Towards reproducibility and transparency of QSARs: Comparison of applicability domain approaches

***Project:* MLTox: Enhancing Toxicological Testing through Machine Learning**

Eawag: Christoph Schür, Marco Baity-Jesi, Kristin Schirmer
SDSC, ETH Zürich: Lilian Gasser, Fernando Perez-Cruz

Background

Chemicals can only be put on the market if they are deemed safe both for humans and the environment. This registration process involves extensive animal testing. For instance, in ecotoxicology, the tests are mainly carried out on fish, crustaceans and algae. Given the evident ethical and economical concerns of animal testing, several approaches are taken to reduce it, summarized under the term New Approach Methods (NAMs). Some NAMs are based on Quantitative Structure Activity Relationship (QSAR) models, which are widely used in cheminformatics to predict activities or properties of chemicals based on their structure.

In the MLTox project, we focus on predicting toxicity, more specifically acute mortality of fish, crustaceans, and algae using *in vivo* experimental data. To foster comparability across studies that predict ecotoxicological outcomes a common dataset was needed. Hence, we created the ADORE dataset, which is based on the ECOTOX database and also contains species-specific and taxonomic information [3]. The response variable for mortality is the lethal concentration 50 (LC50), *i.e.*, the concentration at which half of a population dies. Currently, we are developing machine learning models to predict LC50 that integrate chemical, species-related and experimental features.

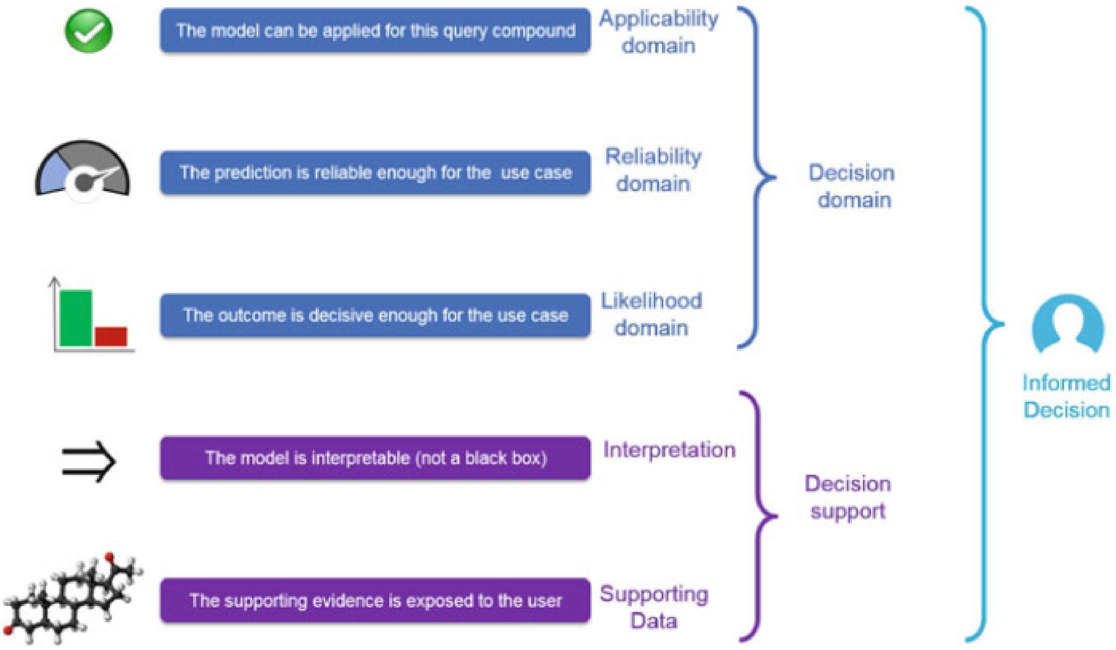
The model can be applied for this query compound
Applicability domain
The prediction is reliable enough for the use case
Reliability domain
The outcome is decisive enough for the use case
Likelihood domain
Decision domain
The model is interpretable (not a black box)
Interpretation
The supporting evidence is exposed to the user
Supporting Data
Decision support
Informed Decision

Figure 1: The TARDIS concept. Figure from Hanser et al. [1]

To be considered for chemical hazard assessment in the regulatory process, QSAR models need to be validated. Within the so-called applicability domain (AD), QSAR models are valid which means that they have reasonable predictive ability. The OECD defines AD as ”the response and chemical structure space in which the model makes predictions with a given reliability” [2]. There exist several approaches to determine the AD, of which one is summarized under the acronym *TARDIS* which stands for transparency, applicability, reliability, decidability, interpretability and support (Figure 1). Two classes of AD can be distinguished, where the first focuses on the feature space and the second class includes the response value [4].

A better understanding of AD is not only relevant in ecotoxicology but can be applied to every model using chemical structures as input to predict a chemical activity or property.

In this MSc project, the aim is to

- generate an overview of what applicability domain entails and which approaches and models can be used to describe it
- run selected models on several datasets (ADORE, *in vitro* data from ToxCast/Tox21, ...) to determine their AD
- create a dashboard that loads compounds from a file, calculates suitable AD models for it and present them in an informative and appealing manner.

This MSc project is an essential step to validate the modeling output from the MLTox project, which aims to predict acute mortality of fish, crustaceans and algae using *in vivo* experimental data.

Additional Information

- **What will you learn?**
  - Implementation, optimization and comparison of different applicability domain models
  - Building a dashboard and communicating complex information effectively
- **Requirements:**
  - Good knowledge of Python and git
  - Experience with classification and regression models is an advantage
  - A strong interest in chemistry, cheminformatics, and data visualization
- **Supervisors and collaborators:**
  - Eawag: Christoph Schür, Marco Baity-Jesi, Kristin Schirmer
  - ETHZ: Lilian Gasser, Fernando Perez-Cruz
- Please contact Lili Gasser (lilian.gasser@sdsc.ethz.ch) or Christoph Schür (christoph.schuer@eawag.ch) for further information

References

[1] Thierry Hanser et al. “Applicability Domain: Towards a More Formal Framework to Express the Applicability of a Model and the Confidence in Individual Predictions”. In: *Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science.* Ed. by Huixiao Hong. Challenges and Advances in Computational Chemistry and Physics. Cham: Springer International Publishing, 2019, pp. 215–232. ISBN: 978-3-030-16443-0. DOI: 10.1007/978-3-030-16443-0_11. (Visited on 06/28/2023).

[2] OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD Series on Testing and Assessment. OECD, Sept. 2014. ISBN: 978-92-64-08544-2. DOI: `10.1787/9789264085442-en`. (Visited on 07/07/2023).

[3] Christoph Schür et al. "A Benchmark Dataset for Machine Learning in Ecotoxicology". In: *Scientific Data* 10.1 (Oct. 2023), p. 718. ISSN: 2052-4463. DOI: `10.1038/s41597-023-02612-2`.

[4] Zhongyu Wang and Jingwen Chen. "Applicability Domain Characterization for Machine Learning QSAR Models". In: *Machine Learning and Deep Learning in Computational Toxicology*. Ed. by Huixiao Hong. Cham: Springer International Publishing, 2023, pp. 323–353. ISBN: 978-3-031-20730-3. DOI: `10.1007/978-3-031-20730-3_13`. (Visited on 07/07/2023).