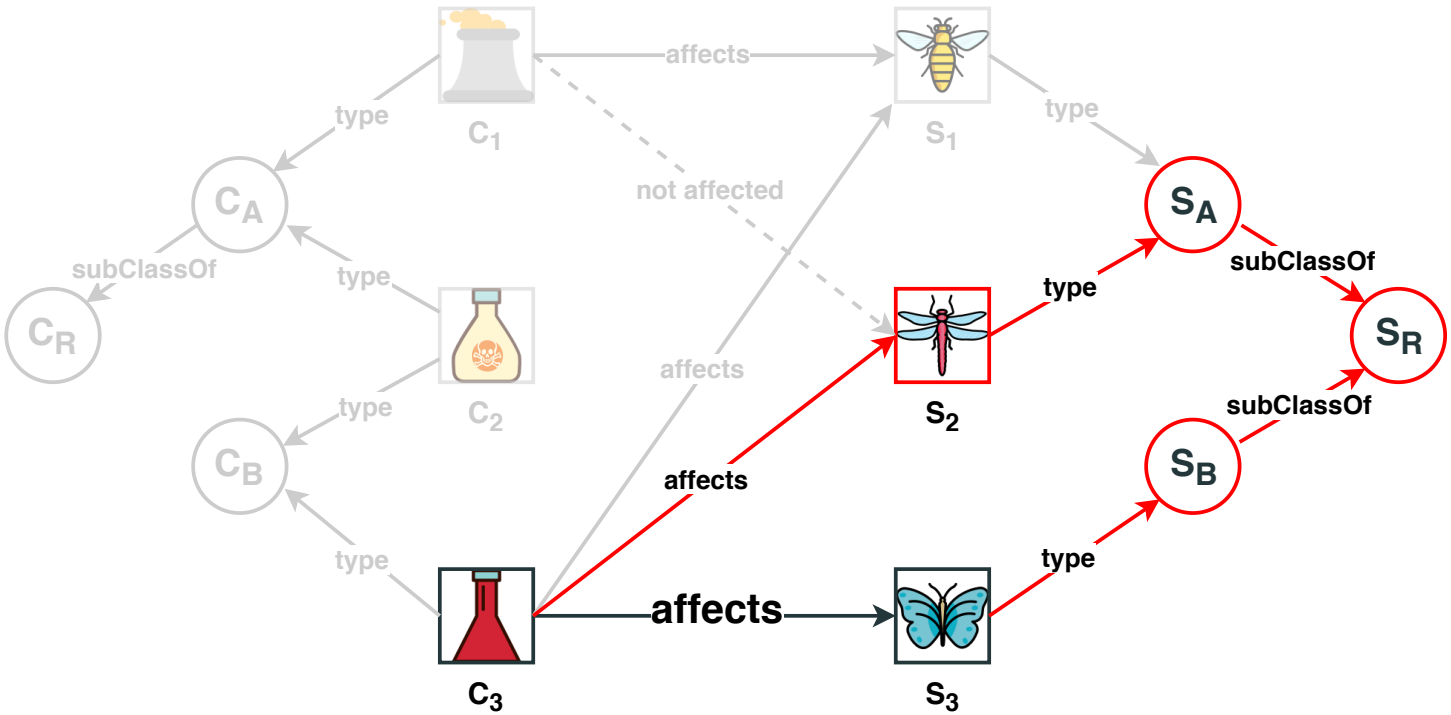
Does C_3 affect S_3?

Taxonomic Distance Model - Baseline (BL)



$$dist(S_3, S_2) = 4$$

Taxonomic Distance Model - Baseline (BL)

