KNOWLEDGE GRAPH EMBEDDING FOR ECOTOXICOLOGICAL EFFECT PREDICTION

Erik B. Myklebust[1,2], Ernesto Jiménez-Ruiz[2,3,4], Jiaoyan Chen[5], Raoul Wolf[1] & Knut Erik Tollefsen[1]
[1]Norwegian Institute for Water Research (NIVA), Oslo, Norway
[2]Department of Informatics, University of Oslo, Oslo, Norway
[3]The 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

Introduction

Ecological risk assessment is the process for evaluating how likely it is that the environment may be impacted by exposure to chemical stressors. The assessments compares effect concentrations and environmental concentrations to estimate cumulative risk. The effect concentrations are gathered through extensive laboratory experiments. Here, we present work that aims at limiting the search space in terms of both chemicals and species, in these experiments. Recall is the preferred metric since risk assessment should be protective of the environment. This poster relates to a paper by the same name, which is linked below (QR-code).

Models

In contrast to commonly used effect prediction models (*e.g.*, QSARs), we aim a developing models that cover the ecotoxicological domain in its entirety (*i.e.*, all species and chemicals). We have developed three models:

- A naive baseline (BL) model based on graph distances in the species taxonomy and chemical classifications graphs.
- A multilayer perceptron (MLP; *Fig. 1*) model using one-hot encodings of species and chemicals as inputs.
- A model replacing the input to the MLP with a knowledge graph embedding model (TransE, DistMult, HolE; *Fig. 2*).

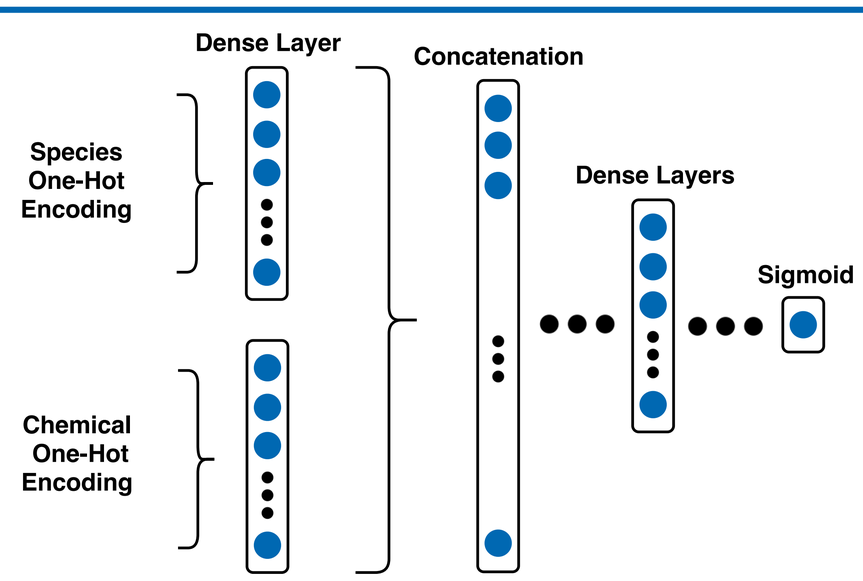

Fig. 1. Multilayer perceptron with one-hot encoding input.

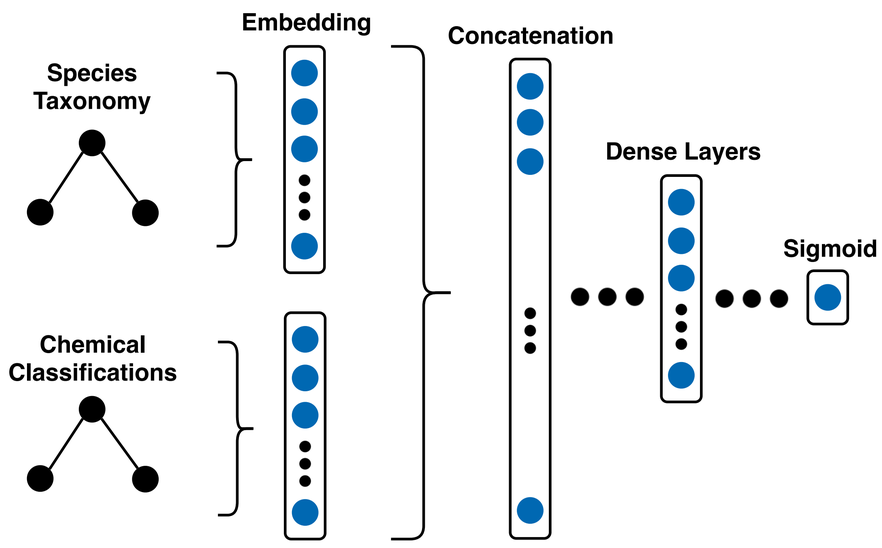

Fig. 2. Multilayer perceptron with knowledge graph embedding as input.

