Prediction of pKa from chemical structure using free and open-source tools

Valery Tkachenko³
Alex Korotcov³
Neal Cariello¹
Kamel Mansouri⁴
Antony Williams²

1. Integrated Laboratory Systems, Research Triangle Park, North Carolina, United States
2. National Center for Computational Toxicology, US-EPA, North Carolina, United States
3. Science Data Software, LLC, Rockville, MD 20850
4. ScitoVation, Research Triangle Park, North Carolina, United States

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA
pKa
What Is It?
Why Is It Important?
How Can We Use It?
Absorption Of Chemicals Into Cells

**Question:** Which kind of chemicals will preferentially partition into the plasma membrane, charged or uncharged (ionized or non-ionized)?

- In general, chemicals that make it into the plasma membrane (lipid bilayer) have a better chance of getting into the cell.
- The plasma membrane facing the extracellular space is hydrophobic and lipophilic so will charged or uncharged molecules cross the membrane best?
What Is pKa?

- pKa is a property that tells us how acidic (or basic) a chemical is. The lower the pKa the stronger the acid.

- The pKa influences the protonation state (charged or uncharged) of the chemical in solution at a given pH value.

![Diagram showing the relationship between pH of solution and dissociation status of acetic acid.](image)

The pKa of acetic acid is 4.75.

At the pH of the pKa, the chemical is 50% ionized.
• $K_a$ is the acid dissociation constant which is a measure of the strength of an acid in solution.

• $K_a$ is an equilibrium constant and $pK_a$ is the $-\log_{10}$ value of $K_a$, therefore for acids

$$HA \leftrightarrow A^- + H^+$$

$$K_A = \frac{[A^-] [H^+]}{[HA]}$$

$$pK_a = -\log_{10} \frac{[A^-] [H^+]}{[HA]}$$
**pKa Importance**

- pKa values reflect the **ionization state** of a chemical

- Why is this important?
  - Ionization affects lipophilicity, solubility, protein binding and the ability of a chemical to cross the plasma membrane
  - This affects ADMET

- pKa can be used, and is many times required, for
  - Physiologically Based Pharmacokinetic (PBPK) modeling
  - In Vitro To In Vivo Extrapolation (IVIVE)
  - Prediction of tissue:plasma partition coefficients
Using Open Source Software and Data to Build a pKa Prediction Algorithm:

Data Quality, Algorithm Development and Applications
Good Cheminformatics Data Is Hard To Obtain, Especially $p$Ka

- Obtaining high-quality data sets is difficult
- Curation is generally VERY time-consuming without optimized workflows
- Many issues exist with available datasets

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling

K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

Pages 911-937 | Received 03 Sep 2016, Accepted 24 Oct 2016, Published online: 25 Nov 2016
7912 Chemicals With pKa In Water Are Available From The Datawarrior Website

- This is not a widely-known dataset
- Datawarrior didn’t list the references for the data
- We checked ~60 DataWarrior chemicals against literature and the results were good (< 0.3 pKa units difference between DataWarrior and the literature)
Dataset Has A Bimodal Distribution

- 7912 structures
  - 3614 acidic
  - 4298 basic
QSAR-ready Workflow

- Remove inorganics and mixtures
- Clean salts and counterions
- Normalize Nitros & tautomers
- Remove duplicates
- Final inspection

KNIME workflow
QSAR-ready analysis

Full dataset:
- 7904 QSAR-ready structures
- 6245 unique QSAR-ready structures

Acidic dataset
- 3610 QSAR-ready total structures
- 3260 unique QSAR-ready structures

Basic dataset
- 4294 QSAR-ready total structures
- 3680 unique QSAR-ready structures

1659 Duplicate structures!

Standard deviation of duplicates

2 as a threshold for averaging?
Modeling Options
To deal with complexity of multiple pKa's for a chemical, three datasets were produced and analyzed:

**Option 1:** Only structures with a unique pKa value were used.
- Pre-categorized Acidic dataset: 2960
- Pre-categorized Basic dataset: 3158
- Combined: 4897 (no amphoteric)

**Option 2:** A unique value/structure (average value if stdDev<2)
- Pre-categorized Acidic dataset: 3095
- Pre-categorized Basic dataset: 3370
- Combined: 5263 (no amphoteric)

**Option 3:** The entire list of QSAR-ready chemicals was used with averaging for similar pKa values.
- if stdDev <= 1: Average value;
- if stdDev > 1: strongest pKa (min acidic/max basic)
- Acidic dataset: 3260 unique QSAR-ready structures
- Basic dataset: 3680 unique QSAR-ready structures
**Machine Learning And Predicting pKa**

- The term Machine Learning was coined in 1959
  Machine learning explores the study and construction of algorithms that can learn from and make predictions on data through building a model from sample inputs.

