A Knowledge Graph for Ecotoxicological Risk Assessment and Effect Prediction

Erik B. Myklebust

[1] Norwegian Institute for Water Research, Oslo, Norway
[2] Department of Informatics, University of Oslo, Norway
erik.b.myklebust@niva.no

Abstract. Exploring the effects a chemical compound has on a species takes a considerable experimental effort. Appropriate methods for estimating and suggesting new effects can dramatically reduce the work needed to be done by a laboratory. In this PhD research we aim at exploring the suitability of using a knowledge graph embedding approach for ecotoxicological effect prediction. A knowledge graph is being constructed from publicly available data sets, including a species taxonomy and chemical classification and similarity. We use ontology alignment techniques to integrate the effect data into the knowledge graph. Our preliminary experimental results show that the knowledge graph based approach improves the selected baselines.

Keywords: Knowledge graph · Semantic embedding · Ecotoxicology

1 Problem statement

Ecotoxicological risk assessment is the task of estimating the risk to a ecosystem by foreign chemicals. The diverse datasets used in risk assessment needs to be aggregated into a common vocabulary before being used in the risk prediction process. This aggregation requires the use of (semi-)manually curated mappings. Creating these mappings is a tremendous task for the domain experts that would benefit from suitable tool support.

At the heart of the data is the *effects*. This data describes the effects compounds has on species. The majority of effect data relates a compound-species pair to a mortality or chronic (*e.g.*, reproductive) effect. Due to the large search space of compound-species pairs, less than 1% of possible combinations has been studied. As a result of the large cost and effort to conduct these experiments, this proportion will not suddenly increase.

Hence, we have two main research tasks, where the latter is reliant on the first. These can be summarized as follows:

(i) Create a knowledge graph by gathering and integrating the relevant biological effect data and knowledge, such that to relieve the (domain) researchers of the manual work.

(ii) Using the knowledge graph together with machine learning techniques to predict effects. The objectives of this task are twofold:

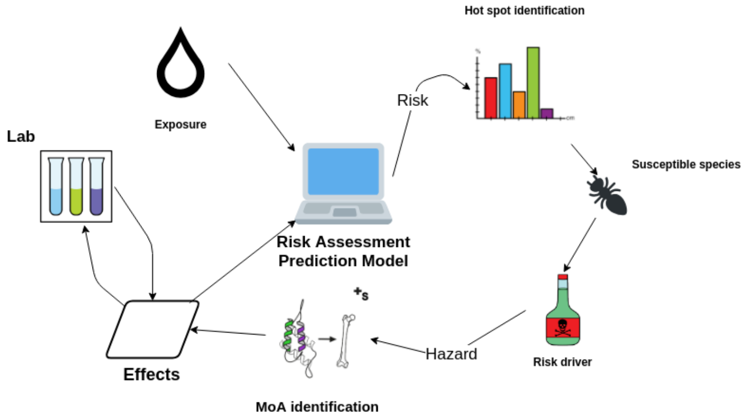


Fig. 1: Risk assessment pipeline.

(a) Limit the search space for the laboratory (binary prediction).
(b) Predict effects outright with a margin of error (regression).

2 Background and related work

In this section we introduce some preliminaries and give insights into the current state of the art efforts applying semantic web technologies within the field of toxicology and risk assessment.

Use case. Ecotoxicology is a multidisciplinary field that studies the ecological and toxicological effects of chemical pollutants on populations, communities and ecosystems. Risk assessment is the result of the intrinsic hazards of a substance combined with an estimate of the environmental exposure (*i.e.*, Hazard + Exposure = Risk).

Figure 1 shows a risk assessment pipeline. *Exposure* is data gathered from the environment, while *effects* are hypothesis that are tested in a laboratory. These two data sources are used to calculate risk, which is used to find (further) susceptible species and the mode of action (MoA) or type of impact a compound would have over those species. Results from the MoA analysis are used as new effect hypothesis.