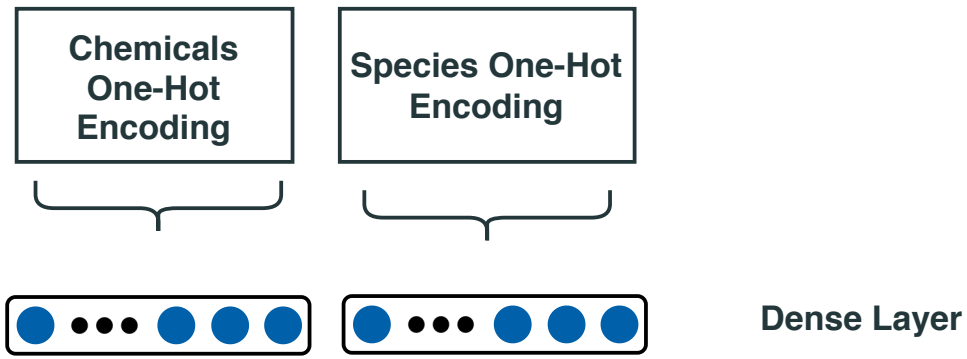


Yes, C_3 affects S_3

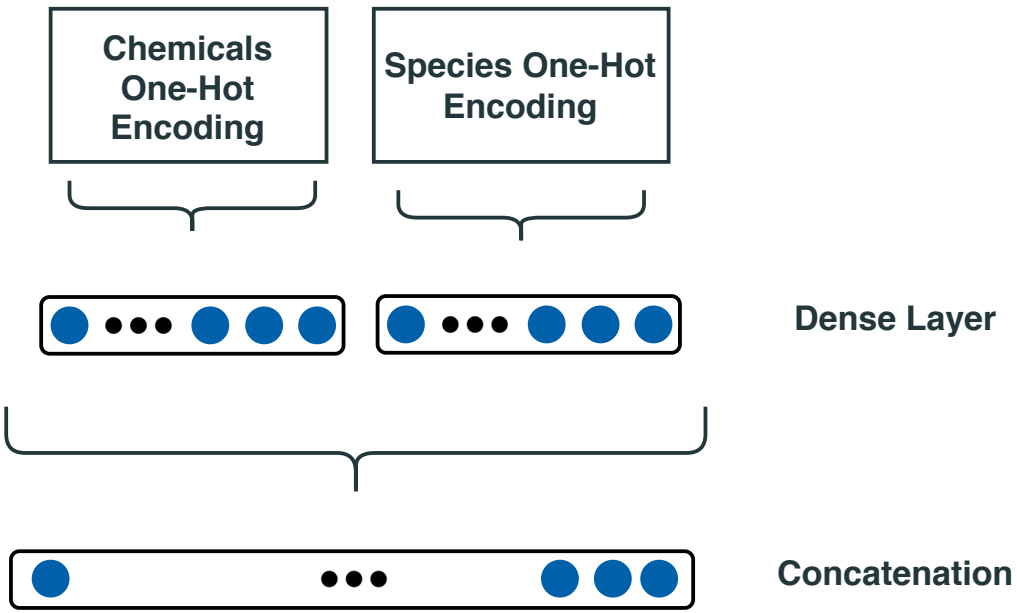
Multi-layer perceptron (MLP)

Multi-layer perceptron (MLP)

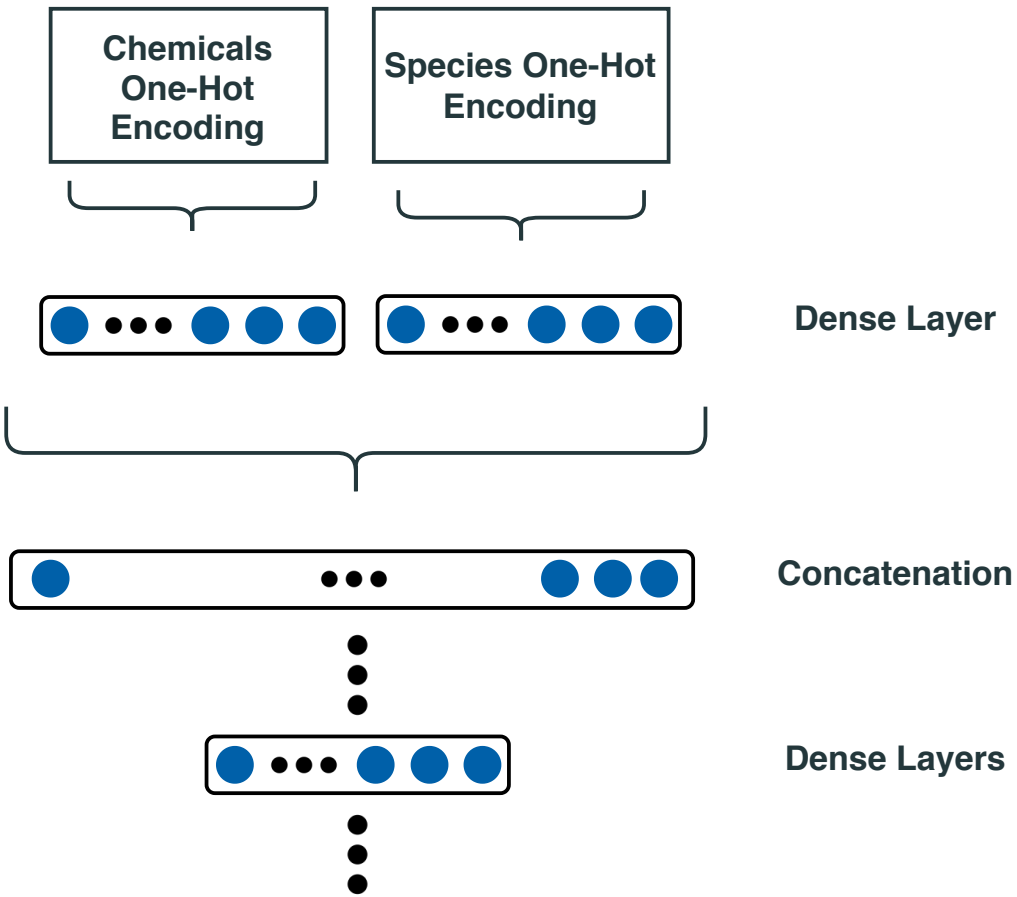
Multi-layer perceptron (MLP)



