ABSTRACT SYMPOSIUM NAME: Cheminformatics Resources & Software Tools Supporting Environmental Chemistry (Oral)
ABSTRACT SYMPOSIUM PROGRAM AREA NAME: CINF
TITLE: Prediction Of pKa From Chemical Structure Using Free And Open-Source Tools
AUTHORS: Valery Tkachenko (1), Neal Cariello (2), Alexandru Korotcov (1), Kamel Mansouri (3), Antony Williams (4)
INSTITUTIONS:
1 Science Data Software, LLC, Rockville, MD 20850
2 Integrated Laboratory Systems, Research Triangle Park, NC 27709
3 ScitoVation LLC, 6 Davis Dr, Research Triangle Park, NC 27709
4 National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, Durham, North Carolina 27711

Abstract: The ionization state of a chemical, reflected in pKa values, affects lipophilicity, solubility, protein binding and the ability of a chemical to cross the plasma membrane. These properties govern the pharmacokinetic parameters such as absorption, distribution, metabolism, excretion and toxicity and thus pKa is a fundamental chemical property and is used in many models of chemical toxicity.

Experimentally determining pKa is not feasible for high-throughput assays. Predicting pKa is challenging and existing models have been developed only using restricted chemical space (e.g., anilines, phenols, benzoic acids, primary amines) and lack of a generalized model impedes ADME modeling.

No free and open source models exist for heterogeneous chemical classes, however, several proprietary programs exist. In this work, pKa open data bundled with DataWarrior (http://www.openmolecules.org/) were used to develop predictive models for pKa. After data cleaning, there were ~3100 and ~3900 monoprotic chemicals with an acidic or basic pKa, respectively. 1D and 2D chemical descriptors (AlogP, Topological polar surface area, etc) in addition to 12 fingerprints (presence or absence of a chemical group) were generated using PaDEL software. Three datasets were used: acidic, basic and acidic and basic combined.

13 datasets were examined, the 1D/2D descriptors and 12 fingerprints. Using the Extreme Gradient Boosting algorithm showed that the MACCS and Substructure Count fingerprints yielded the best results, with models showing an R-Squared of ~0.78 and a RMSE of 1.42.

Recently, Deep Learning models have showed remarkable progress in image recognition and natural language processing. To determine if the Deep Learning algorithms would increase model performance we examined the datasets and found that the Deep Learning models were somewhat superior than Extreme Gradient Boosting with an R-Squared of ~0.80 and an RMSE of ~1.38.

This work does not reflect U.S. EPA policy.