Effect prediction. Estimating the effect a compound has on a species is a large research field within ecotoxicology. Currently, state-of-the-art solutions such as Quantitative Structure-Activity Relationship (QSAR) models (*e.g.*, [7, 13, 14]) exists. However, these are limited in scope. Each QSAR consider small groups of compounds and a single or a few species. Therefore, a general approach suited for a larger subset of the domain is favourable.

Knowledge graphs. We follow the RDF-based notion of knowledge graphs [4] which are composed by RDF triples $\langle s, p, o \rangle$, where s represents a subject (a class or an instance), p represents a predicate (a property) and o represents an object (a class, an instance or a data value *e.g.*, text, date and number). RDF entities (*i.e.*, classes, properties and instances) are represented by an URI (Uniform Resource Identifier). A knowledge graph can be split into a TBox (terminology), often composed by RDF Schema constructors like class subsumption and property domain and range,[3] and an ABox (assertions), which contain relationships among instances and semantic type definitions. RDF-based Knowledge Graphs can be accessed with SPARQL queries, the standard language to query RDF graphs.

There is emerging work in improving the usability of ecotoxicological data by mapping to knowledge graphs or ontologies, *e.g.*, [10], however, currently this work is limited. We are unaware of work incorporating the vast array of sources that is required from beginning to end by a risk assessment system.

Ontology alignment. Ontology alignment is the process of finding mappings or correspondences between a source and a target ontology or knowledge graph [9]. These mappings are typically represented as equivalences among the entities of the input resources.

Currently, mapping ecotoxicological data to different sources are under construction. The ECOTOX web search interface[4] now contains mappings to a external taxonomy source [20] (for a limited number of taxons). Fay et al. [10] indicates a full mapping to external sources exists, however, this is not yet publicly available.

We are not aware of efforts toward mapping taxonomic classes, *e.g.*, *genus*, *family*, etc. which can reveal inconsistencies in the datasets.

Embedding models. Knowledge graph embedding [24] plays a key role in link prediction problems where the goal is to learn a scoring function $S : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$. $S(s, p, o)$ is proportional to the probability that a triple $\langle s, p, o \rangle$ is encoded as true. Several models has been proposed, *e.g.*, Translating embeddings model (TransE) [5]. These models are applied to knowledge graphs to resolve missing facts in largely connected knowledge graphs, such as DBPedia [17].

There is previous work investigating modelling of chemical effects, *e.g.*, [16, 12]. The prediction of ecotoxicological effects can be seen as a sub-problem. These works investigate models that use the chemical structures to determine their effect on species. Yet, we are not aware of approaches where multiple knowledge graph embeddings are used to model the interaction between knowledge graphs.

3 Relevance

The relevance of the research to be conducted during the PhD can be summarized as follows:

[3] The OWL 2 ontology language provides more expressive constructors. Note that the graph projection of an OWL 2 ontology can be seen as a knowledge graph (*e.g.*, [1]).

[4] https://cfpub.epa.gov/ecotox/

(i) Manually integrating background knowledge into risk assessment systems is cumbersome since a common vocabulary does not exists. Our approach will reduce the time spent organizing data, and increase the number of case studies than can be conducted. A common vocabulary will enhance the interoperability between several risk assessment systems, increasing the confidence in the assessments.

(ii) The effect data used in risk assessment models is the result of time-consuming laboratory work. By using machine learning techniques with background knowledge, in the form of a knowledge graph, we aim at being able to limit the search space for new tests to be analysed in the laboratory. For example, we aim at recommending the top-ten compounds to test on a specific species, rather than conducting experiments using thousands of possible compounds.

(iii) Design and implementation of a fully-fledged recommender system to predict the level of effect on a species. For example, DEET (pesticide) has the potential to kill 50% of the population of the common house fly. Such *generalization* using the available data and knowledge is the main target of the research, which aims at reducing to a minimum further laboratory analysis.

4 Research questions and hypothesis

This work aims to address the following questions questions:

a. Can the disparate data sources used in ecotoxicological risk assessment be integrated into a knowledge graph to improve accessibility?

b. Can the knowledge graph be used to improve (or diversify) ecotoxicological effect prediction over current state-of-the-art models?