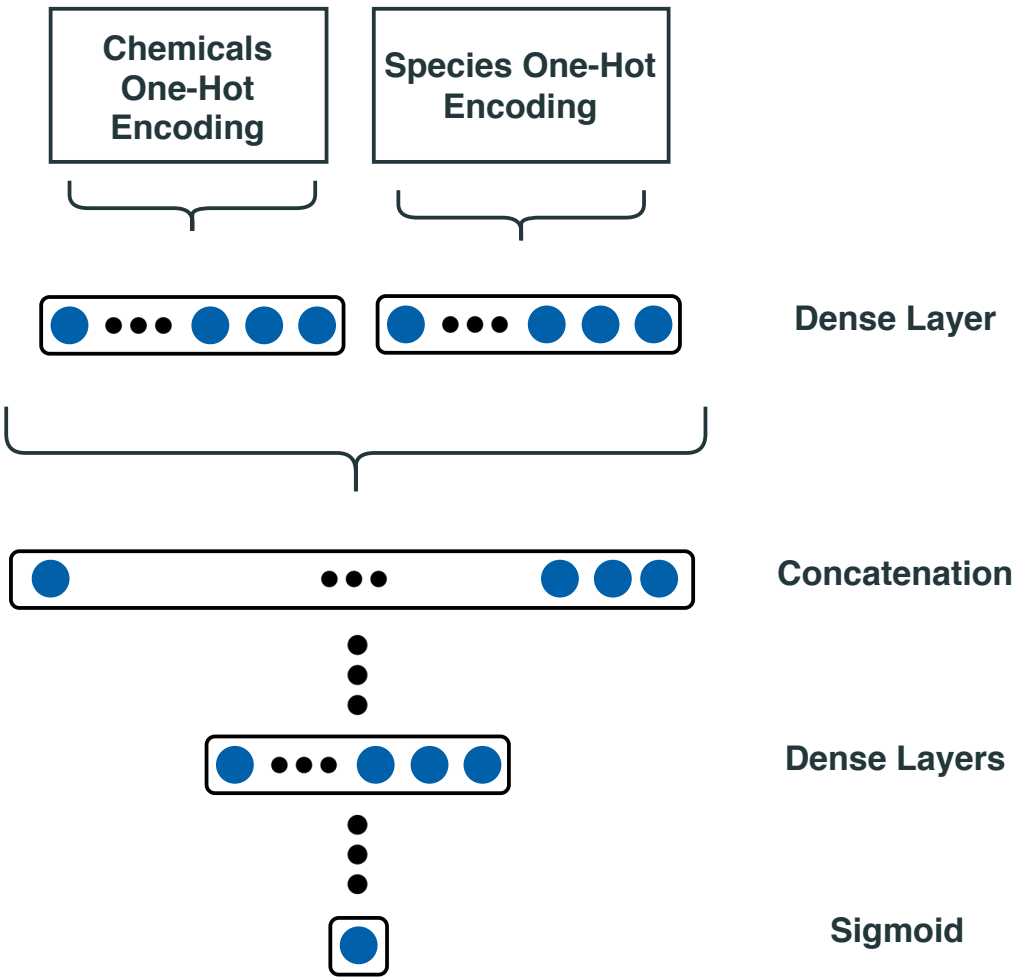
Multi-layer perceptron (MLP)



