A Novel Trustworthy and User-friendly Chemical Hazard Prediction Toolbox

Ziye Zheng[1,2,3], Swapnil Chavan[2], Ulf Norinder[4], Ian Cotgreave[2]

[1]Cytiva, Björkgatan 30, 75 323 Uppsala, Sweden

[2]Chemical and pharmaceutical safety, Research Institute of Sweden (RISE), Forskargatan 18, 15 136 Södertälje, Sweden

[3]IVL Swedish Environmental Research Institute, 10 031 Stockholm, Sweden

[4]Department of Computer and Systems Sciences, Stockholm University, P.O. Box 7003, 16 407 Kista, Sweden

INTRODUCTION

Machine learning prediction for chemical hazard screening has been highlighted in recent years, and many hazard prediction models and tools have been developed. However, there are several problems with the currently available models or tools, including:

- Outdated training sets with which the models were trained
- Limited chemical space since recently published records are not considered
- Outdated tools/workflows which are incompatible with current version of language and operating system
- Lack of graphical user interface (GUI) which limits their use by non-programmers
- Lack of an uncertainty quantification feature.

In the Mistra SafeChem Programme, we aim at developing an *in silico* toolbox that is easy to use and provides chemical hazard prediction with high trustworthiness.

Data collection

- Most recent data were collected for 35 hazard endpoints (Figure 1).
- large (over 1000) training set for each endpoint, covering a wide chemical space.

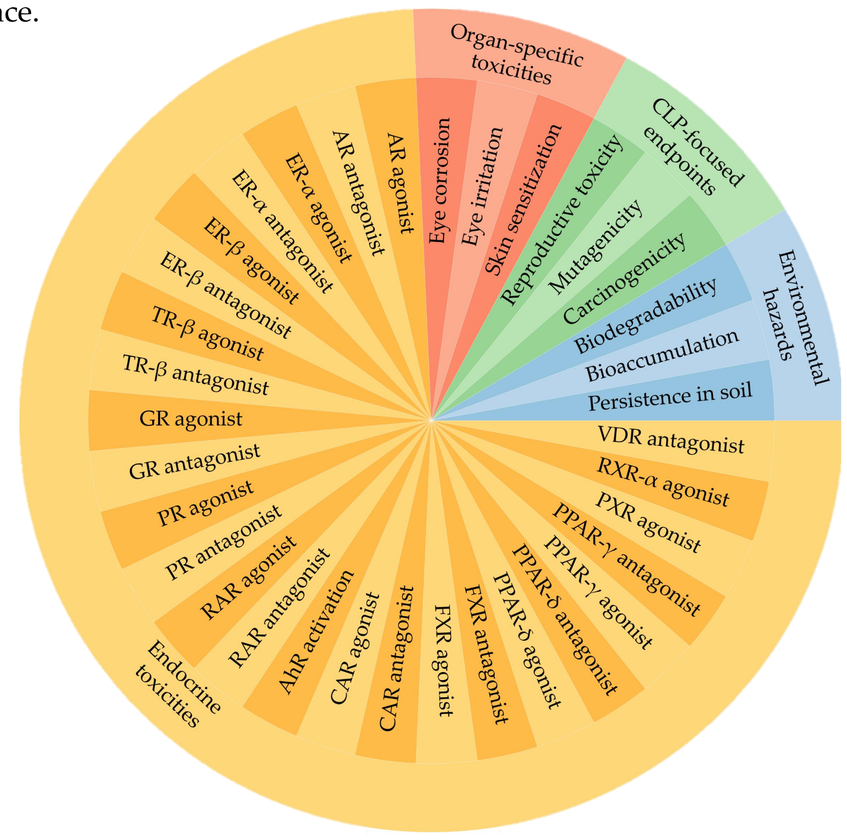

Figure 1. Hazard endpoints included in Mistra SafeChem *in silico* Toolbox

Uncertainty calibration and model validation

- All model predictions are calibrated with mathematical methods like the conformal prediction framework (Figure 2).
- Model quality are validated both internally (cross validation) and externally.
- For most endpoints, three models are available for each endpoint, and a consensus prediction is also provided along with a quantified uncertainty.

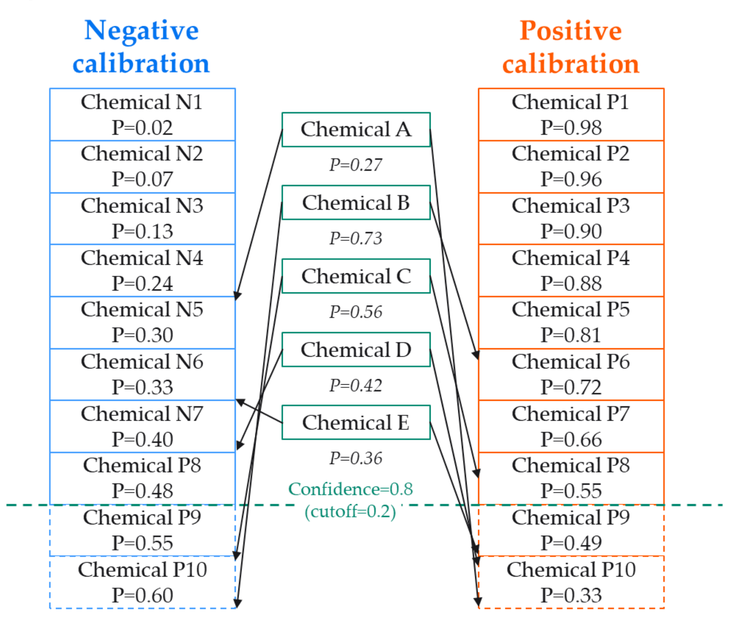

Figure 2. Illustration of the conformal prediction framework

Model training

- Three sets of models were trained, including descriptor-based machine-learning models and descriptor-free deep-learning models (Figure 3).
- In total, 67 newly trained models for hazard prediction were included.

