A KNOWLEDGE GRAPH FOR ECOTOXICOLOGICAL RISK ASSESSMENT AND EFFECT PREDICTION

Erik B. Myklebust
Norwegian Institute for Water Research (NIVA), Oslo, Norway
Department of Informatics, University of Oslo, Oslo, Norway
ebm@niva.no

Funding by Research Council of Norway (272414).
Supervised by Ernesto Jiménez-Ruiz (U. of Oslo; City, U. of London), Jiaoyan Chen (U. of Oxford), Raoul Wolf (NIVA), Knut Erik Tollefsen (NIVA; NMBU) & Martin Giese (U. of Oslo).

Introduction

Ecological risk assessment is the process of 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. To limit the use of test organisms and reduce experimental effort this project aims at:

1. Constructing a knowledge graph by gathering and integrating the relevant biological effect data and knowledge, to relieve the (domain) researchers of the manual work.
2. Using the knowledge graph together with machine learning techniques to predict effects of a compound on a species.

Task 2. is two-fold:

a. Limit the search space for laboratory experiments (binary prediction).
b. Predict effect concentrations (regression).

Note that, in Task *2a* recall is the preferred metric since ecological risk assessment should be protective of the environment.

Models

In contrast to commonly used effect prediction models (*e.g.*, QSARs), we aim at developing a conceptual model (*Figure 1*) that cover the ecotoxicological domain in its entirety (*i.e.*, all species and chemicals). In the preliminary work 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) model using one-hot encodings of species and chemicals.
- A model replacing the input to the MLP with a knowledge graph embedding model (*e.g.*, TransE, DistMult, HolE).

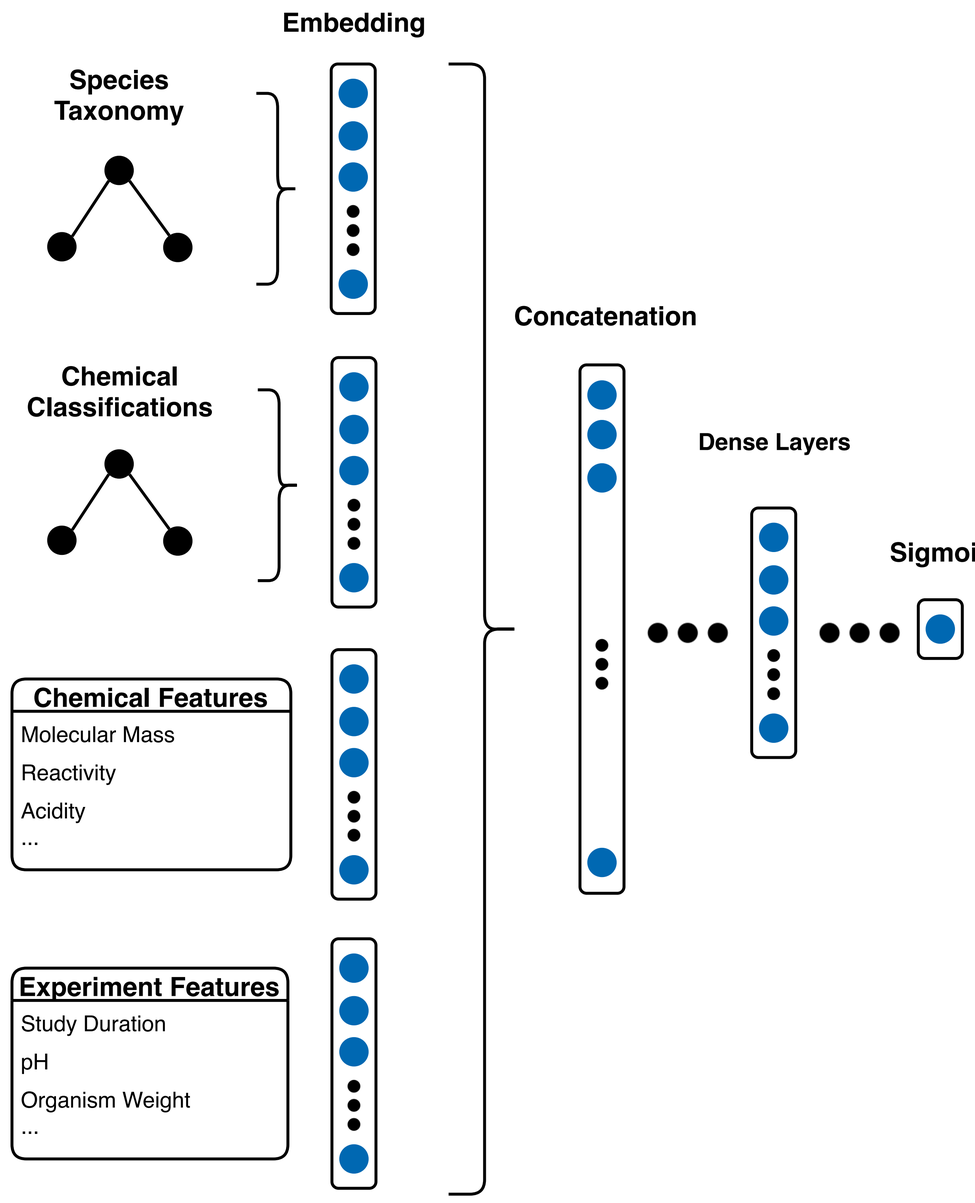

Figure 1. Conceptual model proposed for effect prediction.

