Predictive Models for Acute Oral Systemic Toxicity

Background & Scope. One of ICCVAM's high priority efforts is to develop alternative test methods for the U.S. Environmental Protection Agency (EPA) "six pack" tests: acute oral, dermal, and inhalation systemic toxicity tests and tests to determine eye and skin irritation and skin sensitization. These tests are required by regulatory agencies worldwide and represent the highest cumulative animal use across chemical sectors. As part of the ICCVAM effort, NICEATM in collaboration with EPA's National Center for Computational Toxicology (NCCT) have compiled a large body of rat acute oral lethality data that can be used to develop predictive *in silico* models of acute oral systemic toxicity.

Regulatory acceptance of a computationally derived prediction is unlikely to be achieved with a single model developed by a single lab, as a single model may not be able to cover the chemical space of all available structures and address all endpoints. With this in mind, NICEATM is organizing an international modeling effort to predict acute oral toxicity endpoints using the compiled rat acute oral lethality data. The collaboration is not intended to be a competition between member teams. Rather, the desired outcomes are to leverage each model's strengths and to overcome the limitations of any individual approach to ultimately establish consensus predictions combining the submitted models. In addition to the importance of the collaborative approach to developing the model, it has been established in the scientific literature that ensemble models have the highest statistically significant performance.

Objective. The objective of this collaboration is to integrate the collective expertise of the international modeling community to develop predictive models for acute oral toxicity based on regulatory needs put forward by ICCVAM member agencies. Submissions will be evaluated using external datasets, and those meeting defined criteria will be used to produce consensus predictions for acute oral toxicity endpoints of interest to regulatory agencies. All participants that submit models for any endpoint(s) will be invited to present poster(s) at a workshop in April 2018, and participants selected by the organizing committee (based on the model evaluation criteria below) will receive sponsored invitations to give platform presentations. The outcomes of the collaboration will be submitted for publication in the peer-reviewed literature, and the model predictions will be hosted on the EPA's Chemistry Dashboard.

Endpoints to be modeled. Based on the range of regulatory criteria and decision contexts used by ICCVAM agencies, a total of five different modeling endpoints have been identified. Participants can build models to predict one or more of the following endpoints:

- Very toxic (<50 mg/kg vs. all others)
- Nontoxic (>2000 mg/kg vs. all others)
- LD50 point estimates
- EPA hazard categories (n=4)
- GHS hazard categories (n=5)*

*GHS category 5 and "not classified' have been combined into a single group representing any LD50 value >2,000 mg/kg for the purpose of this modeling effort due to the overabundance of limit tests and uncertainty associated with values above 2,000 mg/kg

Dataset. NICEATM and NCCT have compiled and curated a rat acute oral toxicity database of systemic toxicity (LD50) values. The database was mined for chemicals with QSAR-ready structures, yielding a chemical inventory of 11,992 CASRN to be used for this modeling effort. This dataset was split semi-randomly into a training set (75%) and evaluation set (25%) to ensure equivalent coverage with respect to LD50 distribution. The evaluation set will be embedded within a large prediction set to facilitate blinded evaluation of model outputs.

The training dataset, comprising 8,994 chemicals, is provided for participants to train/optimize their models together with the QSAR-ready structures in two formats (SMILES in tab-delimited training and MOL in SDF files). Additionally, a file containing all compiled LD50 values for all training set chemicals is provided for additional transparency. Three files are available that summarize the training set:

1. **Training Dataset Tab-Delimited Format** (TrainingSet.txt): a file providing CASRN, DTXSID (where available), QSAR-ready structures in the form of SMILES, and a single value per chemical corresponding to each of the requested endpoints, respectively. Values for each endpoint are provided in separate columns.
2. **Training Dataset in SDF Format** (TrainingSet.sdf)**:** a file providing CASRN, DTXSID (where available), QSAR-ready structures in MOL format (2D), and a single value per chemical corresponding to each of the requested endpoints, respectively. Values for each endpoint are provided in separate fields of the SDF.
3. **Complete LD50 Inventory** (TrainingSet_full_LD50.txt): a file providing CASRN and all LD50 values (point estimates and limit tests) for all chemicals in the training set. Some chemicals had multiple LD50 values identified during the compilation of the data, which are all provided in this file for complete transparency.

A brief description of data processing to compute the endpoint values for the training dataset is summarized below:

- To mitigate potential misclassification due to limit test values on the border of hazard categories, any value with a greater than (>) or less than (<) symbol was adjusted. For example, >2,000 mg/kg was retained as 2,001 mg/kg and <2,000 mg/kg was evaluated as 1,999 mg/kg for the purpose of category identification.
- To compute a single LD50 value when chemicals had multiple LD50s reported, the median of the lower quantile was computed across LD50 values omitting any limit test data. If a chemical only had limit test data reported, no point estimate for LD50 is provided, but these chemicals were still considered for the other modeling endpoints.
- Designations of “very toxic” and “nontoxic” were based on the computed single LD50 values described above and on the limit test data.
- For EPA categories, all limit tests reporting an LD50 value of >2,000 mg/kg were omitted because there was no way to determine whether the true category assignment was III or IV. Any other reported LD50 values were considered, and chemicals with multiple values were handled as described above.
- For GHS categories, all data were considered. Category 5 (>2,000 mg/kg) and Not Categorized (>5,000 mg/kg) were combined for the purposes of modeling.

