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

K. Mansouri[1], N. Cariello[1], A. Korotcov[2], V. Tkachenko[2], W. Casey[3], N. Kleinstreuer[3], D. Allen[1], C. Grulke[4], A. Williams[4]

[1]ILS, RTP, NC, USA; [2]Science Data Software LLC, Rockville, MD, USA; [3]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA; [4]NCCT, US EPA, RTP, NC

Background

- The logarithmic dissociation constant, pKa, strongly influences a chemical's pharmacokinetic and biochemical properties.
  - pKa reflects the ionization state of a chemical, which affects lipophilicity, solubility, protein binding, and the ability to cross the plasma membrane and the blood-brain barrier. Thus, pKa affects absorption, distribution, metabolism, excretion and toxicity (ADMET).
  - Chemicals with no charge at a physiological pH will passively cross the plasma membrane more easily than charged molecules and are therefore more likely to have biological activity than passively diffused charged chemicals.
- pKa is an important parameter for physiologically based pharmacokinetic (PBPK) modeling, in vitro to in vivo extrapolation (IVIVE), and predicting tissue:plasma partition coefficients.
- Commercial software tools such as ACD/Labs and ChemAxon predict the pKa of individual ionization sites independently of chemical class. However, current publicly available pKa models are limited to certain chemical classes.

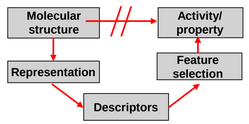
Study Goals and Procedure

- Here we provide free, open-source, fast, and reliable options for predicting pKa for heterogeneous chemical classes.
- Modeling steps:
  - pKa values for 7912 chemicals in water were obtained from DataWarrior, a freely available software package.
  - Chemical structures were standardized for QSAR modeling [1].
  - Continuous molecular descriptors, binary fingerprints and fragment counts were generated using PaDEL.
  - Several machine learning approaches were applied: deep neural networks (DNN), support vector machine (SVM), and extreme gradient boosting (XGB).
  - Models were 5-fold cross-validated and evaluated against an external test set.
  - The best models for each algorithm were compared to each other and to predictions from ACD/Labs and ChemAxon.

QSAR modeling

Conceptual basis

QSARs are based on the congenericity principle, which is the assumption that structurally similar compounds will have similar chemical properties.



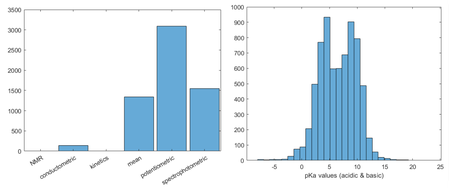
QSARs can be fast and accurate but they depend on the quality of the data used.

General Steps to Develop a QSAR model

- **Curation of experimental data**
- **Standardization of the chemical structures**
- **Preparation of training and test sets**
- **Calculation of an initial set of descriptors**
- **Selection of a machine learning algorithm**
- **Variable selection technique**
- **Validation of the model's predictive ability**
- **Define the Applicability Domain**
- **Interpretation of the selected descriptors, if possible.**

pKa Data

- The pKa data was obtained from DataWarrior (http://www.openmolecules.org/) and included experimentally measured aqueous pKa values and associated SMILES strings for 7912 heterogenous chemicals.

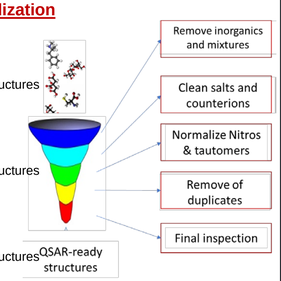


Methods for Measuring the pKa Reported in DataWarrior

Acidic and Basic pKa Values Reported in DataWarrior

Data Preparation for Modeling

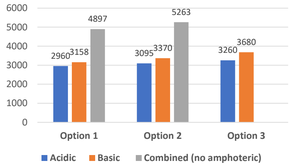
Structure Standardization



Acidic and Basic Datasets

The DataWarrior data set contained a high number of duplicates (1659) and amphoteric chemicals (chemicals with both an acidic and basic pKa). Data were processed in three different ways.