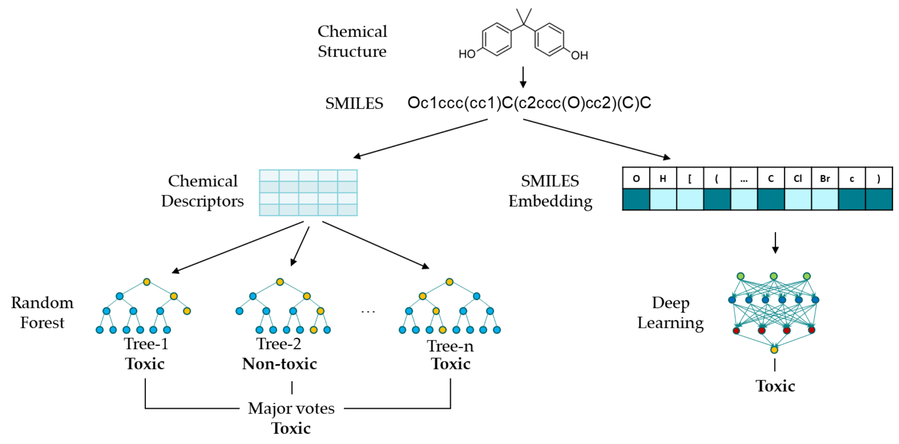

Figure 3. Different models trained in Mistra SafeChem Project

GUI and software developments

- All models are packaged into a Windows executable package (.exe) with a friendly user interface (Figure 4).
- The software is secured as all data are processed on the local machine.

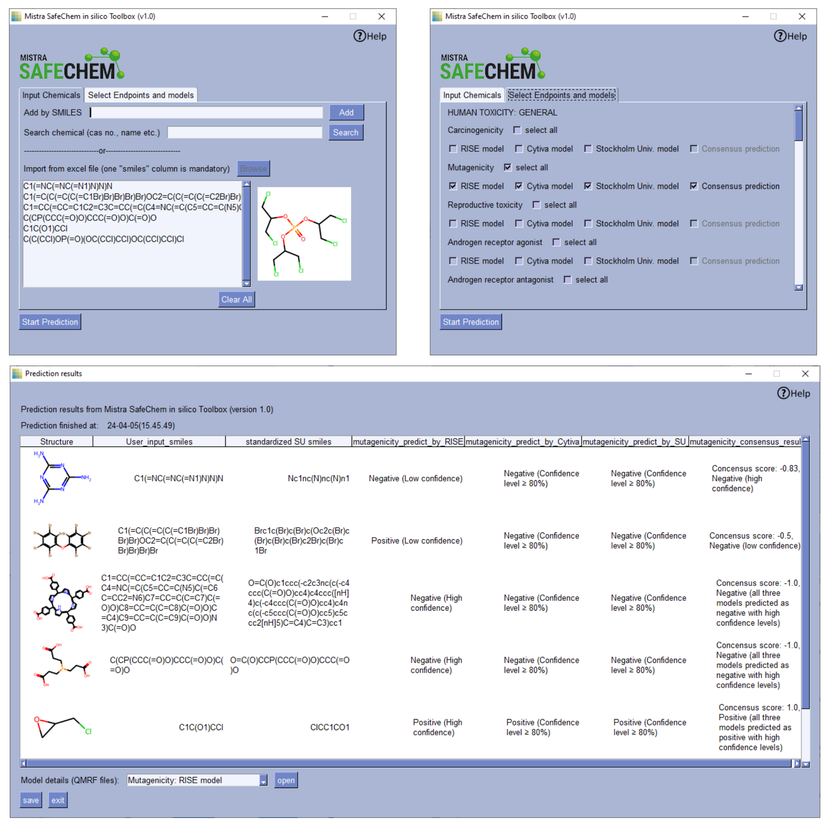

Figure 4. Graphical user interface (GUI) of Mistra SafeChem *in silico* Toolbox

CONCLUSIONS

- The Mistra SafeChem *in silico* Toolbox provides reliable predictions for 35 chemical hazard endpoints.
- The software is safe and user-friendly.
- The toolbox has been tested and proven useful for chemical hazard screening within the Mistra SafeChem programme.

Model Developer: cytiva, RISE Research Institutes of Sweden, Stockholm University

Programme Coordinator: ivl Swedish Environmental Research Institute

Programme website: https://www.ivl.se/projektwebbar/mistra-safechem.html

Presenter: Ziye Zheng, Ph.D. Tel: +46 76 134 22 47 Email: Ziye.Zheng@ivl.se