Development of Free and Open-source QSAR Tools for Predicting pKa

N. Cariello*, K. Mansouri*, A. Korotcov**, V. Tkachenko**, W. Casey***, A.J. Williams****

*ILS, Research Triangle Park, NC, USA

**Science Data Software, LLC, Rockville, MD, USA

***NIH/NIEHS/DNTP/NICEATM, Research Triangle Park, NC, USA

****EPA/ORD/NCCT, Research Triangle Park, NC, USA

The logarithmic dissociation constant, pKa, provides information about the ionization state of a chemical, which affects its lipophilicity, solubility, protein binding, and ability to cross the plasma membrane of a cell. These properties govern pharmacokinetic parameters such as absorption, distribution, metabolism, excretion, and toxicity. Therefore, accurate pKa predictions are critical for the assessment of chemical toxicity and biological activity. Predictions of pKa can be made using empirically based methods such as quantitative structure–activity relationships (QSARs) and quantum mechanical approaches such as density functional theory (DFT). A number of commercial pKa prediction software tools are available but very little exists in terms of open data sets and open prediction models. Predicting pKa is particularly challenging due to the lack of high-quality publicly available experimental data restricting the resultant QSAR models to specific chemical domains. The aim of this work was to provide free and open-source pKa predictors using a large publicly available experimental pKa dataset obtained from DataWarrior (www.openmolecules.org). Chemical structures were standardized for model fitting and validation. Three different machine learning algorithms, support vector machines, extreme gradient boosting and deep neural networks were used to build models using continuous descriptors and binary fingerprints generated by PaDEL. The best performing models for each algorithm were benchmarked using predictions from two commercial tools, ACD/Labs and Chemaxon, on an untested list of chemicals. This comparison showed varying degrees of concordance among the models, including between the proprietary tools. This was funded with U.S. federal funds from the NIEHS/NIH/HHS under Contract HHSN273201500010C.