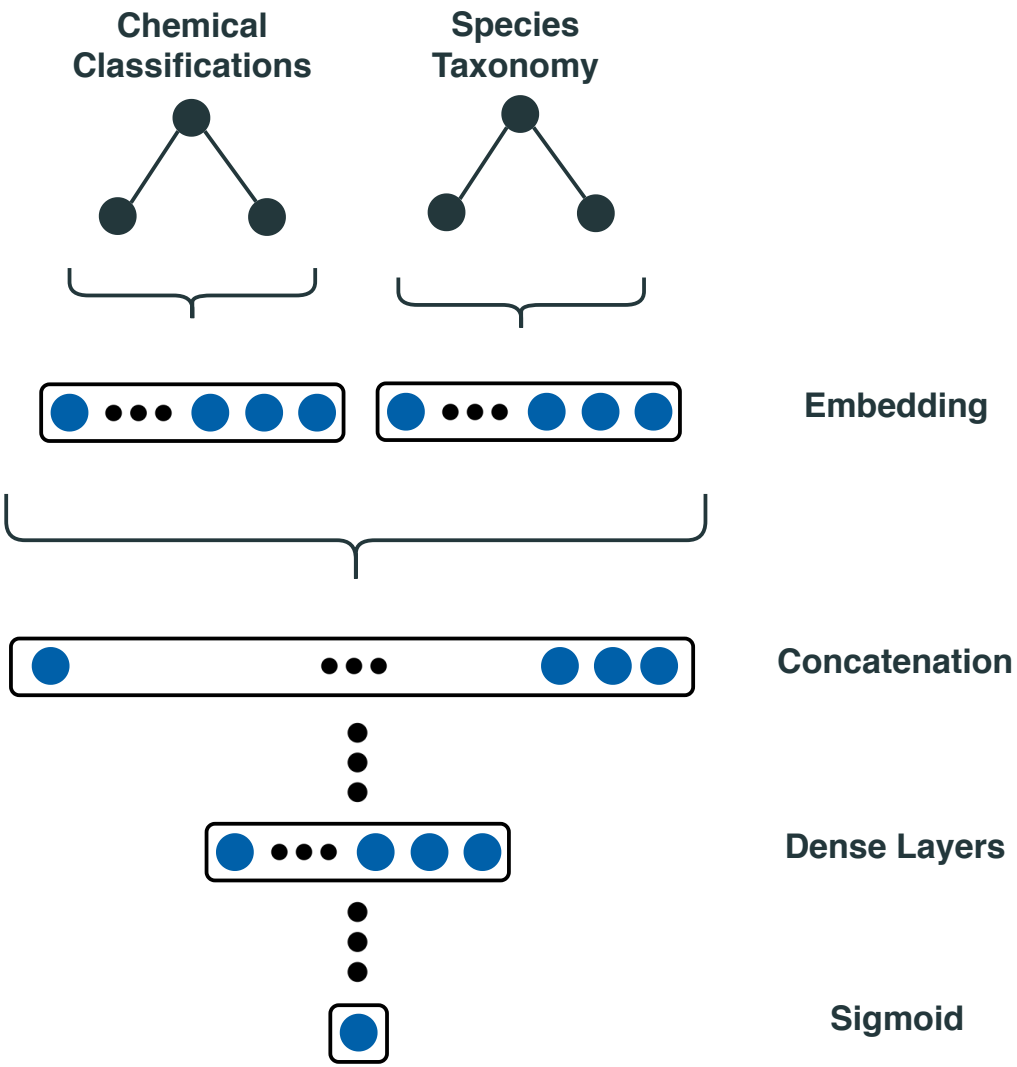
Multi-layer perceptron (MLP)



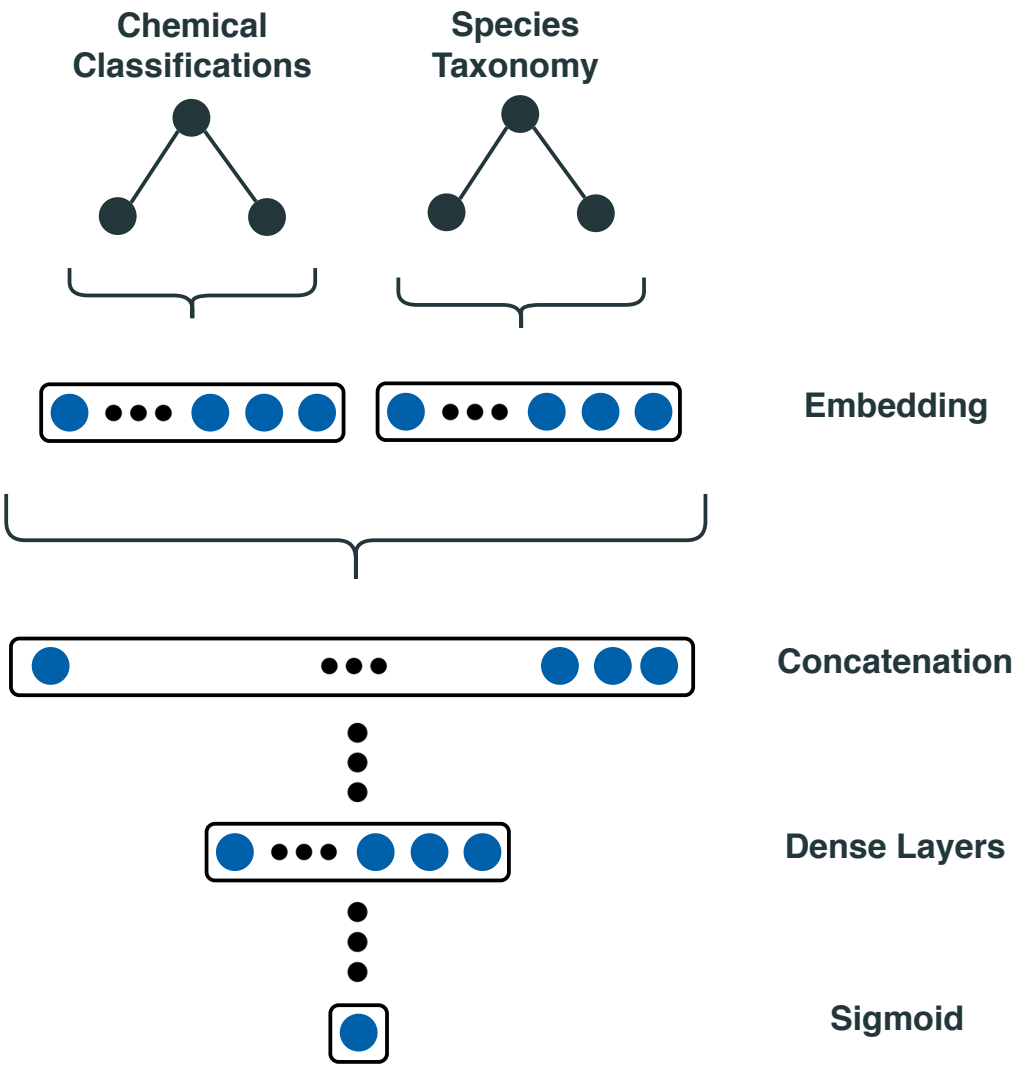
Multi-layer perceptron (MLP)



