Cytiva Chemical Alternatives Assessment:
The Development of Machine-Learning Tool for Toxicity Prediction

Ziye Zheng[1,2], Swapnil Chavan[2], Ulf Norinder[3], Ian Cotgreave[2], Jean-Luc Maloisel[1], Ismet Dorange[1]

[1] *Cytiva, Björkgatan 30, 753 23 Uppsala, Sweden*

[2] *Chemical and pharmaceutical safety, Research Institute of Sweden (RISE), Forskargatan 18, 151 36 Södertälje, Sweden*

[3] *Department of Computer and Systems Sciences, Stockholm University, P.O. Box 7003, 164 07 Kista, Sweden*

Background

Sustainability towards a toxic-free environment is a major chemical strategy in EU. At Cytiva, we continuously search for greener, less toxic alternatives for chemicals used in our products or manufacturing processes. To achieve this, a chemical alternatives assessment tool was required.

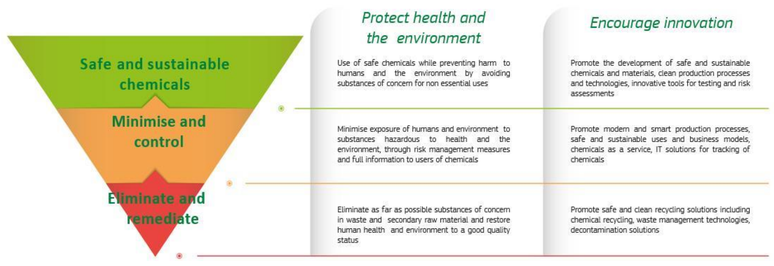

Fig 1. Sustainability towards a toxic-free environment strategy in EU

The Project at Cytiva

A project aiming at developing a chemical hazard prediction toolbox was initiated some years ago and initial toolbox v1.0 was developed. This is based on a multi parameter optimization approach that rank each chemical alternative with a score (0-5) with the consideration of several parameters such as availability, cost, technical performance, physical hazard as well as environmental and health impact.

The sustainability scores (environmental and health impact) are obtained from an equation using hazard labels by CLP Regulation from the ECHA database of REACH dossiers. For chemicals that don't have sufficient data in the REACH database, a penalty score (2.5) is assigned in the equation. The final outcome is a single score assorted to an uncertainty score that gives an idea of how much data is available (see data gap in Fig. 2) for a particular chemical.

To address this gap, a machine-learning model was envisaged to predict chemical toxicity data.

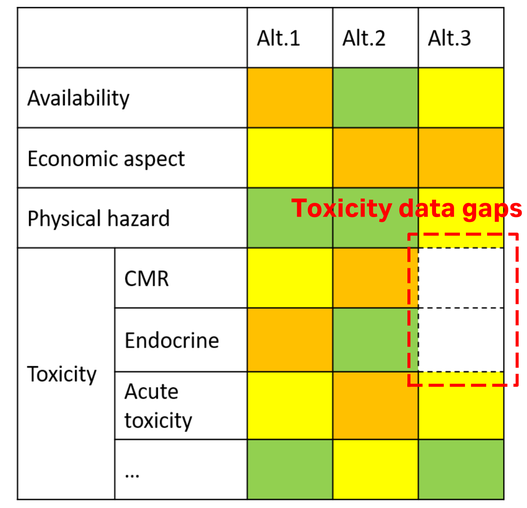

Fig 2. Challenge in chemical alternatives assessment: data gaps

Machine-Learning Models for Chemical Toxicity Prediction

Machine-learning models are trained with known experimental data for target toxicity endpoints and used for prediction of unknown chemicals. Nowadays high-quality models are possible because of the well-developed machine-learning algorithms as well as more available experimental data for model training. In recent years, machine-learning models for toxicity assessment are also more and more accepted by authorities. Compared to Experimental measurements, the modeling tools are cheaper and faster, and rather easy to use.