- Each chemical with a pKa produces ~16.5K data points in 12 datasets

- We need to find the best combination of variables (columns) for pKa prediction

<table>
<thead>
<tr>
<th>basicOrAcidic</th>
<th>pKa</th>
<th>nAcid</th>
<th>ALOGP</th>
<th>ALOGP2</th>
<th>AMR</th>
<th>apol</th>
<th>naAromAtom</th>
<th>FP5</th>
<th>FP6</th>
<th>FP7</th>
<th>FP8</th>
<th>FP9</th>
<th>FP10</th>
<th>SubFPC1</th>
<th>SubFPC2</th>
<th>SubFPC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>acidic</td>
<td>-6.56</td>
<td>1</td>
<td>0.1005</td>
<td>0.0101</td>
<td>16.4215</td>
<td>22.96034</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-1.9</td>
<td>0</td>
<td>0.511</td>
<td>0.261121</td>
<td>8.6916</td>
<td>24.50755</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-1.7</td>
<td>0</td>
<td>-0.8807</td>
<td>0.775632</td>
<td>5.8646</td>
<td>19.32276</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-1.37</td>
<td>0</td>
<td>-0.4603</td>
<td>0.211876</td>
<td>2.5646</td>
<td>21.08276</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-0.9</td>
<td>0</td>
<td>-3.0181</td>
<td>9.108928</td>
<td>41.1246</td>
<td>33.90352</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-0.48</td>
<td>1</td>
<td>-2.144</td>
<td>4.596736</td>
<td>9.7839</td>
<td>18.06676</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>-0.1</td>
<td>0</td>
<td>-0.4603</td>
<td>0.211876</td>
<td>2.5646</td>
<td>12.70917</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>0.2</td>
<td>0</td>
<td>-0.1759</td>
<td>0.030941</td>
<td>16.215</td>
<td>24.97834</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>acidic</td>
<td>0.31</td>
<td>0</td>
<td>-0.4603</td>
<td>0.211876</td>
<td>2.5646</td>
<td>22.40955</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

1D/2D descriptors | binary fingerprints | count substructures
Train and test sets for modeling:

For each one of the data options:

• Split into training (75%) and test (25%)
• Keep similar distribution of pKa values
• Keep similar distribution of acidic and basic pKas for combined datasets
• Descriptors (and fingerprints) are generated for all QSAR-ready structures and can be matched by the generic ID (integers)

A classification model to determine if a molecule will have an **acidic pKa**, **basic pKa** or **both** is trained too.
Training Models

• Create model and estimate performance using only the training dataset
• 5-fold cross-validation was used for training, model performance evaluation and tuning
• Root mean squared error (RMSE) was used as a performance metric for training optimization.
• Choice of machine learning methods:
  • Extreme Gradient Boosting (XGBoost), the advanced traditional (shallow) machine learning (SML) method.
  • Deep Neural Network (DNN), a deep machine learning method (DML).
  • Support Vector Machines (SVM): defines a decision boundary that optimally separates two classes by maximizing the distance between them.
XGBoost training method

• XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
• Coding was done using R. The caret and xgboost packages were used for all analysis.
• RMSE was a metric to be minimized. 5-fold cross validation was used to train the model using the training dataset.
• Highly correlated variables were removed using caret::findCorrelation with a cutoff of 0.90.
• Low variance variables were removed using caret::nearZeroVar with a cutoff of 95/5.
• The following data subsets were modeled using all binary fingerprints:
  • Remove variables that are all 0's (many) and all 1's (few).
  • As above with removal of highly correlated variables.
  • As above with removal of near zero variance variables.
XGBoost training results

- Performance using the basic dataset was substantially better than the acidic dataset
- MACCS and FP (Morgan’s, 1024 bins) binary fingerprints generally gave the best performance

Best RMSE and R-Squared are:
- **Basic** pKa: 1.585 and 0.765
- **Acidic** pKa: 1.737 and 0.737
DNN training method

- The following Deep Neural Network parameters were optimized: optimization algorithm, weight initialization, hidden layers activation function, L2 regularization, dropout regularization, number of hidden layers and nodes in the hidden layers, and learning rate. Keras (https://keras.io/) and Tensorflow (www.tensorflow.org) were used for deep learning models training.
- The final DNN: 3 hidden layers of 256 nodes each followed by a batch normalization and a drop out layer to generalize trained models.
- 5-fold cross validation on training data using mean square error as a loss function with earlier training stopping base on validation loss, thus further improving of the models’ generalization.
DNN training results

• Performance using the acidic dataset was substantially better than the basic dataset, and slightly outperforming XGBoost models

• Combination of RDKit Descriptors+MACCS+FCFC(512 bins, radius 3)+Avalon(512 bins), PADEL continues descriptors+MACCS, and MACCS or MACCS+FP (Morgan’s, 1024 bins) gave the best DNN models performance

Best test RMSE and R-Squared are:

• Basic pKa: 1.506 and 0.789
• Acidic pKa: 1.519 and 0.798
SVM training method

- Used the free and open source package LibSVM3.1 (Chang and Lin 2001).
- Originally designed to solve classification problems then generalized to fit continuous models as well.
- Its algorithm defines a decision boundary that optimally separates two classes by maximizing the distance between them.
- The decision boundary can be described as an hyperplane that is expressed in terms of a linear combination of functions parametrized by support vectors, which consist in a subset of training molecules.
- SVM algorithms search for the support vectors that give the best separating hyperplane using a kernel function.

