Ecotoxicological Effect Prediction Using Knowledge Graph Embedding

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

[1]Norwegian Institute for Water Research (NIVA), Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

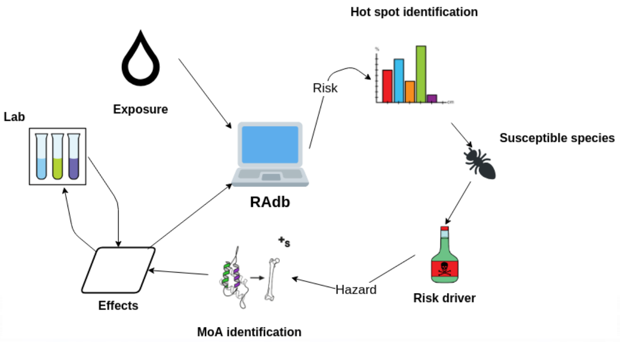
[3]Alan Turing Institute, London, United Kingdom

[4]Department of Computer Science, University of Oxford, United Kingdom

SETAC Helsinki, May 27th, 2019

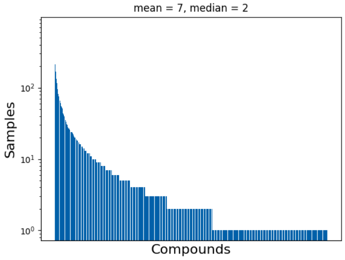
contact: ebm@niva.no

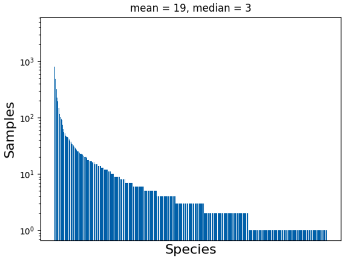
Background



Lack of sufficient effect data for one or more species are currently limiting hazard and risk assessment.

Research question





Needs

Develop robust approaches to perform gap filling using the available information across species and compounds.

Objective

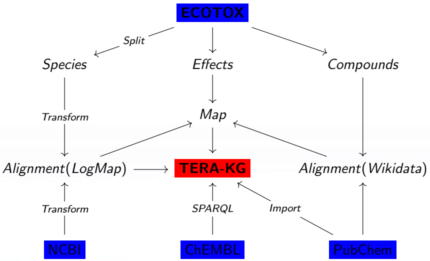
Complement traditional approaches (e.g. QSARs) for predicting ecotoxicological effects.

Proposed solution

i. Materialize a knowledge graph (KG) from disparate sources.

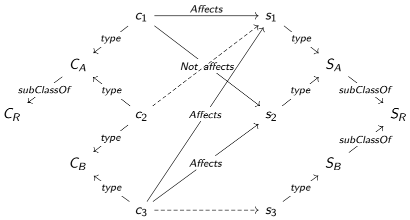
ii. Embed the KG with well proven models (embedding $\Leftrightarrow$ vector space representation).

iii. Embeddings are used to train a model.



Data sources in blue. Details found in [1].

Prediction problem



Gap filling problem. Solid known and dashed unknown.

The *normalized* effect of c_i on s_j is modeled as a function

$$f \colon C, S \mapsto (0, 1) \subset \mathbb{R}$$

$C \equiv$ all entities in compound hierarchy, $S \equiv$ all entities in taxonomy.

Models

The objective of the models is to learn function f. We compare three models with varying complexity.

Baseline

A compound-species pair inherit effects from the *most similar* compound-species pairs.

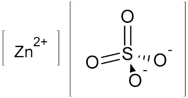
Multilayer perceptron

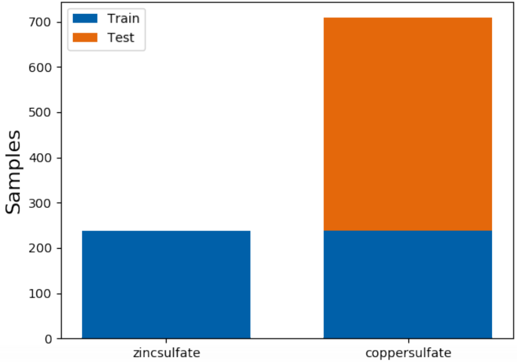
Learning strictly from effect data.

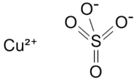
Knowledge graph embedding and multilayer perceptron

Learning embeddings from knowledge graph and using these to train a multilayer perceptron model.

