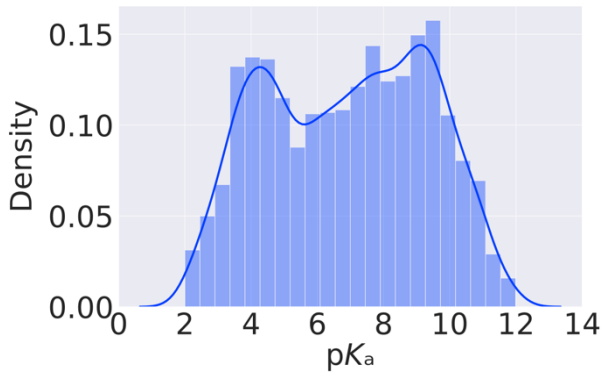
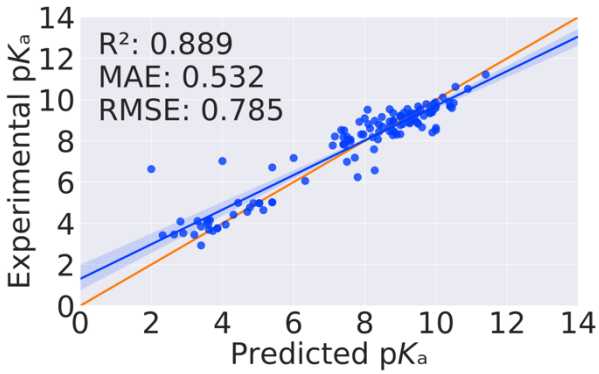
Machine learning meets pK_a

Marcel Baltruschat, Paul Czodrowski

TU Dortmund University, Faculty of Chemistry and Chemical Biology

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 [1,2]. 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* [3] or *Marvin* [4]. Our goal is to develop a new, highly accurate pK_a prediction tool based on freely accessible data and free to use for everyone.





To achieve our goal, we identified a dataset from *DataWarrior*, which contains 7913 pK_a values measured in water. In addition, we could extract 8111 values from the ChEMBL25 database. The values were then preprocessed, filtering out all molecules according to Lipinski's rule of five (one violation allowed), unwanted atoms and bad functional groups. Additionally, any salts were removed and tautomer canonization was performed with *QUACPAC* [5]. Only structures that are considered monoprotic in the pH range of 2 to 12 (determined by *Marvin* [4]) were retained. After preprocessing the structures, multiple measurements were combined and any outliers were removed. The final result was a curated monoprotic training dataset with 5994 unique structures. The distribution of the pK_a values is shown in the left figure.

This training dataset was used to test a total of seven different machine learning configurations with the basic regressors Random Forest, Support Vector Regression, Multilayer Perceptron and XGradientBoost. In addition, six different descriptor/fingerprint sets were examined for each configuration, resulting in a total number of 42 trained and evaluated machine learning models. The models were all 5-fold cross-validated and evaluated using two different external test datasets, one dataset taken from the scientific literature and another dataset provided by Novartis [6]. The predictive quality of the best-performing model as measured by the external test dataset from the literature is shown in the right figure.

In further work all commercial tools used for preprocessing will be replaced by open source applications and the prediction capabilities will be extended to multiprotic molecules.

[1] Charifson, P. S., & Walters, W. P. (2014). Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry*

[2] Manallack, D. T. (2007). The pKa Distribution of Drugs: Application to Drug Discovery. *Perspectives in Medicinal Chemistry*

[3] MoKa, Molecular Discovery Ltd. Borehamwood, United Kingdom

[4] Marvin 20.1.0, 2020, ChemAxon Ltd, *http://www.chemaxon.com*

[5] QUACPAC 2.0.2.2: OpenEye Scientific Software, Santa Fe, NM. *http://www.eyesopen.com*

[6] Richard A. Lewis, Stephane Rodde, Novartis Pharma AG, Basel, Switzerland