A Large Dataset of Acute Oral Toxicity Data Created for Testing in Silico Models

Jeremy Fitzpatrick, Grace Patlewicz

National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC

fitzpatrick.jeremy@epa.gov

Acute toxicity data is a common requirement for substance registration in the US. Currently only data derived from animal tests are accepted by regulatory agencies, and the standard in vivo tests use lethality as the endpoint. Non-animal alternatives such as in silico models are being developed due to animal welfare and resource considerations. We compiled a large dataset of oral rat LD50 values to assess the predictive performance currently available in silico models. Our dataset combines LD50 values from five different sources: literature data provided by The Dow Chemical Company, REACH data from eChemportal, HSDB (Hazardous Substances Data Bank), RTECS data from Leadscope, and the training set underpinning TEST (Toxicity Estimation Software Tool). Combined these data sources yield 33848 chemical-LD50 pairs (data points), with 23475 unique data points covering 16439 compounds. The entire dataset was loaded into a chemical properties database. All of the compounds were registered in DSSTox and 59.5% have publically available structures. Compounds without a structure in DSSTox are currently having their structures registered. The structural data will be used to evaluate the predictive performance and applicable chemical domains of three QSAR models (TIMES, PROTOX, and TEST). Future work will combine the dataset with information from ToxCast assays, and using random forest modeling, assess whether ToxCast assays are useful in predicting acute oral toxicity.