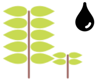
Knowledge Graph Embedding for Ecotoxicological Effect Prediction

Erik B. Myklebust[1,2], Ernesto Jimenez-Ruiz[2,3,4], Jiaoyan Chen[5], Raoul Wolf[1], & Knut Erik Tollefsen[1]

[1] Norwegian Institute for Water Research, Oslo, Norway
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Alan Turing Institute, London, United Kingdom
[4] City, University of London, London, United Kingdom
[5] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ErikBMyklebust
ebm@niva.no

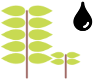
Ecological Risk Assessment



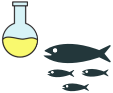
Exposure

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

Ecological Risk Assessment



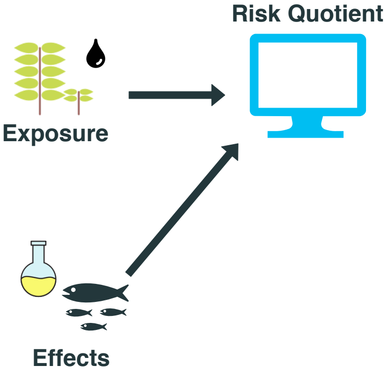
Exposure



Effects

Effect concentrations are found using organism experiments.

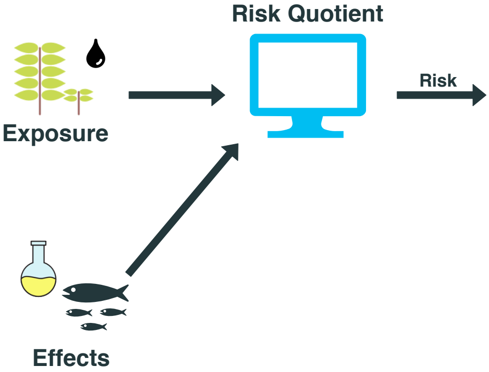
Ecological Risk Assessment



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

RQs coverage is limited by effect concentration experiments.

Ecological Risk Assessment

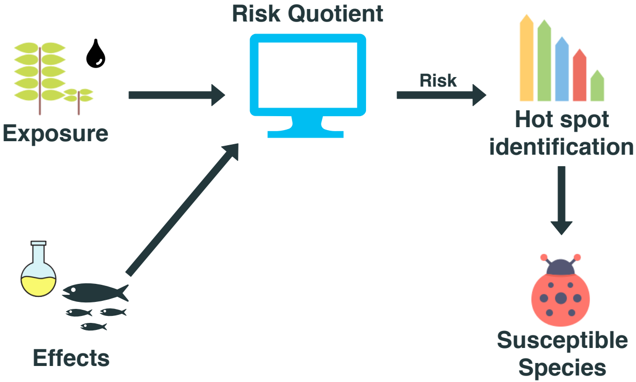


$$\text{risk}_{\text{group}} \approx \sum^{\textit{chemicals}} RQ$$

Risk for a group of species.

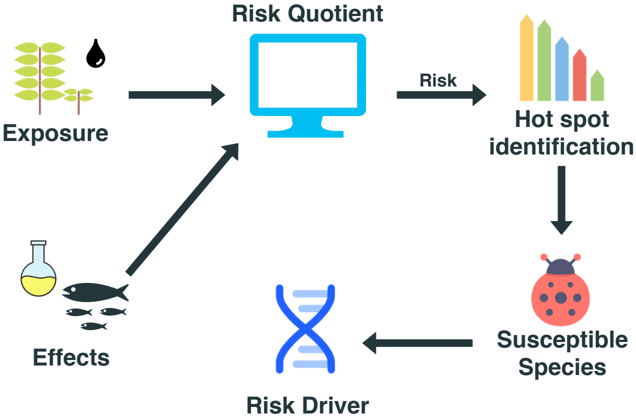
The group can contain all species in the ecosystem.

Ecological Risk Assessment



The risk is used to find further susceptible species.