The hypothesis associated with the above questions are:

A. It is possible to integrate disparate data sources in a toxicological knowledge graph using Semantic Web tools.

B. Extrapolation of effect data increase the reach of risk assessment systems while remaining accurate.

5 Approaches

This section will describe the approaches used to investigate the hypothesis above. The evaluation of the hypothesis is described in Section 7.

Hypothesis A. There are multiple sources, varying from tabular, SPARQL endpoints, REST APIs, and RDF formats, each with its own vocabulary, that needs to be integrated to enable a unified data access. The main sources of data are:

(i) Effect data (ECOTOX [23], example seen in Table 1) in tabular format. Includes limited metadata linked to proprietary identifiers for compounds and species.

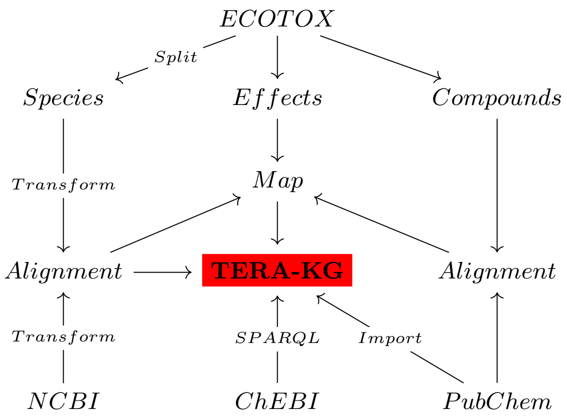


Fig. 2: Data sources in the TERA knowledge graph. Compound classification is available from PubChem. Chemical class hierarchy from the ChEMBL SPARQL endpoint. Compound literals are gathered from PubChem REST API and transformed into triples. ECOTOX and PubChem identifiers are aligned using the Wikidata SPARQL endpoint. ECOTOX and NCBI taxonomies are aligned using ontology alignment.

(ii) Compound data from different sources. Hierarchies available through downloadable RDF files and SPARQL endpoints (PubChem [22] and ChEMBL [6]). Compound features, *e.g.*, Molecular weight, XLogP etc. are available through the PubChem REST API.

(iii) The tabular NCBI taxonomy [20] is used as the hierarchy for species.

We must map the identifiers used in the effect data to open standards to take advantage of the diversity of data sources. The created *Toxicological Effects and Risk Assessment* (TERA) knowledge graph with current sources and aggregation steps is shown in Figure 2. Excerpts of triples from TERA are shown in Table 2.

Improving the knowledge graph can be done with several sources. First, a dataset containing biological activity, *e.g.*, Chemical ontology (CO) [11]. Such datasets would enable finer grained data to be used by the effect predictor. We

test_id	reference_number	test_cas	species_number
1068553	5390	877430 (2,6-Dimethylquinoline)	5156 (Danio rerio)
2037887	848	79061 (2-Propenamide)	14 (Rasbora heteromorpha)

result_id	test_id	endpoint	conc1_mean	conc1_unit
98004	1068553	LC50	400	mg/kg diet
2063723	2037887	LC10	220	mg/L

Table 1: ECOTOX database entry examples.

#	subject	predicate	object
(i)	ecotox:group/Worms	owl:disjointWith	ecotox:group/Fish
(ii)	ncbi:division/2	owl:disjointWith	ncbi:division/4
(iii)	ecotox:taxon/34010	rdfs:subClassOf	ecotox:taxon/hirta
(iv)	ncbi:taxon/687295	rdfs:subClassOf	ncbi:taxon/513583
(v)	compound:CID10198308	rdf:type	obo:CHEBI_134899
(vi)	compound:CID10198308	pubchem:formula	‘‘$C_7H_6O_6S$’’
(vii)	ecotox:chemical/115866	ecotox:affects	ecotox:effect/001
(viii)	ecotox:effect/001	ecotox:species	ecotox:taxon/26812
(ix)	ecotox:effect/001	ecotox:endpoint	LC50
(x)	ecotox:taxon/33155	owl:sameAs	ncbi:taxon/311871

Table 2: Example triples from the TERA knowledge graph

also aim at including an anatomy dataset, such that the biological activity can be aggregated from proteins to individual level.