SVM kernel function maximizing the margin between the classes.
## SVM training results

### Results of Option 1

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>5f CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R2</td>
<td>RMSE</td>
<td>Q2</td>
</tr>
<tr>
<td>Acidic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>870</td>
<td>0.96</td>
<td>0.65</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1548</td>
<td>0.91</td>
<td>1.04</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>556</td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2104</td>
<td>0.94</td>
<td>0.8</td>
</tr>
<tr>
<td>Fingerprint - Continuous</td>
<td>2418</td>
<td>0.99</td>
<td>0.11</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>876</td>
<td>0.96</td>
<td>0.64</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1535</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>544</td>
<td>0.9</td>
<td>1.05</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2079</td>
<td>0.93</td>
<td>0.87</td>
</tr>
</tbody>
</table>

---

### Results of Option 2

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>5f CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R2</td>
<td>RMSE</td>
<td>Q2</td>
</tr>
<tr>
<td>Acidic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>913</td>
<td>0.98</td>
<td>0.49</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1552</td>
<td>0.9</td>
<td>1.05</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>589</td>
<td>0.9</td>
<td>1.09</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2141</td>
<td>0.94</td>
<td>0.85</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>913</td>
<td>0.97</td>
<td>0.52</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1534</td>
<td>0.9</td>
<td>1.02</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>551</td>
<td>0.9</td>
<td>1.02</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2085</td>
<td>0.93</td>
<td>0.88</td>
</tr>
</tbody>
</table>

---

### Results of Option 3

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>5f CV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R2</td>
<td>RMSE</td>
<td>Q2</td>
</tr>
<tr>
<td>Acidic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>510</td>
<td>0.96</td>
<td>0.66</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1580</td>
<td>0.91</td>
<td>1.04</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>815</td>
<td>0.88</td>
<td>1.19</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2395</td>
<td>0.93</td>
<td>0.86</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>510</td>
<td>0.95</td>
<td>0.75</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>1543</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>815</td>
<td>0.89</td>
<td>1.06</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>2358</td>
<td>0.93</td>
<td>0.84</td>
</tr>
</tbody>
</table>

---

### kNN and SVM classification models

<table>
<thead>
<tr>
<th></th>
<th>kNN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>5f CV</td>
</tr>
<tr>
<td></td>
<td>BA</td>
<td>BA</td>
</tr>
<tr>
<td>Continuous</td>
<td>15</td>
<td>0.8</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>15</td>
<td>0.92</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>511</td>
<td>0.98</td>
</tr>
<tr>
<td>Fingerprint Count</td>
<td>1565</td>
<td>0.98</td>
</tr>
<tr>
<td>Fingerprint – FP Count</td>
<td>815</td>
<td>0.96</td>
</tr>
</tbody>
</table>

---

These models are used to decide if a test chemical has an acidic pKa, basic pKa, or both (amphoteric).
Future Work

• Predict pKa values for all ionizable chemicals in the EPA CompTox Chemistry Dashboard (https://comptox.epa.gov)

• Develop web service for pKa prediction – used for calculation on the fly when registering new chemicals

• Integrate web service into online systems: e.g. the CompTox Chemistry Dashboard to allow for real time prediction of pKa values (https://comptox.epa.gov/dashboard/predictions/index)
Summary

• 7912 Chemicals With pKa In Water were scrapped from the public Datawarrior Website: http://www.openmolecules.org/datawarrior/
• Automated QSAR data preparation workflow was developed. Three different options of automated split into Acidic, Basic, and Combined sub-sets was developed and tested.
• A classification model to determine if a molecule will have an acidic pKa, basic pKa or both was trained. Will be used for prediction workflow in a dashboard.
• XGBoost models for pKa predictions were trained. MACCS and FP (Morgan’s, 1024 bins) binary fingerprints gave the best performance with the following best RMSE and R-Squared are: basic pKa: 1.585 and 0.765; acidic pKa: 1.737 and 0.737.
• The DNN exhibited very good performance and generalization characteristics. The best performance with the following best RMSE and R-Squared are: basic pKa: 1.506 and 0.789; acidic pKa: 1.519 and 0.798.
• For SVM: the results for the acidic dataset reached an R2 test of 0.76 and for the basic dataset, an R2 test of 0.78.