Ecological Risk Assessment



Risk Quotient

Exposure

Risk

Hot spot identification

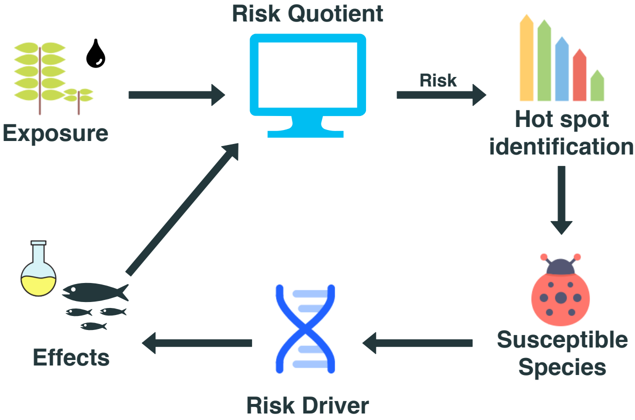
Effects

Susceptible Species

Risk Driver

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

Ecological Risk Assessment

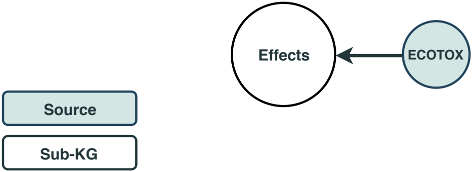


New effect hypotheses are then tested in the laboratory.

The TERA Knowledge Graph

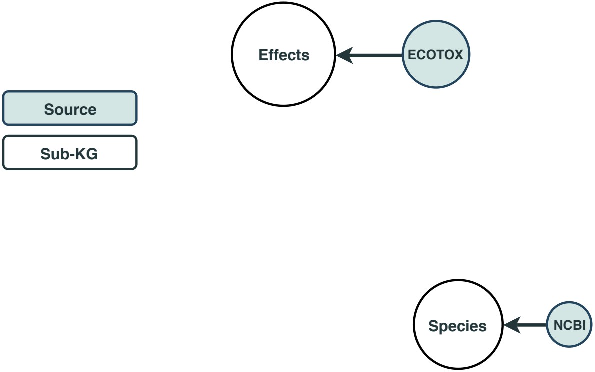
The Toxicological and Risk Assessment (TERA) knowledge graph integrates data sources varying in format.

The TERA Knowledge Graph



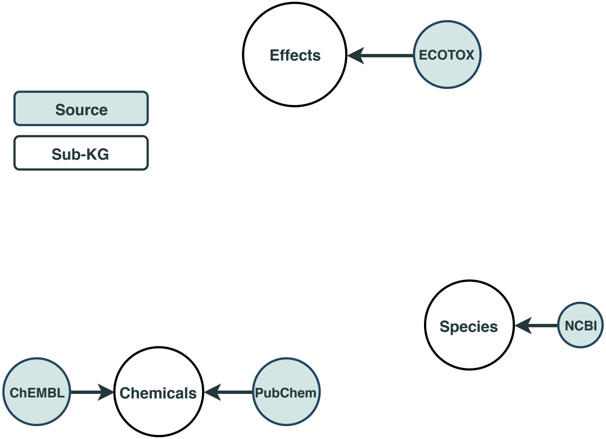
ECOTOX is the largest (public) source of effect data.

The TERA Knowledge Graph



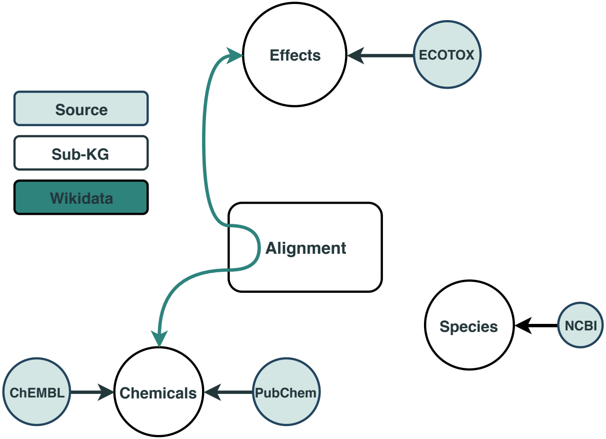
NCBI’s tabular taxonomy is converted to a hierarchy.

