Questions and Answers for Participants in the Predictive Models Project

Q1: Is registration required? Is there a deadline for registration?
A1: Registration is encouraged, as those who have provided contact information will be included on any announcements circulated to participants. However, registration is not required and groups who have not registered may submit models, nor is there a registration deadline. The registration will stay open until the deadline to submit the predictions (extended to February 16).

Q2: Were changes or updates posted after the initial release?
A2: One update was made to the training set files on November 27, 2017, which was announced via email to registered participants. A correction to the non-toxic endpoint designations was posted on November 27, 2017, resulting in an increase of 616 more chemicals (11,974 chemicals in total) in the training set having a non-toxic endpoint designation. The format of the tab-delimited text file was updated on November 30, 2017, to correct misalignment of the header row. No other changes were made since the initial data release to the data set used in the modeling.

Q3: How were endpoint values derived for training set vs. what is in the "Complete LD50 Inventory" file?
A3: Acute oral toxicity data provided for this effort comprises an extensive compilation of LD50 values from a wide variety of sources, resulting in multiple LD50 values for some chemicals. To facilitate model building, we have provided a single representative LD50 value per chemical in the training set files (TXT and SDF). For details on how each endpoint's representative value was obtained, please refer to the "detailed information for model submitters" PDF file. The supplementary file named "Complete LD50 Inventory" is provided for those groups that are interested in reviewing or integrating the entire LD50 inventory (including replicate LD50 values per chemical); this file contains only unique LD50 values per chemical comprising both point estimates and limit test values.

Q4: What will the prediction set look like and when will it be released?
A4: The prediction set will be a large list of chemicals (~50,000 CASRNs and corresponding structure information provided) to be virtually screened by participants' models, yielding predictions for the acute oral LD50 endpoints. That prediction set will contain the evaluation set (~3,000) within it, which the organizing committee will use to evaluate results. Participants are asked to provide predictions for as many of the prediction set chemicals as possible (it is understood that not all models will be amenable to accomplishing predictions for all ~50,000 chemicals); detailed directives will be provided upon release of the prediction set on December 15, 2017.

Q5: Can we use 3D structures instead of the provided 2D structures?
A5: Yes. Minimized 3D structures in the form of an SDF file are available upon request.

Q6: Can we get the original, pre-QSAR-ready structures?
A6: Yes. Original structures (pre-QSAR processing) are available upon request.

Q7: What are the "structure sources" provided in the training file?
A7: There are two structure sources associated with the training set chemical structures, which can be utilized as each participant deems fit for their model development. "EPA_DSSTox" structures are of highest quality. They are associated with active CASRNs, confirmed chemical names, and are available on the EPA CompTox Chemistry Dashboard. They represent 80% of the list. "Public_CrossChecked" structures are of lower certainty. They were mined and cross-checked between online sources. They could be misrendered and/or associated with deleted CASRNs. No chemical names are provided. They represent 20% of the list.

Q8: Was there any deduplication performed for the training set?
A8: Deduplication was performed on CASRNs only. This means that deleted/alternate CASRNs and active CASRNs pointing to the same structures but coming from different experimental sources were kept in the full dataset. This was done intentionally by request of some organizing committee/stakeholder members for transparency reasons, as all LD50 data were obtained based on reported CASRNs. Because the LD50 data were initially compiled by CASRN only, the structures were added later to facilitate modeling efforts. Additionally, after the QSAR-ready standardization procedure, different original structures may result in the same QSAR-ready structure. There are 158 duplicate QSAR-ready structures in the data set based on "InChI_Code_QSARr" and "Salt_Solvent", that the participants can include or exclude as they see fit for their approach. This list can be provided upon request.

Q9: What are the DTXSID values provided?
A9: The DTXSID values provided are DSSTox substance identifiers. These help participants access DSSTox to retrieve any chemical information. It is important to note that the CASRN, chemical name, and DTXSID provided in the training set and prediction set files all map to the original structures (pre-QSAR ready process), whereas the structure information provided (ie. SMILES, InChI key, etc.) are all for the QSAR-ready (QSARr) structures that were generated using the standardization workflow (as described in the "detailed information for model submitters" document). The original structures (Pre-QSAR-ready) for the training and prediction sets are available upon request.

Q10: Must LD50 predictions be made for the entire prediction set?
A10: Participants are encouraged to use their models to predict as many LD50 values as possible for the prediction set. It is understood that not all models are amenable to making so many predictions, therefore please make as many predictions as possible in order through the provided prediction set file (in order by the provided chemID).

Q11: Why is there overlap between the training and prediction set chemicals?
A11: The prediction set was designed separately based on lists of chemicals of interest to the workshop organizers, which were deduplicated across each other based on QSAR-ready structures. Any overlap with the training set in CASRNs or QSAR-ready structures is not relevant to the evaluation of the models. As explained in the "detailed information for model submitters" PDF file, the evaluation set is a small fraction of the prediction set and has no overlap of any kind (CASRN, QSAR-ready structure, or name) with the training set.