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

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

Background

Effect modelling is an important part of 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.

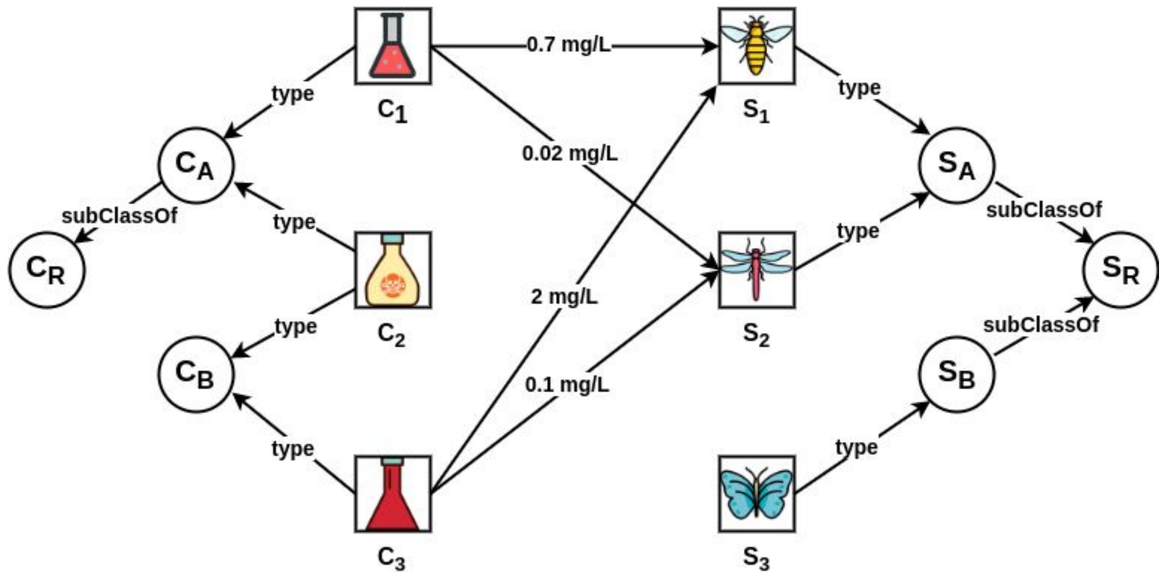

Figure 1: Problem definition with example concentrations.

Approach

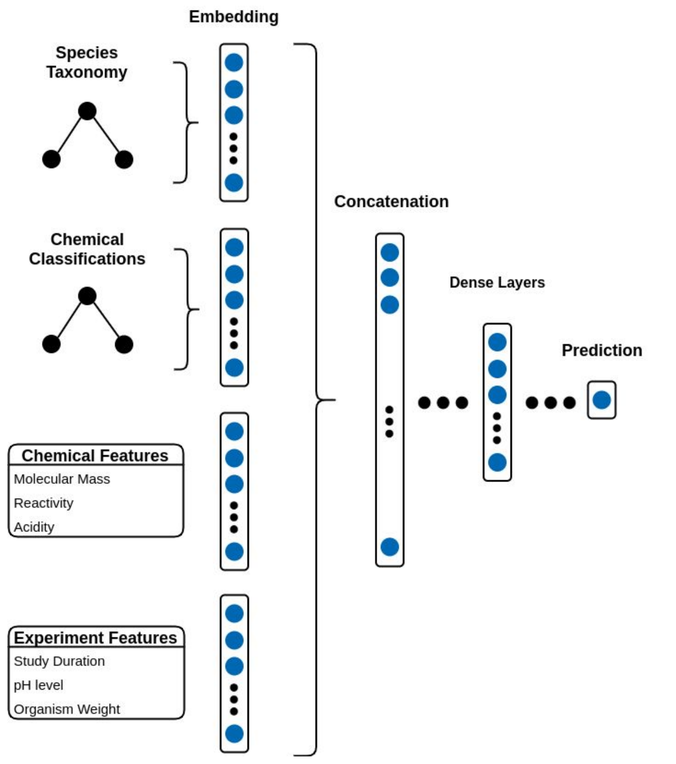

Figure 2: Probabilistic model.