The TERA Knowledge Graph



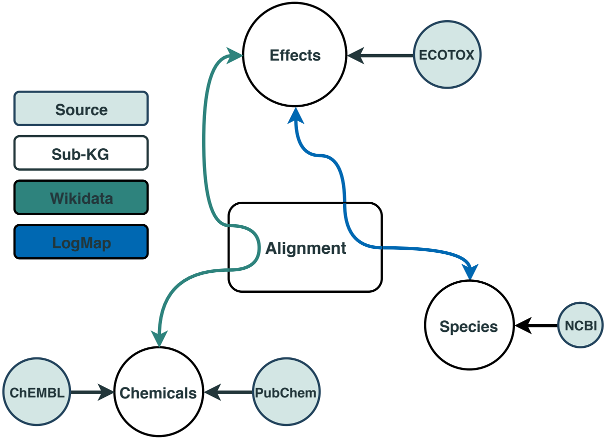
Importing the ChEMBL and PubChem knowledge graph.

The TERA Knowledge Graph



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

The TERA Knowledge Graph



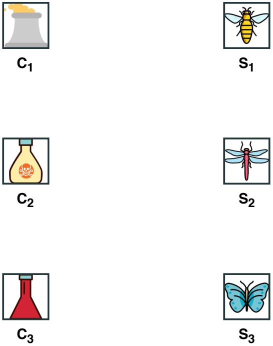
Aligning taxonomies using ontology alignment tool LogMap.