Another aspect important to effect prediction is the habitat of the species, *e.g.*, [8]. Including the species habitat data will limit the effect prediction search space further, *e.g.*, heavy insoluble compounds (sinks in water) would have little/no effect on fish.

Hypothesis B. The prediction task at hand is depicted in Figure 3. Initially, we use a naive approach, which is to assume that *similar* compounds has a comparable effect on the same species and *vice versa*. The state of the art in risk assessment systems implement akin solutions. However, it is not clear what constitutes similarity in this context. The similarity between compounds are quantifiable using different methods, however, similarity does not imply similar biological activity [18]. For species, the naive solution is to calculate the taxonomic distance, but again the classifications of species is not defined by the susceptibility to compounds. Consequently, additional sources that describe these phenomena need to be added to the knowledge graph. When the knowledge graph is enriched with this data we can explore modelling techniques for embedding the knowledge graph for the purpose of predicting effects. We aim at applying simple embedding methods, TransE [5], DistMult [25], and HolE [21], until their performance is exhausted. These model may preform adequately for producing recommendation to the lab, however, as shown in the next section these models cannot be fully trusted to predicting effects outright. Therefore, we intend to include more expressive models, such as Graph Convolution Networks (GCN) [15]. Current approaches in knowledge graph embedding do not consider sparsely connected knowledge graphs, such as the hierarchical structures that make up TERA. Therefore, we aim at using the classification power of GCNs to embed groups of species or compounds more accurately. This will include the use of the vast array of chemical properties (experimental or computed) and the protein classification available for most species.

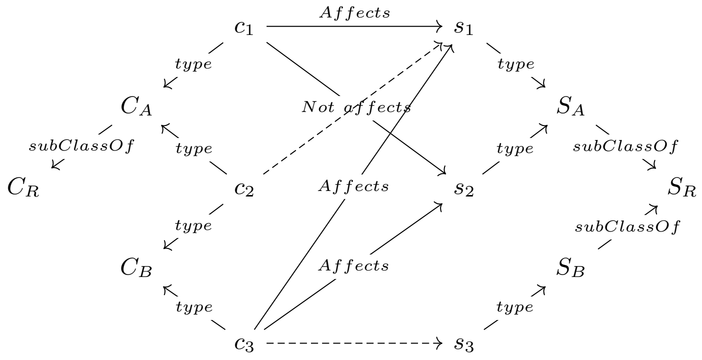


Fig. 3: The effect prediction problem. Lowercase s_j and c_i are instances of species and compounds, while uppercase denote classes in the hierarchy. Solid lines are observations and dashed lines are to be predicted. *i.e.*, does c_2 affect s_1?

	30 NN	**MLP**	**MLP + Embedding**		
Accuracy	0.53	0.81	**0.82**	**0.82**	0.81
Precision	0.45	**0.78**	**0.78**	**0.78**	0.75
Recall	**0.80**	0.74	0.79	0.79	**0.80**
$\mathbf{F_1}$ score	0.58	0.76	**0.78**	**0.78**	0.77
$\mathbf{F}_{\beta=\mathbf{2}}$ score	0.69	0.75	0.78	0.78	**0.79**
AUC	–	0.89	0.89	**0.90**	0.89

