Open Source QSAR Models For pKa Prediction Using Multiple Machine Learning Approaches

K Mansouri1, N Cariello1, A Korotcov2, V Tkachenko2, W Casey3, N Kleinstreuer3, D Allen3, C Grulke4, A Williams1
1ILS, RTP, NC, USA; 2Science Data Software LLC, Rockville, MD, USA; 3NIH/NIEMS/DNTP/NICEATM, RTP, NC, USA; 4NCCT, US EPA, RTP, NC

Background
- The pharmacokinetic properties of a chemical, such as solvophilicity, solubility, and permeability, are important factors in understanding its behavior and activity.
- A variety of physicochemical properties are linked to chemical structure, such as ionization, solubility, and metabolism.
- pKa is a measure of the extent to which a chemical undergoes ionization, and it can vary across different experimental conditions.

pKa Data
- The pKa data were obtained from DataWarrior (https://comptox.epa.gov) and industrial databases, including ChemAxon and ACD/Labs.

Machine Learning Algorithms
- Deep Neural Networks (DNN): These models are capable of capturing complex patterns in data.
- Extreme Gradient Boosting (XGB): A powerful model for classification and regression tasks.
- Deep Neural Networks (DNN): A type of neural network that can learn hierarchical representations of data.

Benchmark with the Commercial Tools
- To allow the user to navigate from the commercial tools to a benchmark for our models, we added the concordance of our models with DataWarrior.

Range of Predictions and Limitations
- The models were evaluated using a range of pKa values, from -15 to 15.
- The best models were compared to each other and to predictions from ACD/Labs and ChemAxon.

Study Goals and Procedure
- The goal of this study was to develop open-source QSAR models for pKa prediction that are comparable to commercial tools.
- The models were developed using multiple machine learning approaches.

Data Preparation for Modelling
- The data set contains a high number of duplicates and was preprocessed to ensure a more accurate representation of the data.

Model Performance
- The models were evaluated using root mean squared error (RMSE) and coefficient of determination (R²).

Range of Predictions and Limitations
- The range of predictions was from -5 to 15.
- The best models were compared to each other and to predictions from ACD/Labs and ChemAxon.

QSAR Modeling
- The models were trained using a variety of machine learning algorithms, including SVM and DNN.

Acknowledgements
- The authors acknowledge the financial support of the National Institute of Environmental Health Sciences (NIH/NIEMS/DNTP) and the National Institute of Standards and Technology (NIST) through the NICEATM project.
- The authors also acknowledge the efforts of the DataWarrior team and the ChemAxon and ACD/Labs teams.

References
- Mansouri K et al. 2016. Environ Health Perspect 124(7):1023-33. doi:10.1289/ehp.1510267

Summary and Next Steps
- The models and source codes will be available for download and use.
- The models will be benchmarked against other commercial tools to assess their performance.

To get announcements of NICEATM activities, visit the NIH mailing list page and click “Subscribe.”