Effect Prediction Problem Definition

Effect Prediction Problem Definition

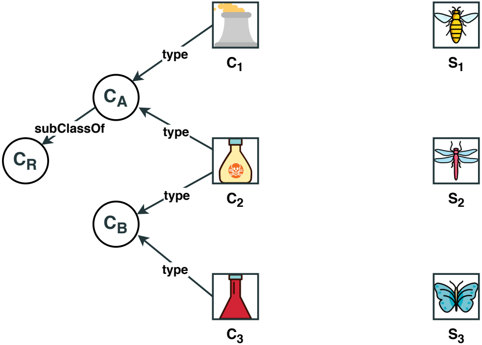
Chemicals

Effect Prediction Problem Definition



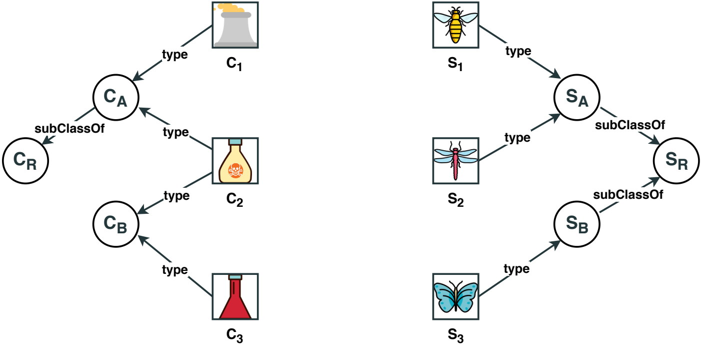
Species

Effect Prediction Problem Definition



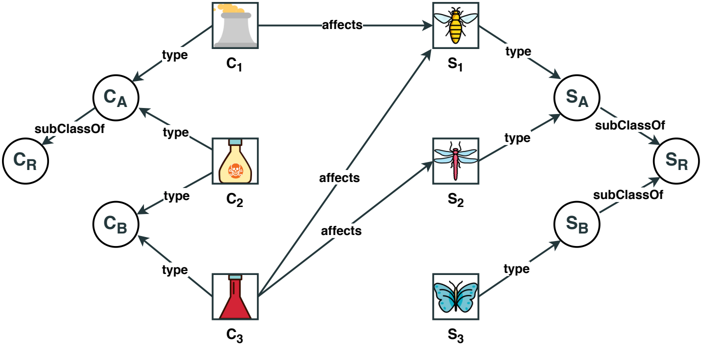
Chemical classification

Effect Prediction Problem Definition



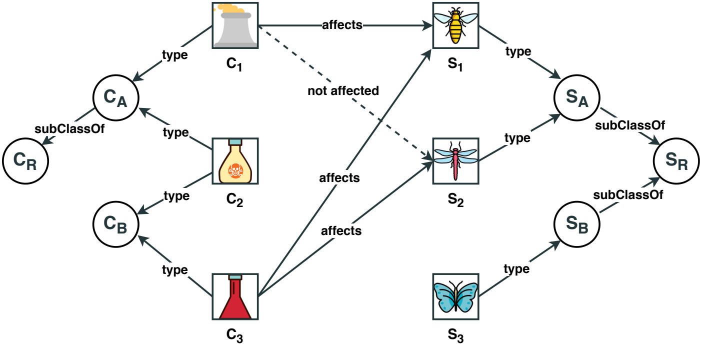
Taxonomy

Effect Prediction Problem Definition



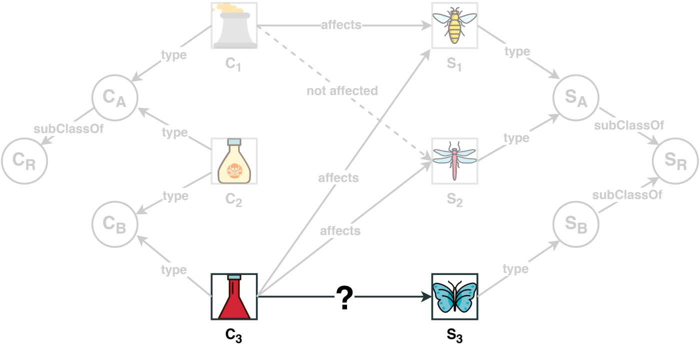
Positive samples

Effect Prediction Problem Definition



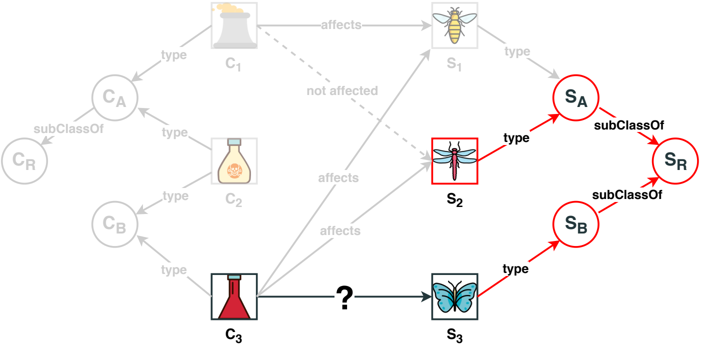
Negative samples

Taxonomic Distance Model - Baseline (BL)



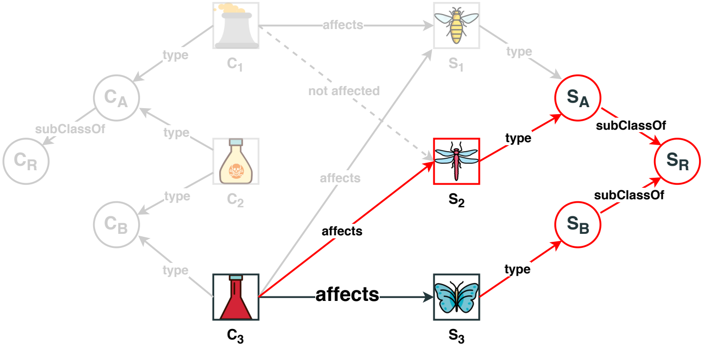
Does C_3 affect S_3?

