The Challenge of Predicting pKa Values of Organic Compounds. Part I – Data Curation

Anushka Sen[1], Dragos Horvath[1]

[1]UMR 7140, University of Strasbourg, 4 Rue Blaise Pascal, 67000 Strasbourg, France

Abstract Text:

Accurate prediction of protonation states is essential for understanding drug pharmacodynamics and pharmacokinetics. However, most chemoinformatics methods rely on substructure-based approaches that assign default pKa values, neglecting remote structural effects such as inductive, resonance, and through-space interactions. Although large experimental datasets exist, pKa values are rarely linked to specific functional groups or tautomeric microstates, limiting machine learning approaches to monofunctional molecules and hindering modeling of multiple ionizable sites.

In this work, this challenge is addressed through a dedicated data curation pipeline that maps experimental pKa values to plausible protonation microstates. Curation faces a two-fold challenge in trying to assign an experimental pKa value to a microspecies: first, deciding which atom (to be explicitly marked in the SMILES in which tautomeric form, if applicable) is the one losing its proton at that pH level, and second specifying the current state of the other ionizable groups. Initial assignments are guided by rule-based estimates of typical pKa ranges in enumerated tautomeric forms, followed by expert-driven refinement for ambiguous cases.

Initial data curation of the IUPAC digitized pKa dataset[1] merged with Czodrowski's dataset[2] led to a set of ~16K pKa values associated with >98K marked SMILES. Since manual assessment of this data volume is unfeasible, a model/predict/check-outliers loop was implemented to address inconsistencies. At each iteration, unambiguous assignments (~8K pKa values associated with a single marked SMILES) were used to train Support Vector Regression models based on ISIDA marked-atom descriptors. These models predicted pKa values for all hypothetical microspecies. For each experimental pKa, the microspecies with the closest predicted value was selected; if the deviation exceeded 2 pKa units, the case was flagged for inspection. Unrealistic assignments were corrected, and after five iterations, no major errors remained among the worst predictions.

Eventually, the latest generation of preliminary models served to prune multiple candidate assignments: only the marked SMILES with predicted pKa closest to experiment were retained (top 5 best, extended to top 50 or until prediction error > 3 pKa units). This resulted in 15,189 curated experimental pKa values associated with 45,125 marked SMILES, of which 8,880 are unambiguously assigned.

As a next step, a multi-instance/multi-task learning framework will be developed, grouping candidate microstates into "bags" to identify the most probable contributor to each pKa. Additionally, quantum-mechanics-based descriptors of marked atoms will be used as auxiliary endpoints learned alongside experimental pKa values from the 2D molecular graph.

Bibliography **:**

[1] IUPAC Dissociation Constants Dataset. https://github.com › IUPAC › Dissociation-Constants
[2] Czodrowski Lab, *Machine Learning Meets pKa*. https://github.com/czodrowskilab/Machine-learning-meets-pKa/blob/master/datasets/combined_training_datasets_unique.sdf