KG embedding + MLP



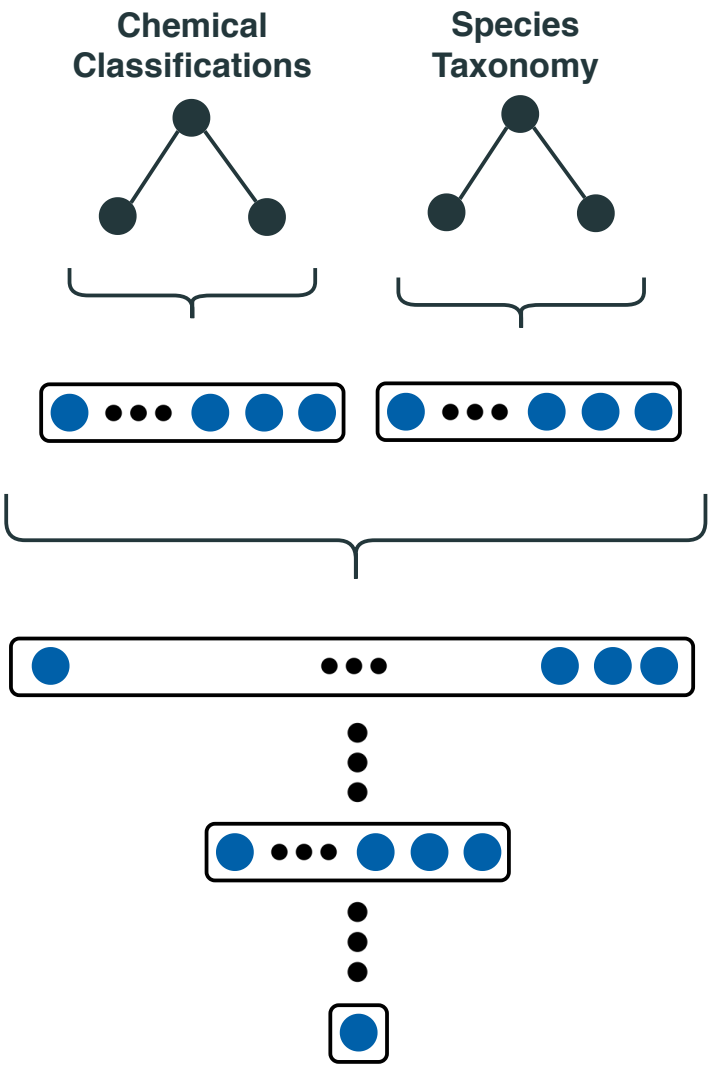
KG embedding + MLP



Three embedding models:

1. TransE
2. DistMult
3. HolE

KG embedding + MLP



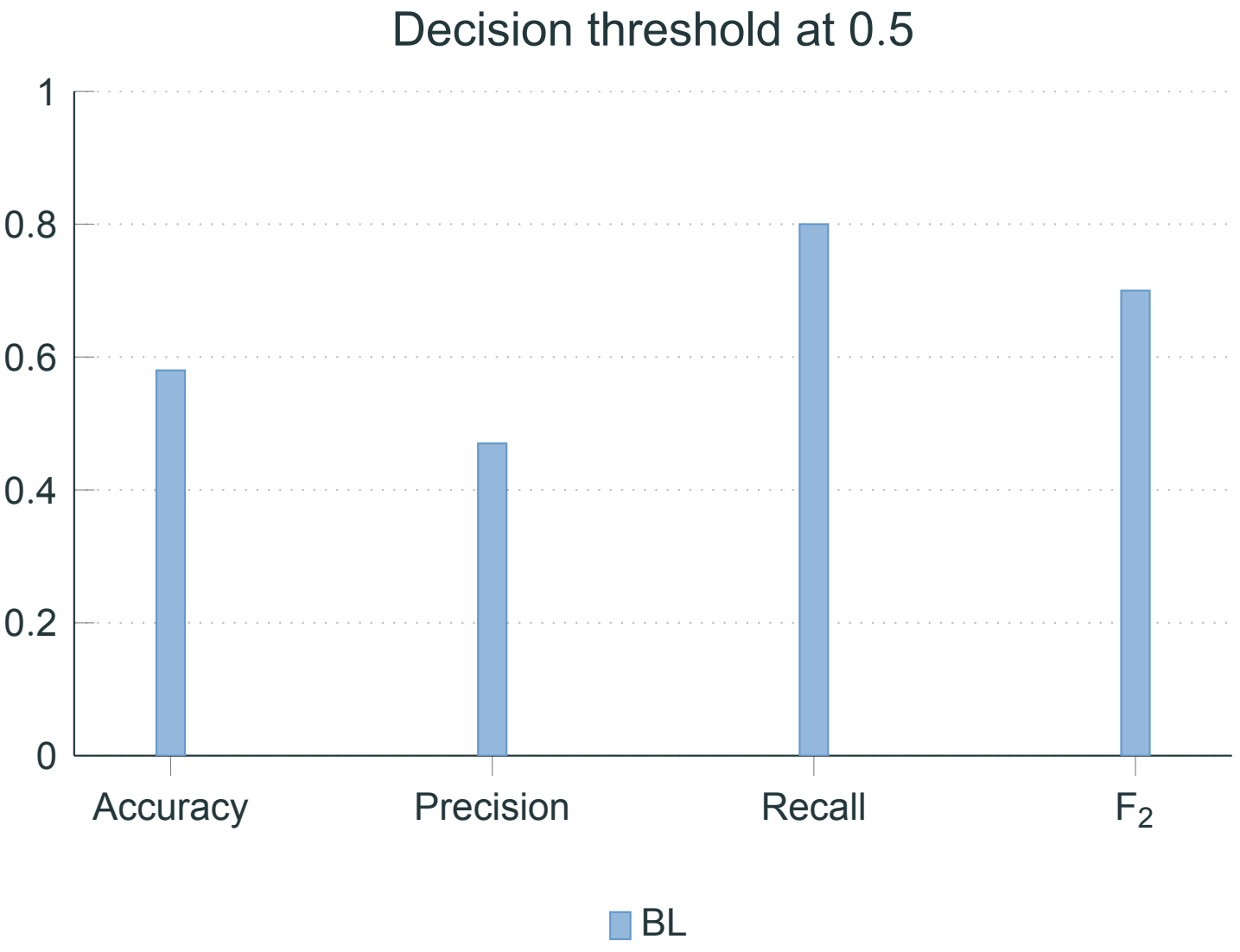
Three embedding models:

1. TransE
2. DistMult
3. HolE

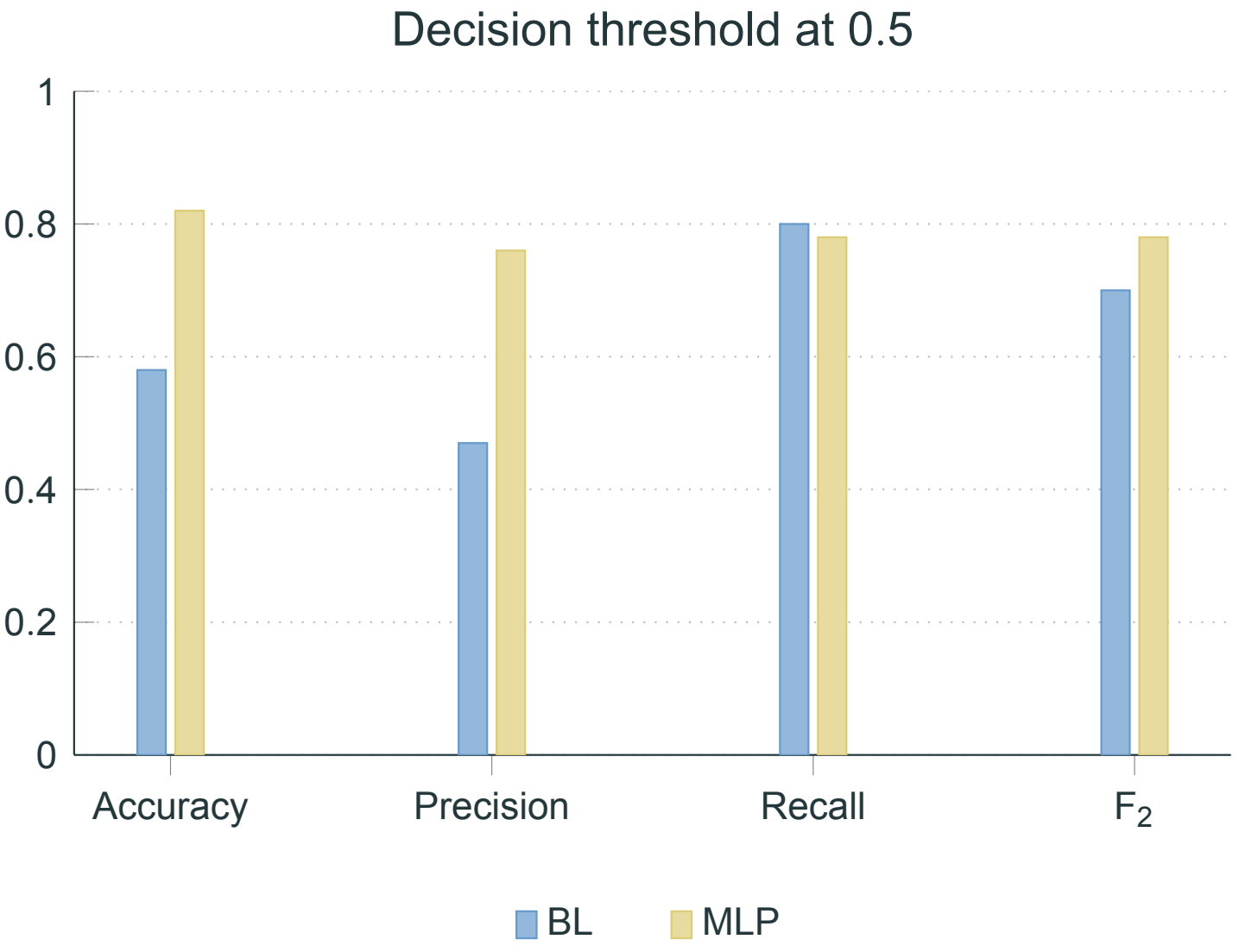
Optimization:

Simultaneous optimization of prediction and embedding models.

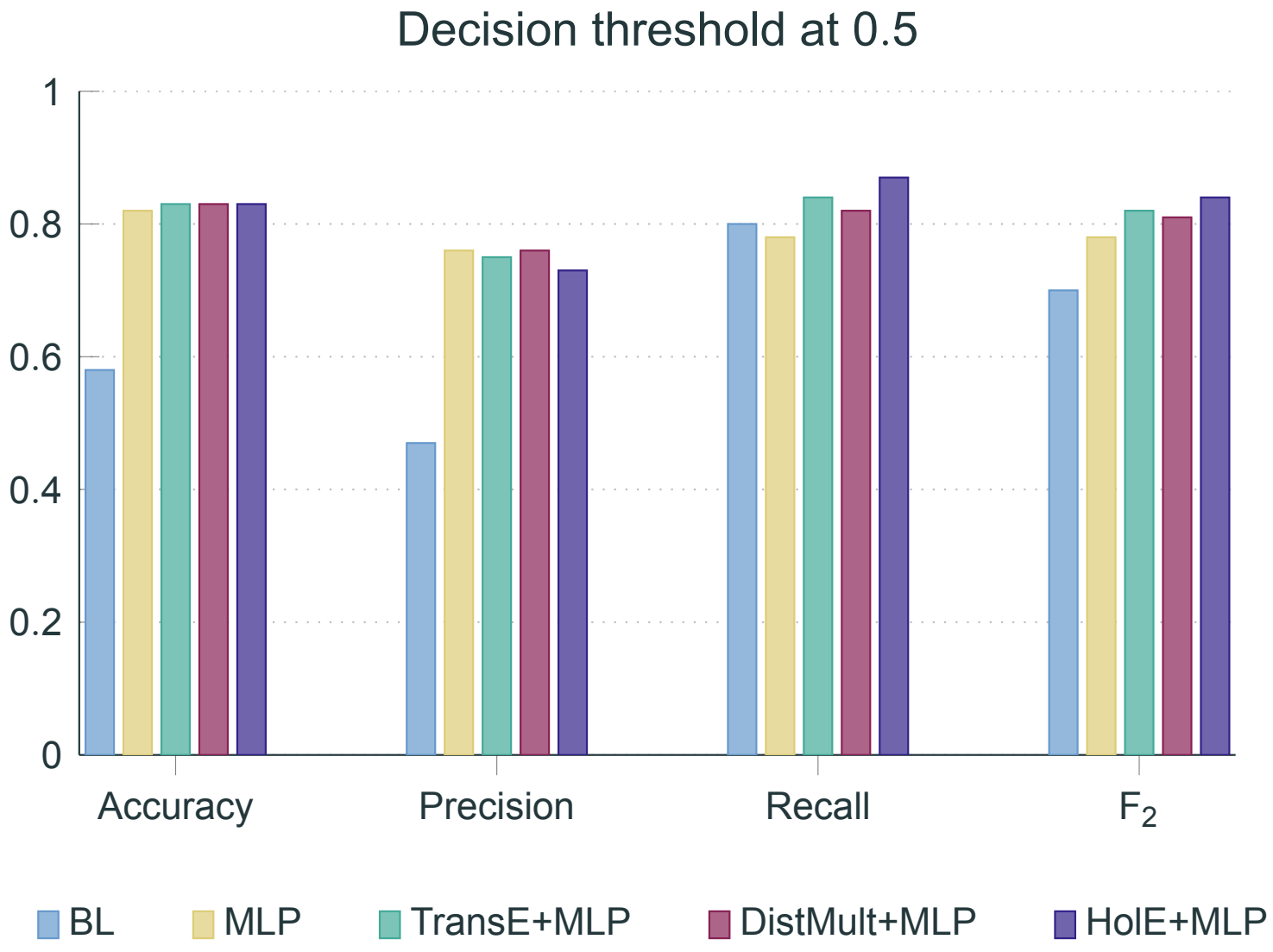
Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL

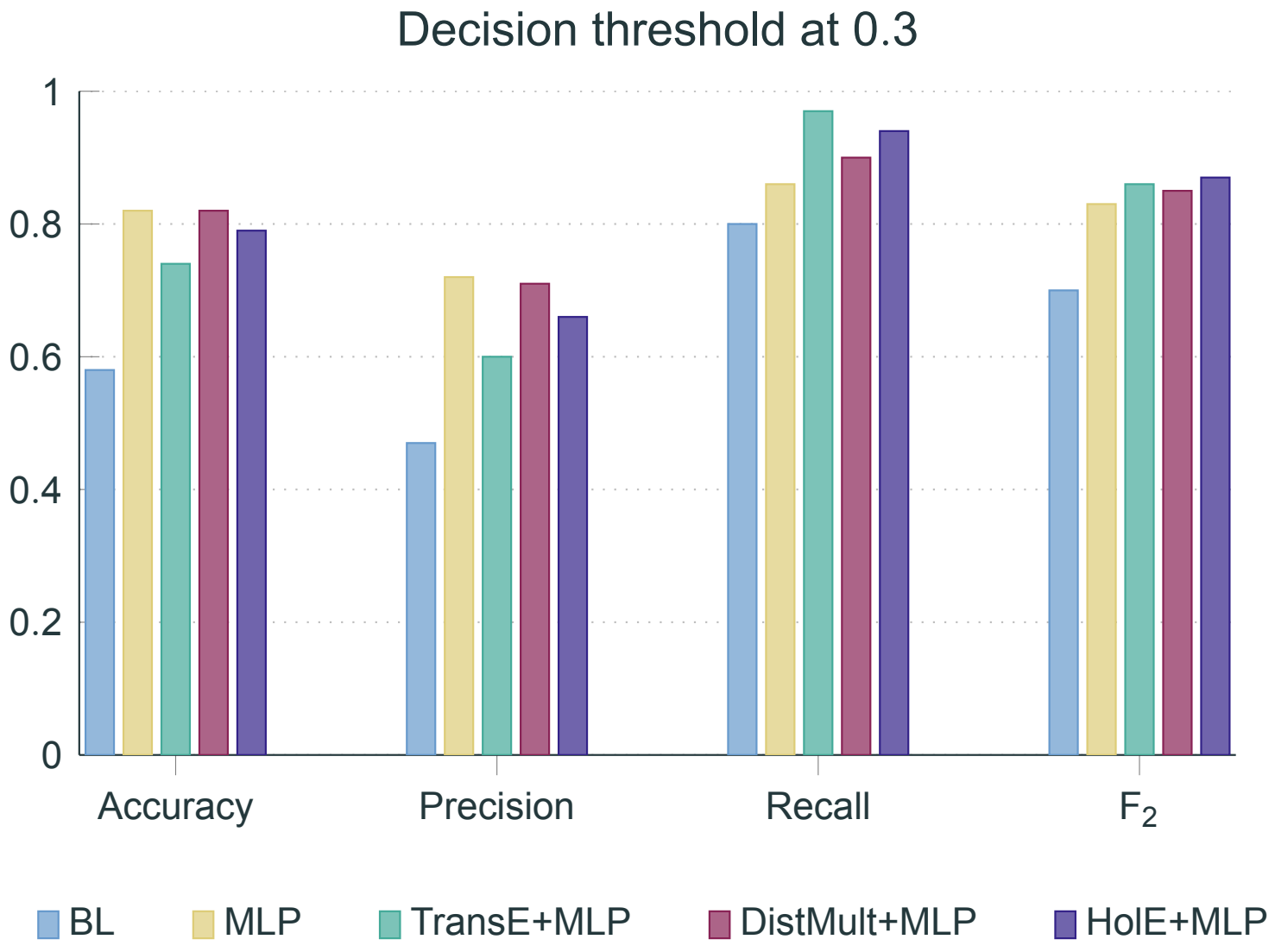
Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL
MLP

Results


Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL
MLP
TransE+MLP
DistMult+MLP
HolE+MLP

Results


Decision threshold at 0.3
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL
MLP
TransE+MLP
DistMult+MLP
HolE+MLP

Summary and Future Work

☑ Improved data access using TERA KG.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

- ☐ Explore the use of more sophisticated models

Summary and Future Work

- ☑ Improved data access using TERA KG.
- ☑ Introducing background knowledge in form of a KG improved the prediction results.

- ☐ Explore the use of more sophisticated models
- ☐ Move from binary labels to chemical concentrations.

Erik B. Myklebust, Ernesto Jimenez-Ruiz, Jiaoyan Chen, Raoul Wolf, Knut Erik Tollefsen. ***Knowledge Graph Embedding for Ecotoxicological Effect Prediction.***

Erik B. Myklebust, Ernesto Jimenez-Ruiz, Jiaoyan Chen, Raoul Wolf, Knut Erik Tollefsen. ***TERA: the Toxicological Effect and Risk Assessment Knowledge Graph.***

ErikBMyklebust

ebm@niva.no