Submitted model predictions will be evaluated by NICEATM using the independent evaluation set, which will be embedded in a large prediction set of chemicals compiled based on interest to the organizing committee. The large prediction set will be released shortly after the training set and will contain CASRN, DTXSID (where available), and chemical structures in machine-readable format but with no activity data.

Model Building. Modelers are encouraged to consider different modeling and machine learning approaches as well as global, local, and hybrid/consensus methods to ensure optimum predictivity with no specific restrictions or recommendations. Models could include focusing on specific chemical structural classes, product use categories, production volumes, etc. Many of these types of information are available via the EPA Chemistry Dashboard at: http://comptox.epa.gov, and the training dataset includes the DTXSIDs needed to run queries against the dashboard (but note that modelers are not restricted to this resource). Models may be based on chemical features, *in vitro* data, physicochemical properties, or any other inputs that are widely available and ideally free and open-source.

Model Dataset Release Dates:

Training set release: **November 17, 2017**

Prediction set release (evaluation set is within this prediction set): **December 15, 2017**

Model Submission Deadline: February 9, 2018

Model submissions should include documentation that is as detailed as possible. Modelers may wish to use the EURL ECVAM JRC's QSAR Model Reporting Format (QMRF Editor and Interactive QSAR database can be found here: http://qsardb.jrc.ec.europa.eu/qmrf/). At a minimum, documentation must include features/data used, how the applicability domain is defined, and a description of the modeling approach. Modelers are asked to submit the following items for each modeled endpoint:

- Documentation of the model: summarize what input data were used, modeling process, domain of applicability, etc. Using QMRF as a guideline is recommended.
- Abstract* summarizing model and results, including contact information for presenting author.
- Training dataset predictions and cross-validated performance in separate tab-delimited files for each endpoint/model, with CASRNs as IDs.
- Model results for the large prediction set in separate tab-delimited files for each endpoint/model with CASRNs as IDs (models will be evaluated using the evaluation set, which is a subset of this large prediction set). Modelers may choose to predict toxicity for all data or part of the data; in the latter case, explanation should be provided.
- For all predictions, an indication of whether each chemical is within the applicability domain of the model should be provided.

*All abstracts will be welcome for poster presentations at the April 11-12, 2018, workshop and will be considered for platform presentations. While workshop attendance is not required to participate in the project, it is strongly encouraged. Regardless of workshop

participation, an abstract summarizing the model(s) is necessary to facilitate model review and evaluation.

Evaluation Criteria. The organizing committee will review submitted abstracts, model documentation, and model performance; standard performance characteristics such as balanced accuracy or R^2 will be used. Different modeling and machine learning approaches can be tested and applied with no restrictions or approach-specific recommendations. However, modelers should find the best compromise between simplicity and predictivity, since the evaluation criteria will be based on the OECD QSAR validation principles as guidance (defined below).

The OECD QSAR validation principles to be considered as evaluation criteria will include:

- A defined endpoint
    - Separate models should be submitted corresponding to the five endpoints defined above.
- An unambiguous algorithm
    - Ensure transparency in the description of the model algorithm. Preference will be given to models using simple algorithms and open-source code.
- A defined domain of applicability
    - Define limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. Including ability to characterize uncertainty/confidence is a plus.
- Appropriate measures of goodness-of–fit, robustness and predictivity
    - Provide measures of quantitative performance, including cross-validated training set performance and external evaluation set performance (will be recalculated by the project organizers).
- Mechanistic interpretation, if possible
    - Describe the mechanistic associations between the descriptors used in a model and the endpoint (mode of action) being predicted.

It is important to note that, although the evaluation criteria are aligned with OECD QSAR validation principles, this modeling exercise is not restricted to QSAR models. Decision trees, read-across, hybrid models incorporating *in vitro* data, consensus, and approaches beyond those that are exclusively structure-based will also be considered. Further, while many QSAR models are high throughput and can process large numbers of chemicals, this project is intended to also include lower-throughput mechanistically driven models. To that end, modelers will be asked to submit predictions on as many chemicals as possible, but will not be required to provide predictions for the entire prediction set.

Timeline and deliverables:

November 17, 2017: Release of Training Data to the public**.** Data are made available as a tab-delimited file and SDF files via the NICEATM website. Curated data include CASRN, DTXSID, LD50 values and hazard categories (corresponding to designated modeling endpoints), coupled with the QSAR-ready structures (SMILES in the tab delimited files and MOL in the SDFs). Note: Mapping to DTXSIDs and Structure is based only on searching CASRNs provided with the original

source of the Acute Tox Data against the CompTox Chemistry Dashboard batch search (https://comptox.epa.gov/dashboard/dsstoxdb/batch_search)

December 15, 2017: Release of Prediction Data to the public. The prediction set will be made available as a tab-delimited file and SDF files via the NICEATM website. The prediction set includes CASRN, DTXSID, and QSAR-ready structures (SMILES in the tab delimited file and MOL in the SDFs). The evaluation set for model assessment and evaluation is within this large prediction set.

February 9, 2018: Deadline for submission of model results and documentation to NICEATM. As detailed above, submissions should include an abstract, model documentation, and results for both cross-validation performance on the training set and model predictions for the prediction set (of which the evaluation set is a subset).

March 9, 2018: Organizing Committee finalizes selection of models to be invited for platform presentations and notifications are sent to presenters.

April 11-12, 2018: Predictive Models for Acute Oral Systemic Toxicity Workshop, NIH Natcher Conference Center, Bethesda, MD.