Taxonomic Distance Model - Baseline (BL)



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

Taxonomic Distance Model - Baseline (BL)



Yes, C_3 affects S_3

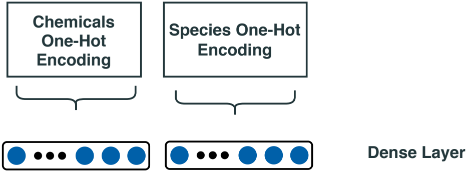
Multi-layer perceptron (MLP)

Multi-layer perceptron (MLP)

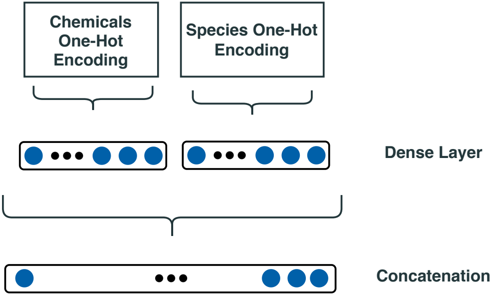
Chemicals One-Hot Encoding

Species One-Hot Encoding

Multi-layer perceptron (MLP)



Multi-layer perceptron (MLP)



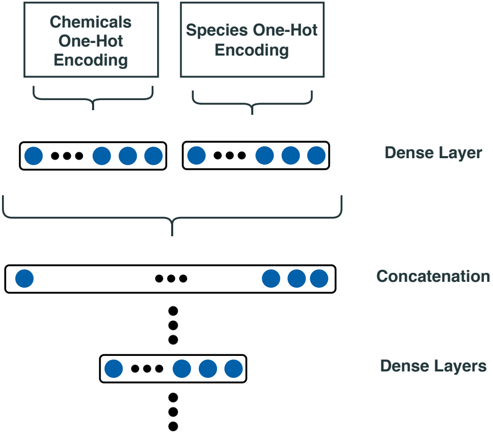
Chemicals One-Hot Encoding

Species One-Hot Encoding

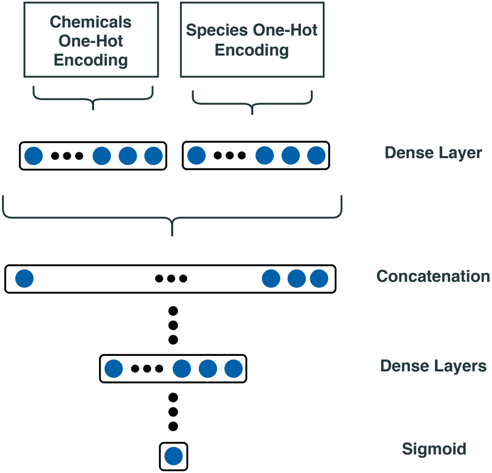
Dense Layer

Concatenation

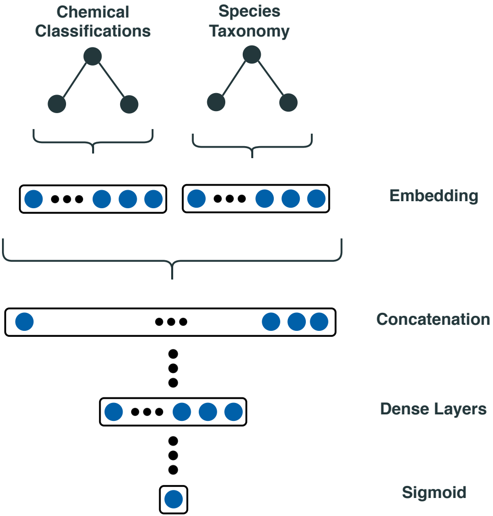
Multi-layer perceptron (MLP)



