Open Source pK_a Predictions

Marcel Baltruschat and Paul Czodrowski
Faculty of Chemistry and Chemical Biology, TU Dortmund University

technische universität dortmund
fakultät für chemie und chemische biologie

Introduction

The acid-base dissociation constant (pK_a) of a drug has a far-reaching influence on pharmacokinetics by altering the solubility, membrane permeability and protein binding affinity of the drug. Therefore, the pK_a values of drugs must be determined for approval by regulatory authorities. That's why high quality pK_a prediction methods are very important within the drug discovery process. However, there is no publicly available, open source and license-free pK_a prediction tool that can reach the quality of licensed programs like *MoKa*[1] or *Marvin*[2]. Our goal is to develop a new, highly accurate pK_a prediction tool based on freely accessible data and free to use for everyone.[3,4]

Data Basis

There are several freely accessible databases available, each containing hundreds to thousands of pK_a values. Those have to be curated and cleaned in a consistent manner to retrieve a combined dataset suitable for machine learning.

Database	pK_a Values	Unique Valid Structures
ChEMBL25	8132	6384
DataWarrior	7913	7268
Reaxys*	44125	37671

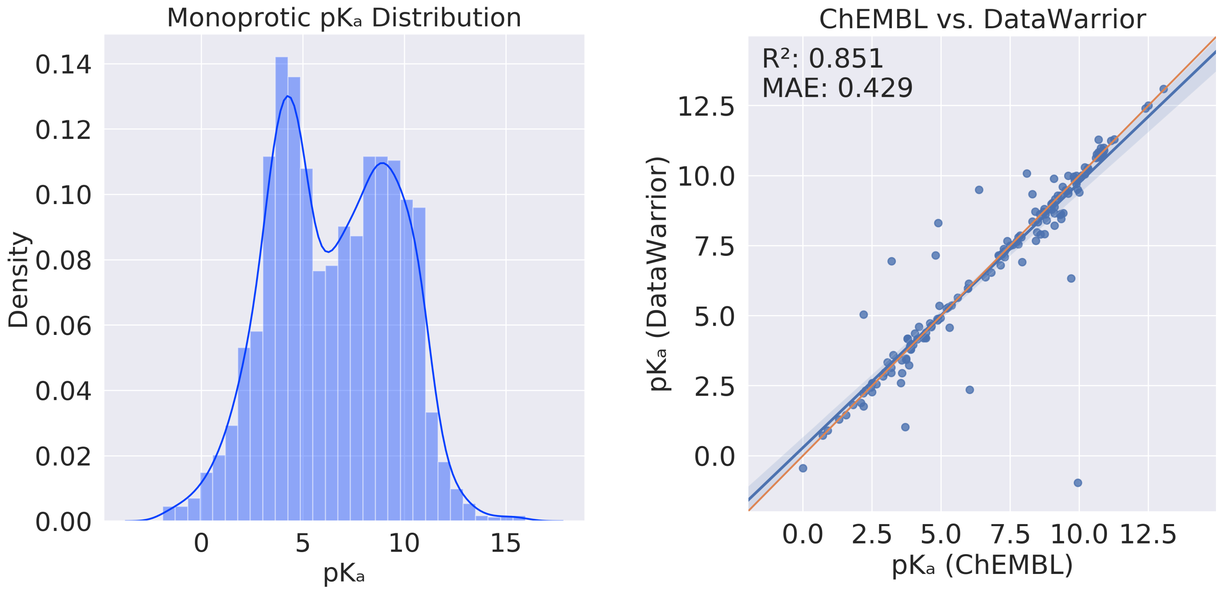
*Not included in further analyses

The following cleaning steps were applied to prepare the datasets:

- Filter by Lipinski's rule of five
- Keep only data points between -2 and 16
- Remove all salts from molecules
- Uncharge molecules if possible
- Convert pK_b to pK_a
- Combine data points from duplicated structures while removing outliers

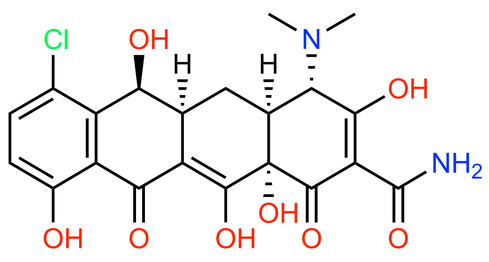
After all cleaning and filtering steps, all monoprotic structures, i.e. structures containing only one titratable group, were extracted.