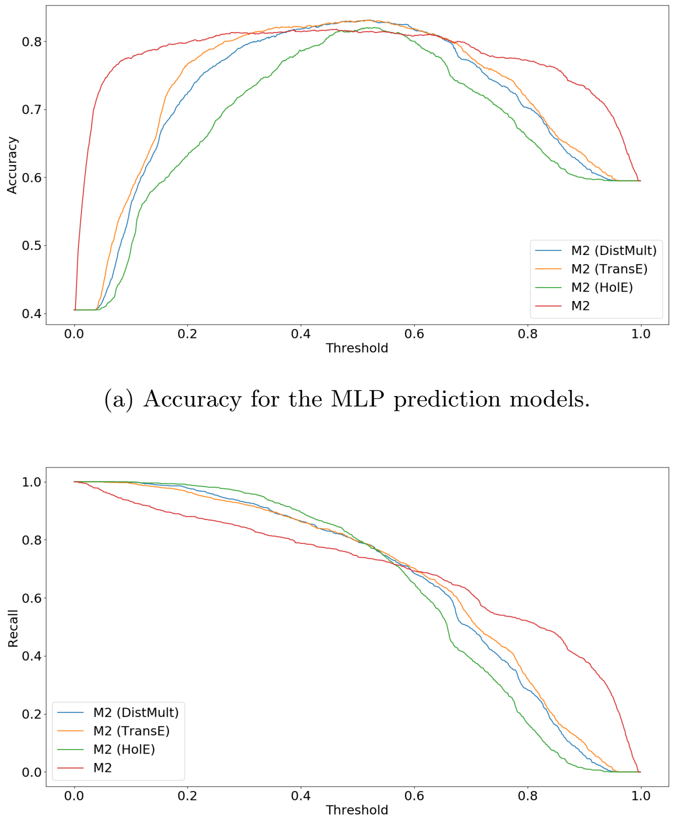
Table 3: Performance of the prediction models. All values are averages over 10 clean test runs. NN is the graph nearest-neighbour approach (using the closest 30 neighbours). MLP is the multi-layer perceptron model. The three values under MLP + Embedding are the results using TransE, DistMult, HolE embedding methods, respectively.

6 Preliminary results

We have evaluated three (plus variants) prediction models based on the effect data and the TERA knowledge graph. Note that currently the TERA knowledge graph has been created with the bare minimum of sources required to integrate the effect data with external metadata for compounds and species. Selected results are shown in Table 3. Prediction models used in this preliminary evaluation:

(i) A nearest-neighbour approach. A compound-species pair can inherent an effect if another compound or species is *close* in the knowledge graph. This method provides a useful baseline. However, the performance of this method is far from ideal, as it either will over or underestimate effects based on the number of neighbours considered.

(ii) A zero-background-knowledge multi-layer perceptron (MLP) model was applied to the effects data. This model is able to learn simple relations, *e.g.*,



(a) Accuracy for the MLP prediction models.

(b) Recall for the MLP prediction models.

Fig. 4: Accuracy and Recall for the MLP models with various thresholds.

s_1 and s_2 is effected by c_3, therefore, c_3 is toxic and will effect s_3. However, when this model is presented with previously unseen compound-species pairs, it cannot rely on background knowledge, and hence, the prediction will be highly flawed.

(iii) Using knowledge graph embedding ([5, 25, 21]) on TERA, followed by the same MLP model architecture as above yields better results for recall (which is preferred), while accuracy remains similar. In contrast to the above model, this model is more uncertain when unseen combinations are presented to the model (*in dubio pro reo*). As shown in Figures 4a and 4b, lowering the decision threshold (from 0.5 to 0.35) would yield a higher recall (0.93) for the HolE-based model, while reducing the accuracy (0.75).

The obtained predictions are promising and show the potential usefulness of the machine learning models in our setting and the benefits of using the TERA knowledge graph. As mentioned before, we favour recall with respect to precision. One the one hand, false positives are not necessarily harmful, while overlooking the hazard of a chemical may have important consequences. On the other hand, due to the limited experiments in terms of concentration (*i.e.*, effect data may not be complete), some chemicals may look less toxic than others while they may still be hazardous. At the same time the adoption of a RDF-based knowledge graph enables the use of an extensive range of Semantic Web infrastructure that is currently available (*e.g.*, reasoning engines, ontology alignment systems, SPARQL query engines).

7 Evaluation plan

In this section, we introduce the evaluation plan for the success of this project. We can divide the evaluation of both research questions into qualitative and quantitative measures.