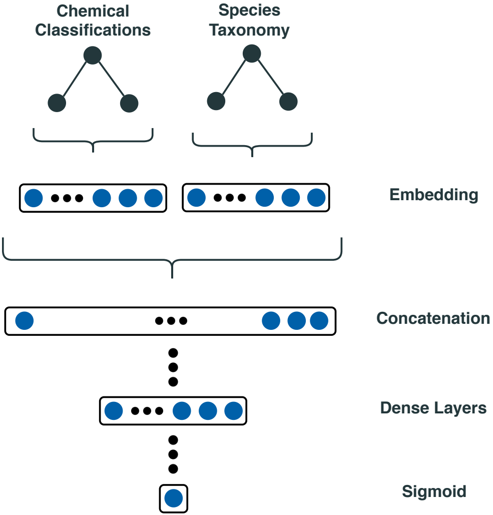
Multi-layer perceptron (MLP)



KG embedding + MLP



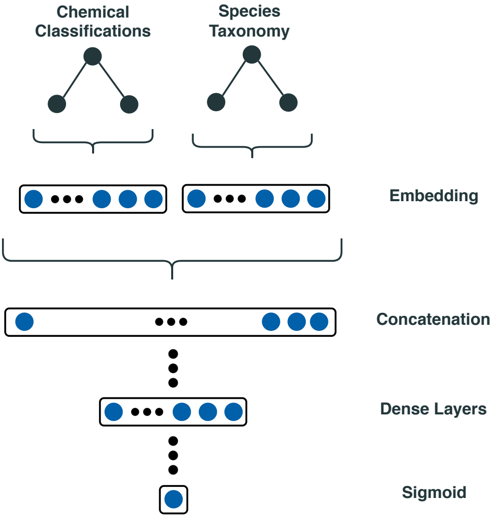
KG embedding + MLP



Three embedding models:

1. TransE
2. DistMult
3. HolE

KG embedding + MLP



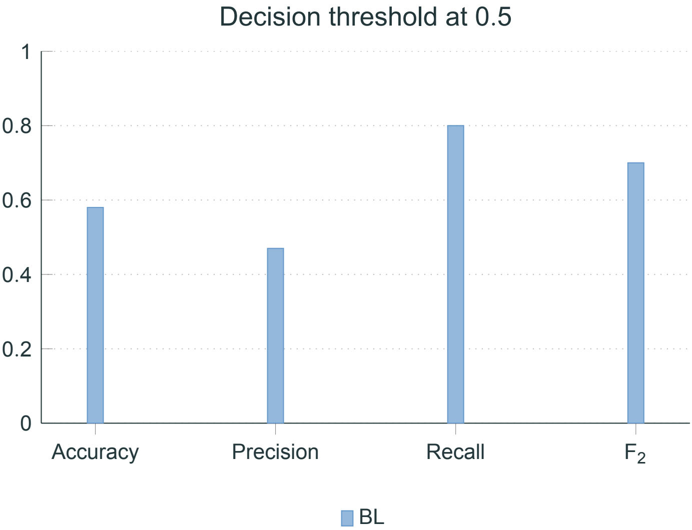
Three embedding models:

1. TransE
2. DistMult
3. HolE

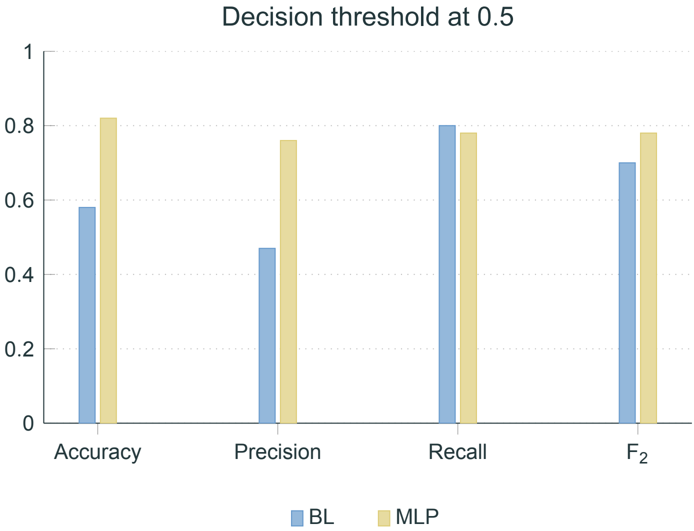
Optimization:

Simultaneous optimization of prediction and embedding models.