Example







Two settings:

i. Only samples for `zincsulfate` and `coppersulfate`.

ii. Use all available additional data to train models.

Example results

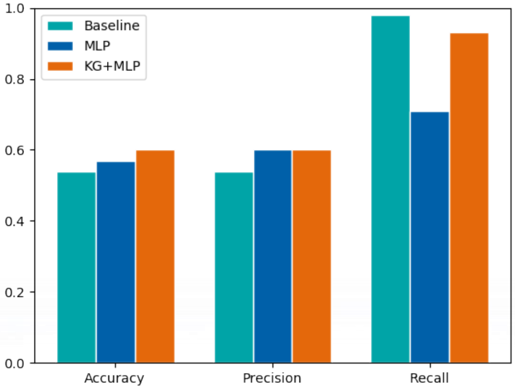


Figure: Setting *i.* results. Higher is better.

Example results

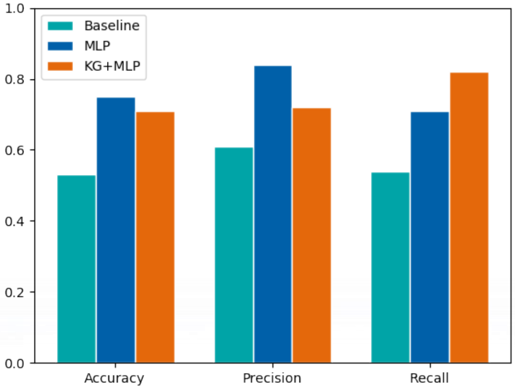


Figure: Setting *ii.* results.

General results

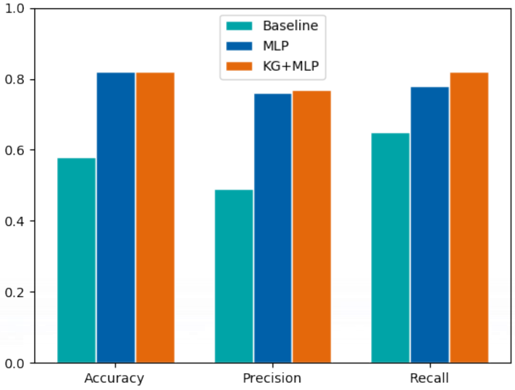
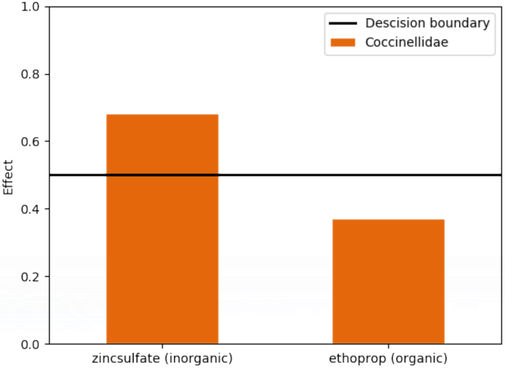
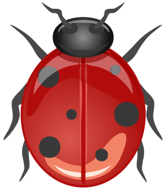


Figure: Random train/test sets (0.7/0.3 split).

Abstraction





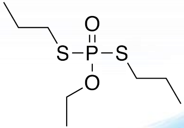


Figure: Effects on family of ladybugs (Coccinellidae).

Conclusion

- Models with background knowledge generally predicts effects with higher metrics.
- We have to sacrifice precision to improve recall, but not linearly proportional.
- The knowledge graph enables prediction on higher taxonomic levels.
- Near future work: Integrate the knowledge graph and prediction models with NIVA's risk assessment system.

Thank you for listening.

contact: `ebm@niva.no`

Acknowledgements

This work is supported by the grant 272414 from the Research Council of Norway (RCN), the MixRisk project (RCN 268294), the AIDA project (UK Government's Defence & Security Programme in support of the Alan Turing Institute), the SIRIUS Centre for Scalable Data Access (RCN 237889), the Royal Society, EPSRC projects DBOnto, $MaSI^3$ and ED^3. We would also like to thank Martin Giese and Zofia C. Rudjord for their contribution in early stages of this project. This work is organized under NCTP (`www.niva.no/nctp`).

Related work

[1] Myklebust, E. B., Jimenez-Ruiz, E., Chen, J., Wolf, R., Tollefsen, K. E.
Knowledge Graph Embedding for Ecotoxicological Effect Prediction
Submitted, 2019. Contact `ebm@niva.no` for preprint.

[2] Additional Resources
https://github.com/Erik-BM/NIVAUC