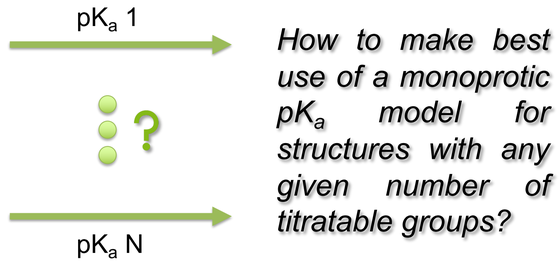
Monoprotic pK_a



In total there are **3939** unique monoprotic structures in the combined ChEMBL25 and DataWarrior dataset. Unfortunately, no reference is available for the DataWarrior dataset. Therefore, we compared the two datasets to validate this dataset. The results show that the values correlate well, but there are also a few outliers.

Working with monoprotic structures makes it relatively easy to develop a machine learning model which predicts the pK_a value of a single titratable group. Based on such a model, how can one apply it to predict pK_a values of multiprotic structures?





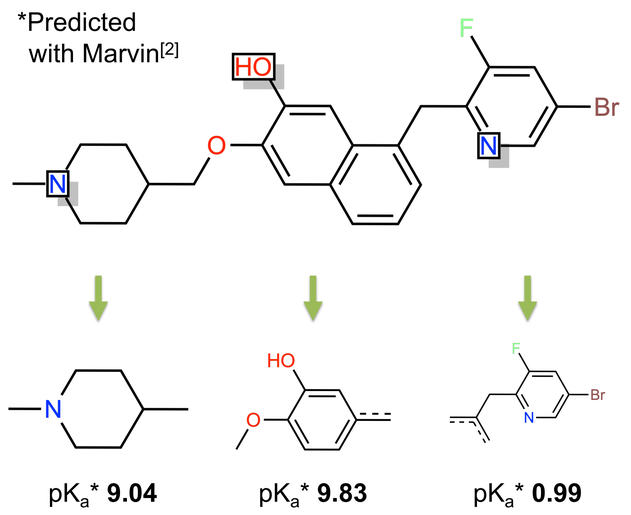
How to make best use of a monoprotic pK_a model for structures with any given number of titratable groups?

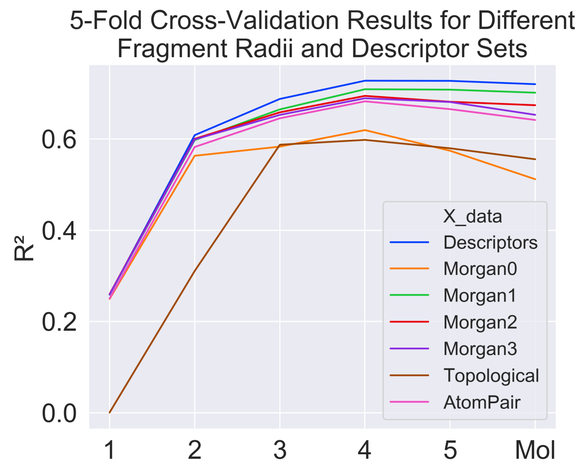
References

[1] MoKa, Molecular Discovery Ltd. Borehamwood, United Kingdom
[2] Marvin 19.15.0, 2019, ChemAxon Ltd, *http://www.chemaxon.com*
[3] Charifson, P. S., & Walters, W. P. (2014). Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry*
[4] Manallack, D. T. (2007). The pKa Distribution of Drugs: Application to Drug Discovery. *Perspectives in Medicinal Chemistry*
[5] Richard A. Lewis, Stephane Rodde, Novartis Pharma AG, Basel, Switzerland

