Domain Wide Effect Modelling to Support Read-Across in Hazard and Risk assessment

Erik Bryhn Myklebust[1,2], Ernesto Jimenez-Ruiz[2,3], Jiaoyan Chen[4],

Raoul Wolf[1], Knut Erik Tollefsen[1,5]

[1]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, 0349 Oslo, Norway
[2]Department of Informatics, University of Oslo, Gaustadalléen 23B, 0373 Oslo, Norway
[3]City, University of London, Northampton Square, Clerkenwell, London EC1V 0HB, United Kingdom
[4]Department of Computer Science, University of Oxford, 15 Parks Rd, Oxford OX1 3QD, United Kingdom
[5]Norwegian University of Life Sciences (NMBU), Universitetstunet 3, 1433 Ås, Norway
E-mail contact: ebm@niva.no

1. Introduction

Effect modelling is an important part of ecological risk assessment to provide environmental safety thresholds without excessive animal usage and resource-demanding experimental efforts. The current suite of QSAR models and other read-across approaches are efficient at filling data gaps in many cases but lack the coverage required to predict diverse toxicity mechanisms, species/taxa and endpoints relevant for ecological exposure scenarios. We propose the use of background knowledge about species and chemicals to improve the reach of read-across techniques. This is illustrated in Fig. 1, where the background hierarchical structures can improve the performance and the understanding of the model results.

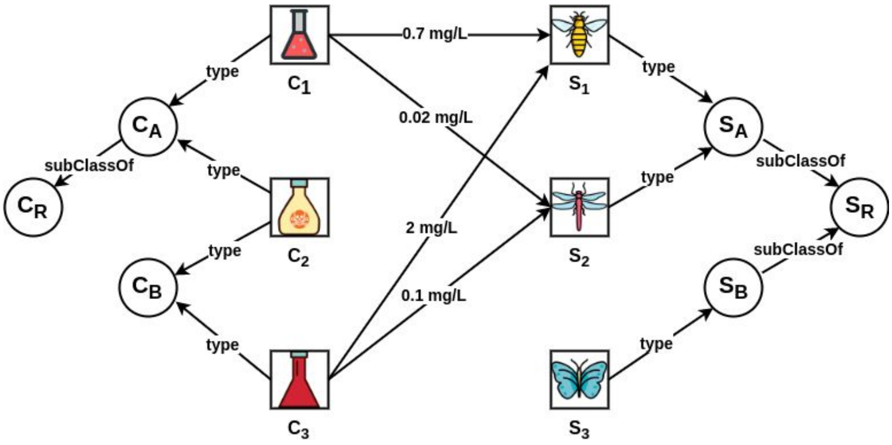

Figure 1: Problem definition with example concentrations.

2. Methods

We describe the two methods that model the problem shown in Fig. 1. These models are an expansion of the models described in elsewhere[1].

We compare the models using two metrics, accuracy (total number of correct classifications in percent) and $f_{0.5}$ which favours precision over recall. This gives emphasis to the importance of the models for unveiling potentially harmful chemicals (i.e., lower concentrations).

Data. We gather the effect data from ECOTOX[2] (https://cfpub.epa.gov/ecotox/). ECOTOX also includes a suite of experimental features (Fig. 2). Data from PubChem (https://pubchem.ncbi.nlm.nih.gov/) and ChEMBL (https://www.ebi.ac.uk/chembl/) are aggregated into a chemical graph (classification hierarchy) while the species taxonomy is generated from the NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy). ECOTOX uses proprietary identifiers for chemicals and species which need to be aligned to the external chemical and species graphs. This is done using Wikidata (https://www.wikidata.org/) and the alignment tool LogMap[3]. These mappings enable us to model effects outside the domain of ECOTOX.

Deterministic Model. We predict unknown concentration for chemical and species combinations without available laboratory data, e.g., at which concentration does C_3 affect S_3 in Fig. 1. To predict the effect concentration we identify the closest pair of chemical and species where the concentration is known, in this case C_3 and $S_2(0.1\ mg/L)$. This method can be generalized to not only include a chemical-species effect pair, but use a set of the closest pairs and use the average as a prediction of the effective concentration.