The value of the knowledge graph in toxicology research is uncertain at this stage. The knowledge graph must provide value for the researchers. We can ensure this by evaluating the quality of the knowledge graph. Our definition of quality is that the knowledge graph should have high levels of:

(i) Coverage. The sources included in the knowledge graph must cover the areas of interest. The coverage also relates to the degree of successful mappings between the sources. There will be a trade-off between completeness and correctness of the mappings.

(ii) Integration. The ease of integrating the various sources. This involves aligning and mapping to attain a consolidated knowledge graph.

(iii) Functionality. The ability of the knowledge graph to be integrated into the risk assessment systems. This includes keeping the flexibility of Semantic Web technology without a commitment to a schema. We can add new triples and extend the knowledge graph, without the need of major changes.

(iv) Embedding enrichment. The semantic enrichment the knowledge graph gives the embeddings compared to embeddings learned from effects.

The quantitative evaluation of the knowledge graph generation is tightly related to that of evaluating the effect prediction. We can evaluate the ability to make good effect predictions quite easily. This can be done with precision, recall, accuracy, etc. for binary effects or with mean squared error, R^2-score, etc. for regression. However, the value of these predictions is integrating them into the risk assessment pipeline. The evaluation metrics must be inline with the ability our methods have to enhance risk assessment. We will compare environmental case studies results before and after the use of our modelling results. Since there is *no* ground truth data for risk assessment we rely on domain experts to determine if our contributions adds value to the assessments.

Risk assessments has currently large margins of errors (experimental errors etc.), and we may introduce new sources of error with our effect predictions.

However, we are confident that errors can also be reduced by greater data coverage. These are different types of errors and part of the evaluation process will be to find the optimal trade-off between them.

The current preliminary results uses random dataset splits for training and testing the models. We aim at introducing highly selective datasets that can test predictive performance in different scenarios. We will also try a completely clean test, where we recommend compound-species pairs to be tested in the lab. This will obviously be limited by the available compounds and test species of the particular laboratory.

The methodologies and knowledge graphs will be publicly available such that feedback from the community can help us evaluate and improve our contributions.

8 Reflections

The conducted work falls into one of the main research lines of toxicology research to enhance the generation of hypothesis to be tested in the laboratory [19]. Furthermore, the data integration efforts and the construction of the TERA knowledge graph is a large contribution to the area of risk assessment. The availability and accessibility of the best knowledge and data will enable optimal decision making.

Knowledge graph embedding models have been applied in general purpose link discovery and knowledge graph completion tasks [24]. They have also attracted the attention in the biomedical domain to find, for example, candidate genes for a disease, protein-protein interactions or drug-target interactions (*e.g.*, [3, 2]). However, we are not aware of the application of knowledge graph embedding models in the context of toxicological effect prediction.

Acknowledgements

This PhD project is supported by grant 272414 from the Research Council of Norway. I would like to thank my supervisors, Ernesto Jimenez-Ruiz (The Alan Turing Institute and University of Oslo), Raoul Wolf (Norwegian Institute for Water Research), and Knut Erik Tollefsen (Norwegian Institute for Water Research) for their feedback on this work. In addition, I would also like to thank Jiaoyan Chen (University of Oxford), Martin Giese (University of Oslo) and Zofia C. Rudjord (Norwegian Institute for Water Research) for their contribution in different stages of this PhD research.

References

1. Agibetov, A., Jiménez-Ruiz, E., Ondresik, M., Solimando, A., Banerjee, I., Guerrini, G., Catalano, C.E., Oliveira, J.M., Patanè, G., Reis, R.L., Spagnuolo, M.: Supporting shared hypothesis testing in the biomedical domain. J. Biomedical Semantics **9**(1), 9:1–9:22 (2018)