- *Option 1: all duplicates removed*
- *Option 2: low variability duplicates averaged*
- *Option 3: all data included (strongest pKa rule)*



QSAR-ready structures in Each of the Data Options

Training and Test Sets

- For each data option, the structures were split into training (75%) and test sets (25%).
- Training/test set splitting was performed semi-randomly to:
  - Keep similar distributions of pKa values
  - Keep similar distribution of acidic and basic pKas for combined datasets

Molecular Descriptors

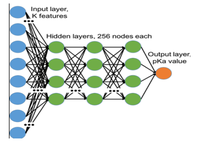
The QSAR-ready structures were used to calculate molecular descriptors and generate binary fingerprints and fragment counts using PaDEL.

- 1D and 2D continuous descriptors: 1444 descriptors.
- Binary fingerprints and counts: 9121 bits (CDK, Estate, MACCS, PubChem, Substructure, Klekota-Roth and 2D atom pairs).

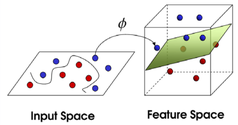
Machine Learning Algorithms

Deep Neural Networks (DNN)

- DNN maps features through a series of nonlinear functions that are linked in a combinatorial fashion to maximize model accuracy
- Tensorflow and Keras packages were used to build a feed-forward DNN with 3 hidden layers of 256 nodes each.



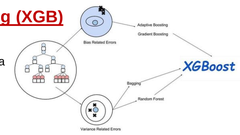
Support Vector Machine (SVM)



- SVM defines a non-linear decision boundary that optimally separates two classes.
- The free and open source package LibSVM3.1 was used for SVM implementation.

Extreme Gradient Boosting (XGB)

- XGB is used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- The R package caret was used to implement XGB.



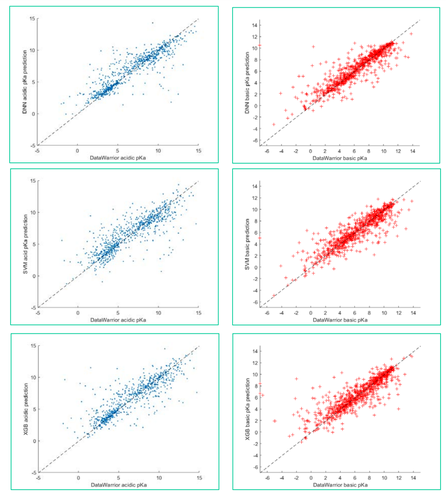
All used tools and resulting models are free and open source.

Model Performance

The pKa dataset was divided into acidic and basic pKa datasets, which were modeled separately.

Models were assessed using root mean squared error (RMSE) and the coefficient of determination (R^2). Test set results are reported here below.

Algorithm	Best Acidic Model RMSE	Best Acidic Model R^2	Best Basic Model RMSE	Best Basic Model R^2
DNN	1.51	0.80	1.57	0.77
SVM	1.80	0.72	1.53	0.78
XGB	1.82	0.71	1.90	0.67



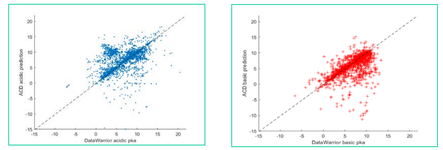
Benchmark with the Commercial Tools

Concordance between the commercial tools and DataWarrior

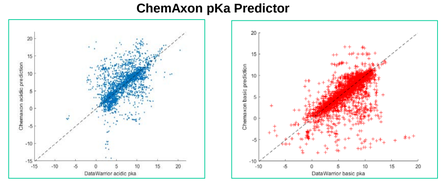
To be able to use the predictions from the commercial tools as a benchmark to our models, we first needed to assess the concordance of their predictions with DataWarrior.

	DataWarrior Acidic Dataset (3260)				DataWarrior Basic Dataset (3680)			
	ACD/Labs		ChemAxon		ACD/Labs		ChemAxon	
Predicted chemicals	3145	3206	1618	3649	3145	3206	1618	3649
R^2	-0.21	-0.11	-0.05	0.23	-0.21	-0.11	-0.05	0.23
RMSE	3.72	3.52	3.00	2.79	3.72	3.52	3.00	2.79

ACD/Labs pKa Predictor



ChemAxon pKa Predictor



External Set Prediction and Model Concordance

- A set of 8904 QSAR-ready structures (non overlapping with DataWarrior) from the TSCA-actives list (https://comptox.epa.gov/dashboard), was used as benchmark to compare the predictions of the models from this work and the commercial tools.
- For this analysis, the SVM model was implemented in OPERA (https://github.com/kmansouri/OPERA) [2].