Knowledge Graph

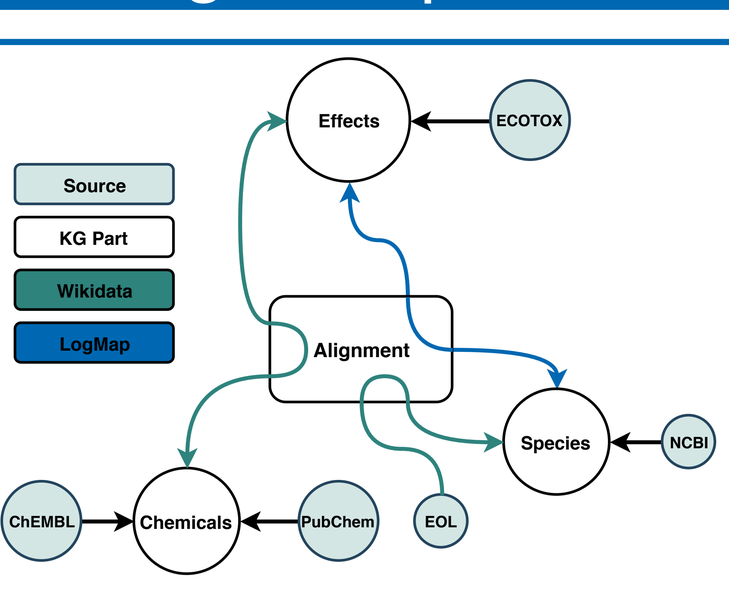

Fig. 3. The Toxicological and Risk Assessment (TERA) knowledge graph.

TERA (*Fig. 3*) is constructed from various disparate datasets. These include tabular (U.S. EPA ECOTOX, NIH NCBI Taxonomy, Encyclopedia of Life (EOL)) and RDF data (NIH PubChem), as well as SPARQL endpoints (EMBL ChEMBL). The sources are aligned using LogMap and Wikidata.

Results

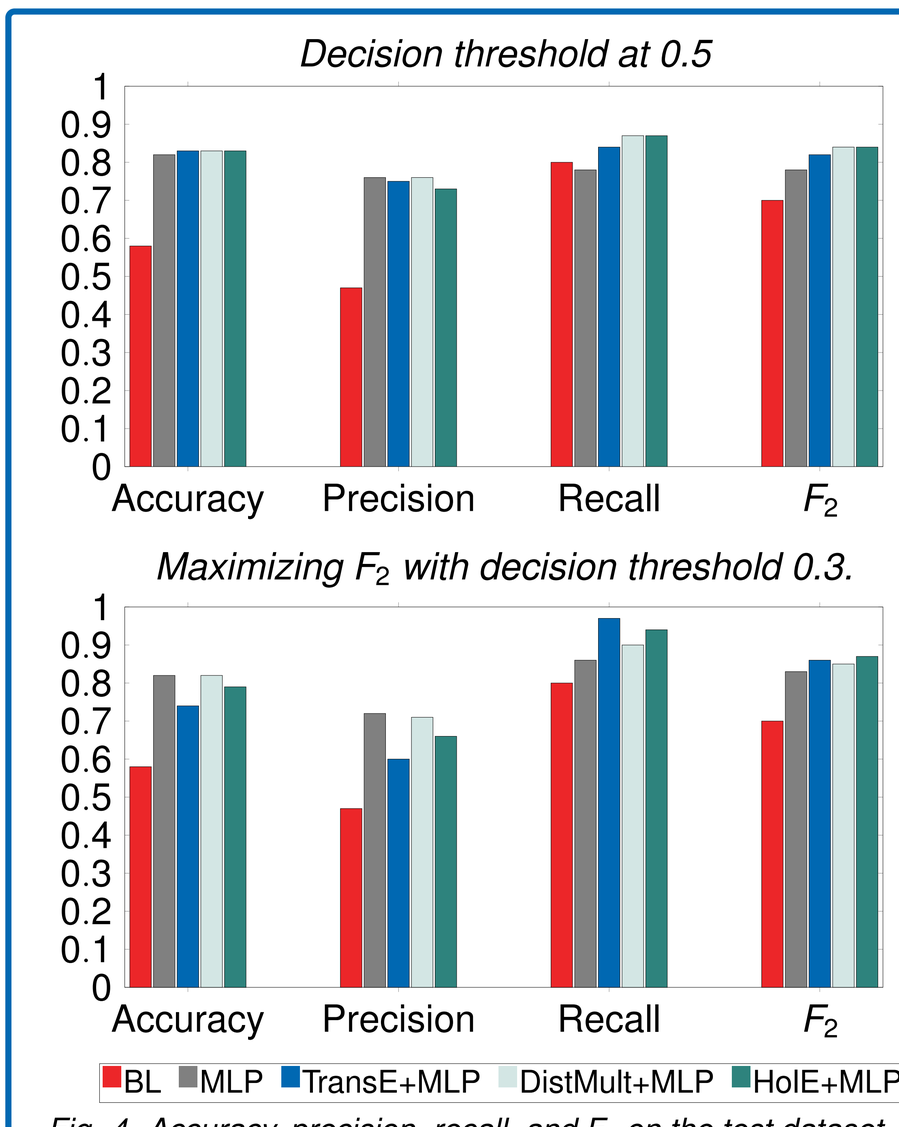

Fig. 4. Accuracy, precision, recall, and F_2 on the test dataset. F_2 gives twice the weight to recall over precision.

Conclusion

The TERA knowledge graph is an important contribution to the ecotoxicological community, as it incorporates a large amount of relevant data. The obtained predictions are promising and show the benefits of the TERA knowledge graph.
The ultimate goal of this work is integration into the ecological risk assessment pipeline, where the contributions will expand and diversify the assessments. The datasets, evaluation results, documentation and source codes are available from the following GitHub repository: https://github.com/Erik-BM/NIVAUC. Contact via email: ebm@niva.no.