2. Agibetov, A., Samwald, M.: Global and local evaluation of link prediction tasks with neural embeddings. In: 4th Workshop on Semantic Deep Learning (ISWC workshop). pp. 89–102 (2018)
3. Alshahrani, M., Khan, M.A., Maddouri, O., Kinjo, A.R., Queralt-Rosinach, N., Hoehndorf, R.: Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics **33**(17), 2723–2730 (2017)
4. Arnaout, H., Elbassuoni, S.: Effective Searching of RDF Knowledge Graphs. Web Semantics: Science, Services and Agents on the World Wide Web **48**(0) (2018)
5. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26, pp. 2787–2795. Curran Associates, Inc. (2013)
6. ChEBI-ontology: The european bioinformatics institute (2019), https://www.ebi.ac.uk/chebi/
7. Escher, B.I., Baumer, A., Bittermann, K., Henneberger, L., Knig, M., Khnert, C., Klver, N.: General baseline toxicity qsar for nonpolar, polar and ionisable chemicals and their mixtures in the bioluminescence inhibition assay with aliivibrio fischeri. Environ. Sci.: Processes Impacts **19**, 414–428 (2017)
8. European Environment Agency: Linkages of species and habitat types to maes ecosystems (2015), https://www.eea.europa.eu/data-and-maps/data/linkages-of-species-and-habitat
9. Euzenat, J., Shvaiko, P.: Ontology Matching, Second Edition. Springer (2013)
10. Fay, K., Elonen, C., Hoff, D., Skopinski, M., Pilli, A., Wang, R., LaLone, C.: Enhancing the Utility of the ECOTOX knowledgebase (ECOTOX KB) via ontology-based semantics mapping. In: SETAC Europe, Rome, ITALY, May 14 - 18, 2018. (2018)
11. Feldman, H.J., Dumontier, M., Ling, S., Haider, N., Hogue, C.W.: Co: A chemical ontology for identification of functional groups and semantic comparison of small molecules. FEBS Letters **579**(21), 4685 – 4691 (2005)
12. Forbes, V.E., Calow, P., Sibly, R.M.: Are current species extrapolation models a good basis for ecological risk assessment? Environmental Toxicology and Chemistry **20**(2), 442–447 (2001)
13. Khan, K., Benfenati, E., Roy, K.: Consensus qsar modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the drugbank database compounds. Ecotoxicology and Environmental Safety **168**, 287 – 297 (2019)
14. Khan, K., Khan, P.M., Lavado, G., Valsecchi, C., Pasqualini, J., Baderna, D., Marzo, M., Lombardo, A., Roy, K., Benfenati, E.: Qsar modeling of daphnia magna and fish toxicities of biocides using 2d descriptors. Chemosphere **229**, 8 – 17 (2019)
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. CoRR **abs/1609.02907** (2016), http://arxiv.org/abs/1609.02907
16. Laender, F.D., Morselli, M., Baveco, H., den Brink, P.V., Guardo, A.D.: Theoretically exploring direct and indirect chemical effects across ecological and exposure scenarios using mechanistic fate and effects modelling. Environment International **74**, 181 – 190 (2015)
17. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015)
18. Martin, Y.C., Kofron, J.L., Traphagen, L.M.: Do structurally similar molecules have similar biological activity? Journal of Medicinal Chemistry **45**(19), 4350–4358 (Sep 2002)

19. Myklebust, E.B., Jimenez-Ruiz, E., Rudjord, Z.C., Wolf, R., Tollefsen, K.E.: Integrating semantic technologies in environmental risk assessment: A vision. In: 29th Annual Meeting of the Society of Environmental Toxicology and Chemistry (SETAC) (2019)
20. NCBI-Taxonomy: The national center for biotechnology information (2019), https://www.ncbi.nlm.nih.gov/taxonomy
21. Nickel, M., Rosasco, L., Poggio, T.A.: Holographic embeddings of knowledge graphs. CoRR **abs/1510.04935** (2015), http://arxiv.org/abs/1510.04935
22. PubChem: National institutes of health (nih) (2019), https://pubchem.ncbi.nlm.nih.gov/
23. U.S. EPA: Ecotoxicology knowledgebase (ecotox) (2019), https://cfpub.epa.gov/ecotox/
24. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Trans. Knowl. Data Eng. **29**(12), 2724–2743 (2017)
25. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. CoRR **abs/1412.6575** (2015)