Data. We gather the effect data from ECOTOX. ECOTOX also includes a suite of experimental features (Fig. 2). Data from PubChem and ChEMBL are aggregated into a chemical graph as classification hierarchy, while the species taxonomy is generated from the NCBI Taxonomy. All these datasets are aligned using information from Wikidata.

Deterministic Model. We predict unknown effect concentration for chemical and species combinations without available laboratory data, *e.g.*, at which concentration does C3 affect S3 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 C3 and S2 (0.1 mg/L).

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. The simplest examples of which is LogP. Moreover, 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.

Results

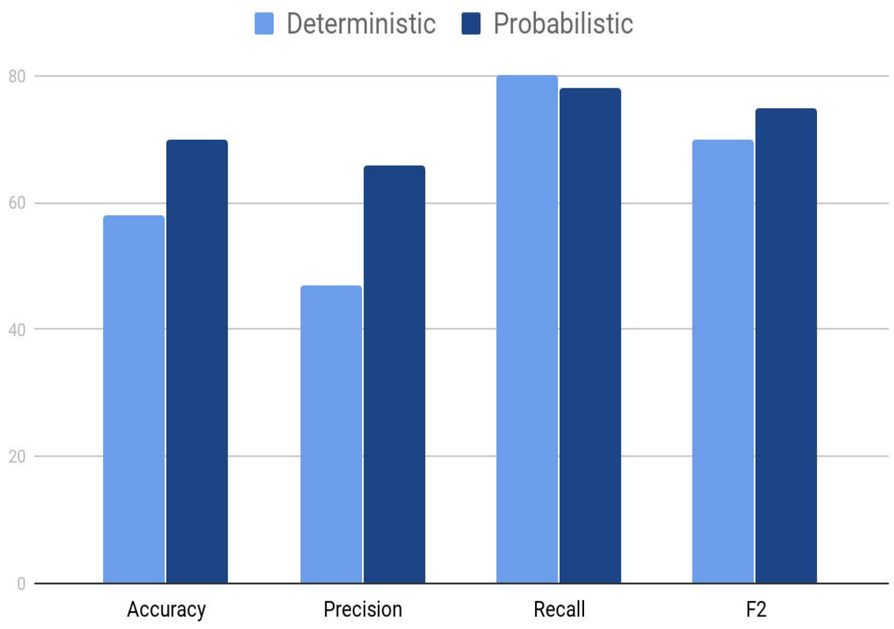

Figure 3: Metrics in percent.

Results. The results consider 434 chemicals and 225 species used in 1736 experiments. Fig. 3 shows the results of the the deterministic and probabilistic models. The probabilistic model outperforms the determinist in all metrics, except for recall. This indicates that the deterministic model overestimates the effects of certain chemicals.

The results support the notion that models cannot only consider the hierarchical structures, but must include the relations between the hierarchies, which is achieved with the probabilistic model.

Conclusion. 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, *e.g.*, regression. This method enables background knowledge to be encoded into the model and thereby improve the application domain and supports read-across where effect information does not exist. This will be instrumental to the assessment of chemicals and species combinations where data is sparse or do not exist.

Future perspectives

We are continuing efforts to integrate and align more disparate datasets, which will enable better performance of the models. Moreover, the extension of the data sources will give better insight into how models make predictions. The work will in the future be integrated into risk assessment data tools such as NIVA's Risk Assessment database, RAdb. Resources can be found at https://github.com/NIVA-Knowledge-Graph/ .

Contact info:
Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, N-0349 OSLO, Norway
Email: ebm@niva.no ; ket@niva.no

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).

With funding from The Research Council of Norway