Comparison of All Models for the Acidic pKa Predictions

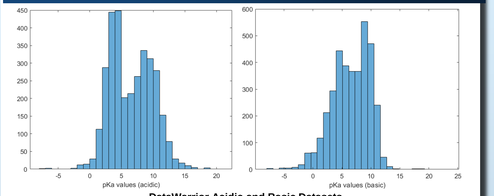
Predictions	ACD/Labs		ChemAxon		OPERA		DNN		XGB	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
ACD/Labs	1	0	0.47	4.48	0.60	3.46	0.34	4.95	0.23	5.36
ChemAxon	0.60	4.48	1	0	0.52	4.55	0.45	5.41	0.30	6.09
OPERA	*	*	*	*	1	0	0.51	2.09	0.44	2.27
DNN	*	*	*	*	0.74	2.09	1	0	0.51	2.39
XGB	*	*	*	*	0.43	2.27	0.15	2.39	1	0

Comparison of All Models for the Basic pKa predictions

Predictions	ACD		ChemAxon		SVM		DNN		XGB	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
ACD/Labs	1	0	0.48	2.88	-0.14	4.57	-0.62	5.67	-0.80	5.99
ChemAxon	0.61	2.88	1	0	0.02	5.66	-2.62	9.77	-2.36	9.42
SVM	*	*	*	*	1	0	-0.90	3.41	0.37	1.97
DNN	*	*	*	*	0.15	3.41	1	0	0.35	2.99
XGB	*	*	*	*	0.28	1.97	-0.49	2.99	1	0

* Our models are not used as refence to evaluate ChemAxon and ACD/Labs predictions.

Range of Predictions and Limitations



DataWarrior Acidic and Basic Datasets

- pKa predictions generated by our models are ranging between about -5 and 15 for both the acidic and basic datasets.
- The narrow predictions range of our models (in comparison with the two commercial tools) is certainly linked to DataWarrior data that has the same range as shown by the distribution of its acidic and basic pKa values (histograms above).

The different ranges in pKa predictions may also explain why:

- The disagreement between our models and the commercial models on the benchmark dataset (TSCA actives) is higher for the basic pKa predictions. This is particularly noticeable with ChemAxon, which generated a high number of predictions of pKas lower than -5 for the basic data set.
- For the TSCA-actives list, the divergence between ACD/Labs and ChemAxon is higher for the basic pKa predictions compared to the acidic pKa predictions. Interestingly, this is the opposite of what occurred for the DataWarrior dataset.

The predictions of our models can be considered more accurate in the range of -5 to 15 for both the acidic and basic pKas.

Summary and Next Steps

- An automated QSAR data preparation workflow was applied to a public data set of 7912 chemicals, created three data subsets, Acidic, Basic and Combined. Model performance was evaluated using all data subsets with the DNN, SVM and XGB algorithms.
- The best models were compared and benchmarked with two commercial predictors showing different levels of concordance.
- The models and source codes will be available for download and use.
- This modeling effort will help provide predicted pKa values for all ionizable chemicals in the EPA DSSTox database.
  - Predictions will be available on the EPA's CompTox Chemistry Dashboard (https://comptox.epa.gov)
  - Predictions will also be used by the NICEATM's Integrated Chemical Environment (ICE) Dashboard (https://ice.ntp.niehs.nih.gov/) in various pharmacokinetic calculations.

Acknowledgements

We thank Caroline Stevens at EPA/ORD (Athens, GA) for providing ChemAxon predictions and Catherine Sprankle, ILS, for editing the poster text.

The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contract HHSN273201500010C.

The views expressed above do not necessarily represent the official positions of any federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.

References

[1] Mansouri et al. EHP, 2016. (doi:10.1289/ehp.1510267)

[2] Mansouri et al. J. Cheminfo., 2018 (doi:10.1186/s13321-018-0263-1)

Subscribe to the NICEATM News Email List

To get announcements of NICEATM activities, visit the NIH mailing list page for NICEATM News at https://list.nih.gov/cgi-bin/wa.exe?SUBED1=niceatm-l&A=1 and click "Subscribe."