Probabilistic Model. Fig. 2 describes the probabilistic model. The experimental features are described in ECOTOX, features such as experimental duration can be used directly. Moreover, categorical properties such as organism life stage are assigned a value (e.g. juvenile as 0, smolt as 1, and adult as 2 for fish). The chemical features used here are derived descriptors from chemical formula like simple QSAR models. However, the chemical features can easily be replaced with higher dimensional properties (e.g., fingerprints). Finally, the chemical and taxonomic graphs are embedded into a high dimensional vector space using knowledge graph embedding models. The aim of these models is to mimic the graph structure in a continuous space.

After the inputs are represented as vectors, we concatenate the vectors and pass them through a series of hidden multi-layer perceptron (dense) layers, before making a prediction of the effect concentration.

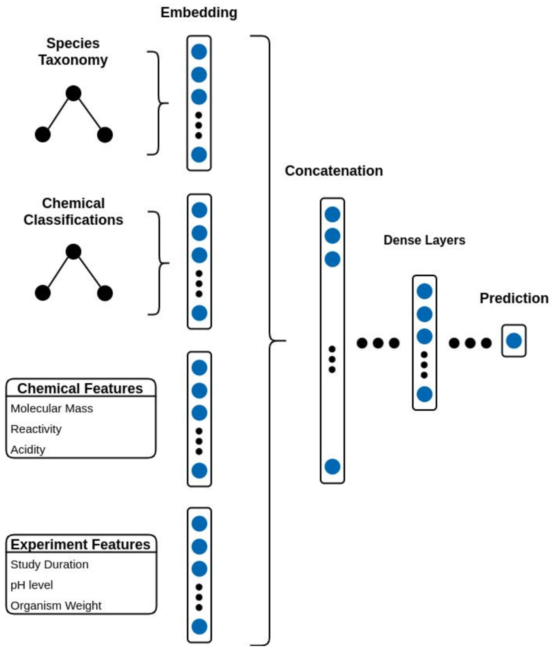

Figure 2: Conceptual probabilistic model.

3. Results

We perform an experiment using binary LC50 values, i.e. the predictors task is to separate highly toxic chemicals from less toxic chemicals.
The results for this setting show an improvement of 25 points (58 to 83%) using the probabilistic model over the deterministic one in terms of accuracy. The probabilistic model also gives a large improvement for the $f_{0.5}$ metric, 14 points.

4. Conclusions

We propose a method for including the chemical classification hierarchy and the species taxonomy in effect modelling. We expect to extend the methods to applications outside classification (regression). This method enables background knowledge to be encoded into the model and thereby improve the application domain of the model and supports read-across where effect does not exist.

5. References

[1] Myklebust EB, Jimenez-Ruiz E, Chen J, Wolf R, Tollefsen KE. 2019. Knowledge Graph Embedding for Ecotoxicological Effect Prediction. In: Ghidini C. et al. (eds) The Semantic Web – ISWC 2019. Lecture Notes in Computer Science, vol 11779. Springer, Cham

[2] Myklebust EB, Jimenez-Ruiz E, Chen J, Wolf R, Tollefsen, KE. 2019. Enabling Semantic Data Access for Toxicological Risk Assessment. CoRR, abs/1908.10128.

[3] Jiménez-Ruiz E, Cuenca Grau B, Zhou Y, Horrocks I. 2012. Large−scale Interactive Ontology Matching: Algorithms and Implementation. In the 20th European Conference on Artificial Intelligence (ECAI 2012).

Acknowledgement - This work is supported by grant 272414 from the Research Council of Norway (RCN), the MixRisk project (RCN 268294), and the SIRIUS Centre for Scalable Data Access (RCN 237889). The work is organised under NIVA’s Computational Toxicology Program (NCTP).