Knowledge Graph

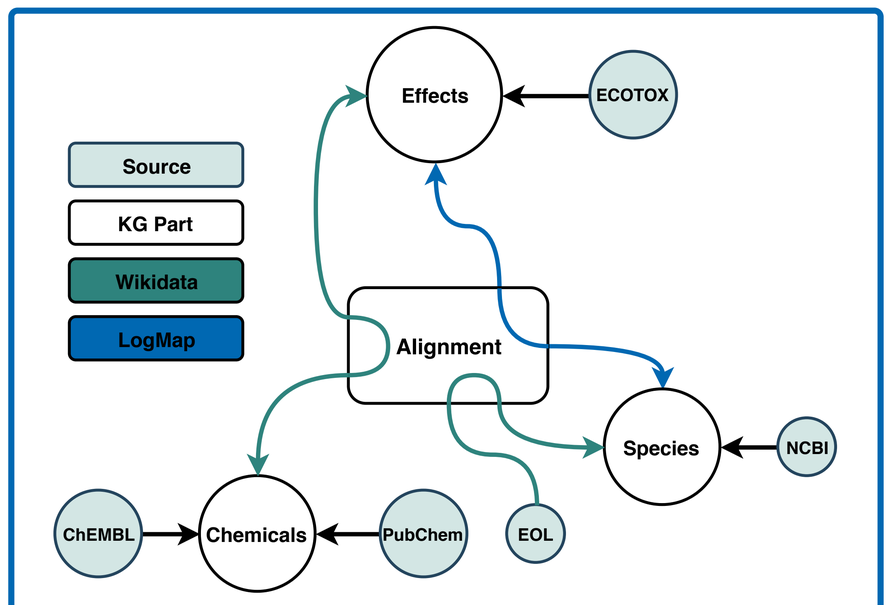

Figure 2. The Toxicological and Risk Assessment (TERA) knowledge graph.

TERA (*Figure 2*) 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.

Preliminary Results

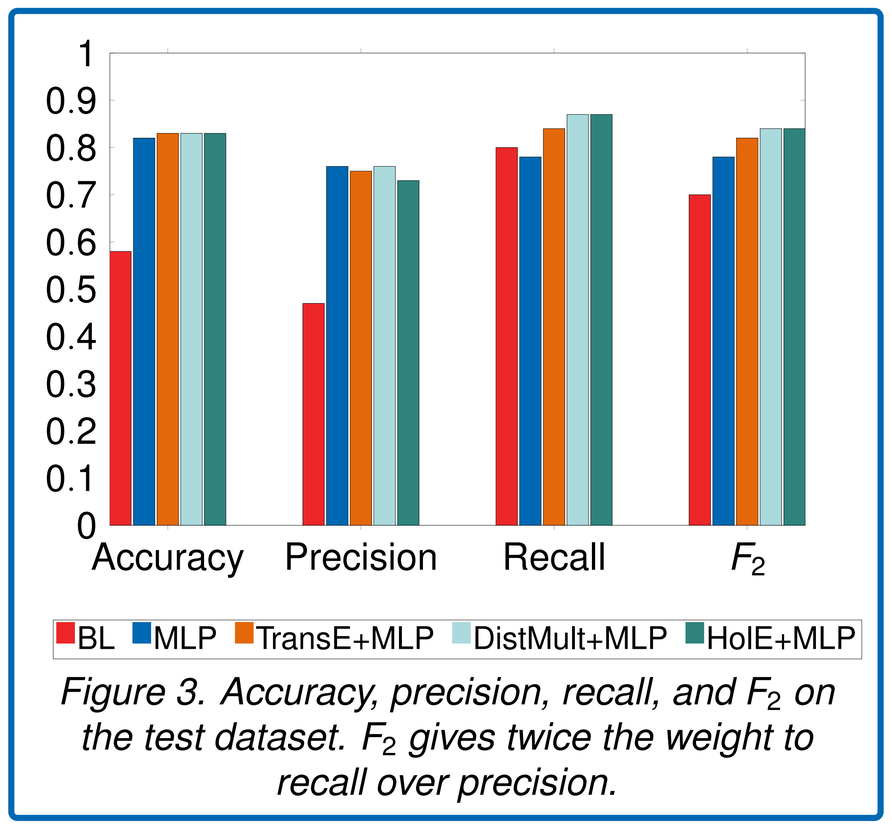

Figure 3. Accuracy, precision, recall, and F_2 on the test dataset. F_2 gives twice the weight to recall over precision.

Additional results in paper linked below (QR-code).

Future Work

1. Task *2b* has certain challenges, most notably, the inconsistent data, *e.g.*, effective concentrations of a chemical can vary by orders-of-magnitude from experiment to experiment.
2. The preliminary study also indicated that the chosen embedding models are sub-optimal for learning hierarchical structures. We will explore more sophisticated models, *e.g.*, Graph Convolution Networks. Moreover, we will develop/tune models for the specific hierarchical structure of the knowledge graph.

NIVA Norwegian Institute for Water Research
UiO : University of Oslo
SIRIUS