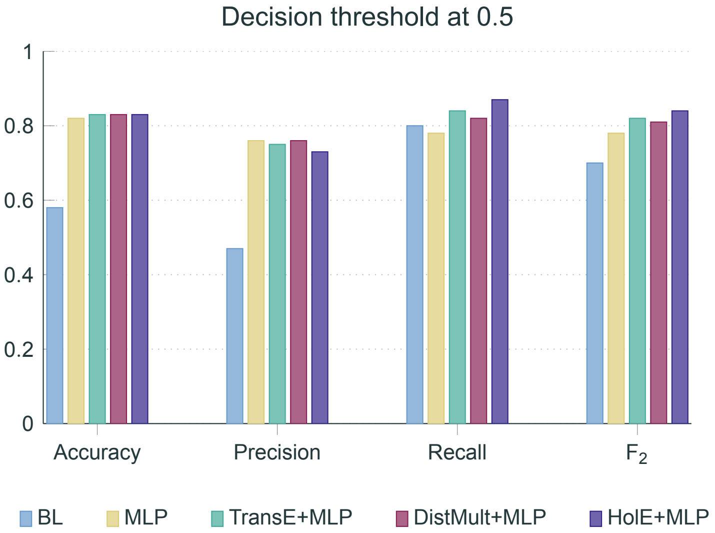
Results


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

Results

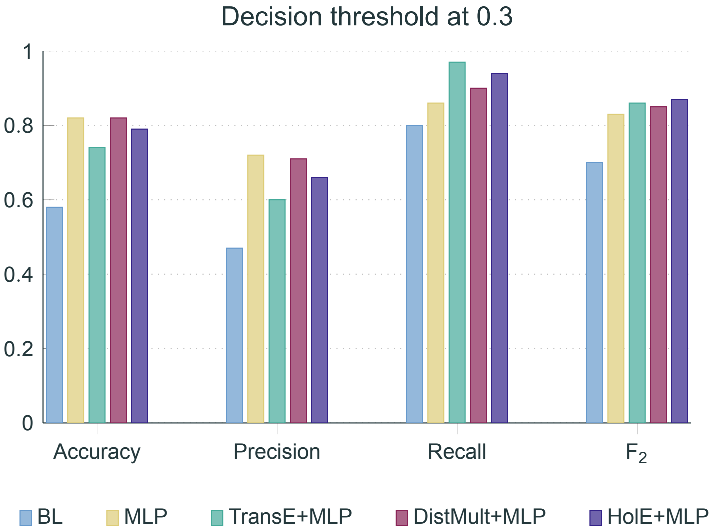

Decision threshold at 0.5
1
0.8
0.6
0.4
0.2
0
Accuracy
Precision
Recall
F_2
BL
MLP

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.

☐ Expand the TERA knowledge graph with other relevant data, *e*.*g*., habitat.

Summary and Future Work

☑ Improved data access using TERA KG.

☑ Introducing background knowledge in form of a KG improved the prediction results.

□ Expand the TERA knowledge graph with other relevant data, *e*.*g*., habitat.

□ 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.

☐ Expand the TERA knowledge graph with other relevant data, *e*.*g*., habitat.

☐ Explore the use of more sophisticated models

☐ Move from binary labels to chemical concentrations.

Thank you.

Please visit us at poster 466.

Questions?

Acknowledgements:

Forskningsrådet
The Research Council of Norway

272414 & 268294 & 237889

EPSRC
Pioneering research and skills

EP/N510129/1 & DBOnto & MaSI^3 & ED^3

NCTP at NIVA & Martin Giese & Zofia C. Rudjord

🐦 ErikBMyklebust

✉ ebm@niva.no