Fragmentation Concept

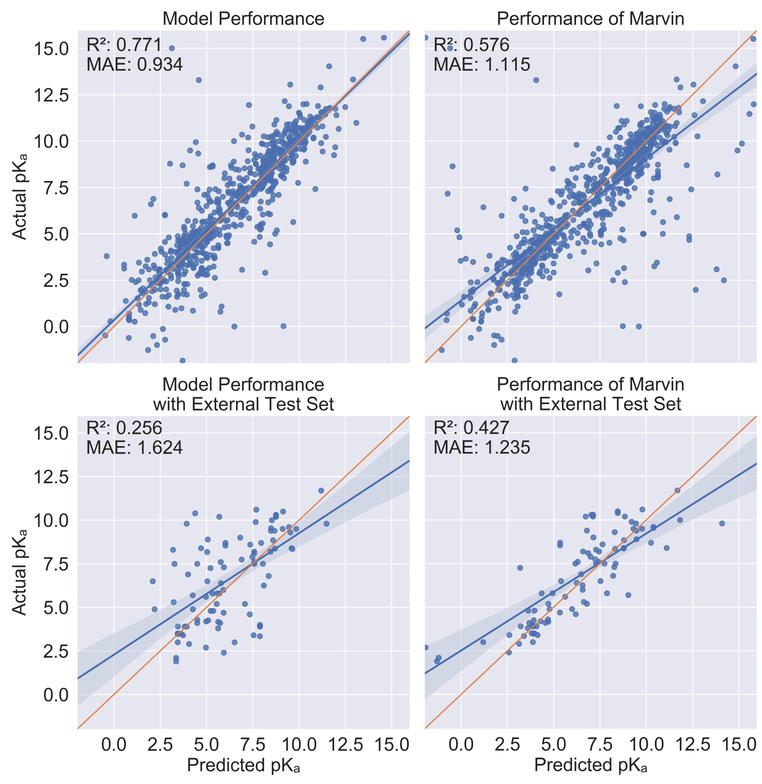
To be able to predict the pK_a values of a molecule with any given number of titratable groups, the individual groups including a defined binding radius are extracted. Such an approach allows a 1 to 1 structure-value relationship between substructure and pK_a. The model is then trained with the resulting fragments.
By cutting out fragments, structural information that is important to determine the pK_a value can be lost. This makes the choice of the radius for extraction crucial. To determine the optimal radius, we performed a 5-fold cross validation with Random Forest for all binding radii from 1 to 5 and 7 different descriptor sets over the monoprotic dataset.





Validation

The test of several machine learning algorithms lead us to the decision to use a multilayer perceptron (MLP) with two hidden layers of 500 neurons each. For the training the fragments with radius=4 were used, for which 196 RDKit descriptors were calculated. The monoprotic dataset was divided into 80% training data and 20% test data. To further validate the model quality, an external data set provided by Novartis[5] was used.
For the test data, the MLP shows significantly better performance (R^2=0.77) than the prediction of Marvin[2] (R^2=0.58). However, comparing the results obtained with the external test dataset provided by Novartis[5], the outcome is completely different. This suggests that the data from ChEMBL25 and DataWarrior may have annotation errors, which leads to wrong predictions after the training. The data basis must be examined more closely and the overall quality of the model has to be improved.



Future Work

- In-depth analysis of ChEMBL25, DataWarrior, Novartis and Reaxys data
- Development of suitable data correction mechanisms
- Improvement of model architecture and quality
- Development of a method for recognition and localization of titratable groups
- Validation of the fragmentation concept for structures with any given number of titratable groups