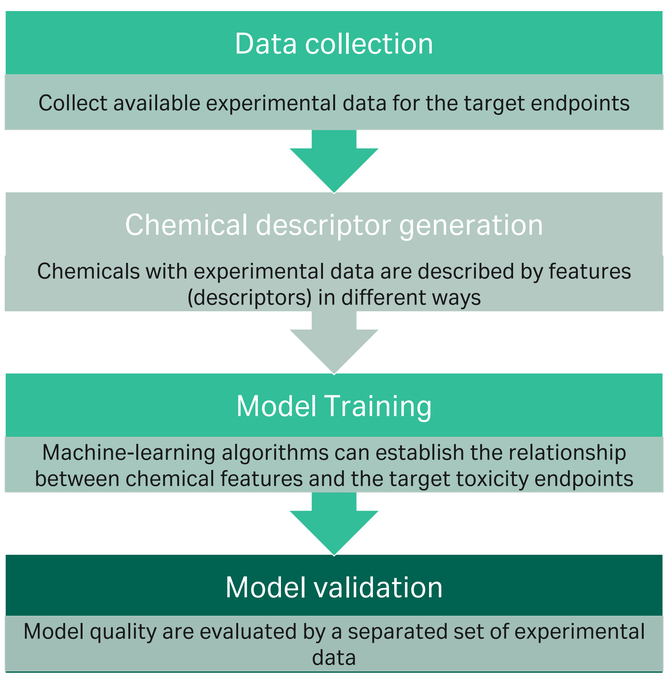

Fig 3. Summary of the process to develop machine-learning models

Results

Model development

Machine-learning models have been developed for four human toxicity endpoints: carcinogenicity, mutagenicity, reproductive toxicity and skin sensitization. Large amount of experimental data (more than 1000 for each endpoint) were collected for model training and validation. Different chemical descriptors (RDKit, PaDeL, cddd and chemical reactivity calculated by quantum chemistry) and different machine-learning (Random Forest and Support Vector Machine) were tested. As a result, Random Forest algorithm combined with RDKit descriptors sticks out, as it has good performance for all four endpoints, and require less computational powers.

Carcinogenicity	Mutagenicity	Reproductive toxicity	Skin sensitization
• trained with **1535** experimental data • **balanced accuracy 70-72%**	• trained with **8249** experimental data • **balanced accuracy 85-88%**	• trained with **1823** experimental data • **balanced accuracy 85-86%**	• trained with **1004** experimental data **balanced accuracy 70-79%**

Fig 4. Four developed machine-learning models

Cytiva Computational Toxicity Prediction Tool

With the four developed models, a user-friendly prediction tool was developed using Jupyter Notebook. The advantages of this tool includes:

- Secure : the tool does not require internet connection or any third-party software
- Accurate: balanced accuracy 70-85%, higher than most of the current open-sourced models
- Broad coverage: each model trained with >1000 compounds, which covers a broad chemical space
- Easy to use: only needs a text file with target chemical structure (SMILES) as an input file
- Fast: Batch prediction for 100 compounds in less than 10 seconds on a standard office laptop

Fig 5. Cytiva Computational Toxicity Prediction Tool

Conclusion

A prediction tool with four machine-learning models was developed which can provide prediction results for carcinogenicity, mutagenicity, reproductive toxicity and skin sensitization with good accuracy. With the data provided by the prediction tool, the chemical hazard prediction toolbox can make more accurate selection of greener chemical alternatives.

Future Perspective

- More machine-learning models will be developed in the prediction tool to cover other important toxicity endpoints (*e.g.* acute oral toxicity, endocrine disruption properties and ecotoxicities)
- A feature to evaluate prediction accuracy will be added with the conformal prediction method
- For the chemical hazard prediction toolbox , data from other open-sourced tools (*e.g.* VEGA and QSAR Toolbox) will also be taken into consideration

Contact Information
Ziye Zheng
postdoc researcher in computational chemistry
ziye.zheng@cytiva.com
ziye.zheng@ri.se

In collaboration with

RI.SE Research Institutes of Sweden

MISTRA SAFECHEM

cytiva