ChemAxon pKa Predictor – IUPAC pKa Dataset Benchmark

Why this benchmark

The evaluated IUPAC pKa dataset originates from ~24k records, which after curation, structural normalization, and duplication filtering resulted in ~10.6k unique molecules. The source data is heterogenous. It mixes several pKa-types, contains numerous ambiguous labels, and in many cases does not clearly indicate which protonation step a reported pKa corresponds to.

Key limitation of the source data

The pKa types in the original series (pKa, pKaH, pKb, etc.) are not consistently defined. The authors themselves note that pKa types were often assigned automatically, based on numerical ordering, rather than stated experimentally. Consequently:

- ~1,916 molecules had ambiguous protonation labels and were excluded.
- In many remaining entries, the pKa type is not fully aligned with the underlying chemical process.

This matters because our model returns Strongest Acidic pKa 1-n (A1, A2, ..., An) and Strongest Basic pKa 1-m (B1, B2, ..., Bm) values using a chemically consistent, structure-based definition.

Data set processing

Environment

- KNIME v5.4.0
- JChem Extension v25.3.4

Approach 1: Mapping IUPAC data to our A1/A2/B1/B2 system

We retained only the most consistent subset: pKa1, pKa2, pKaH1, pKaH2 with Reliable & Approximate assessment. This allowed a meaningful comparison for ~7k molecules. The reported experimental pKa values were assigned to the individual molecules in the dataset by types.

→ *Considered pKa-types / assessment: pKa1, pKa2, pKaH1, pKaH2 / reliable, approximate*

Statistical metrics

n = 7,119	**Pearson**	**R^2**	**MAE**	**RMSE**
V25.5.0	0.899	0.791	0.810	1.378

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	2,334	33%
$0.25 \leq \Delta < 0.5$	1,494	21%
$0.5 \leq \Delta < 1$	1,553	22%
$1 \leq \Delta \leq 2$	1,085	15%
$\Delta > 2$	653	9%

Approach 2: Type-agonistic alignment

We ignored pKa types entirely due to their inconsistencies. We assigned each experimental pKa to the closest predicted value across A1/A2/B1/B2. For the sake of comparison, I also created a similar type of filter.

→ *Considered pKa-types / assessment: pKa1, pKa2, pKaH1, pKaH2 / reliable, approximate*

Statistical metrics

n = 6,856	**Pearson**	**R^2**	**MAE**	**RMSE**
V25.5.0	0.897	0.786	0.829	1.401

Ranged delta distribution

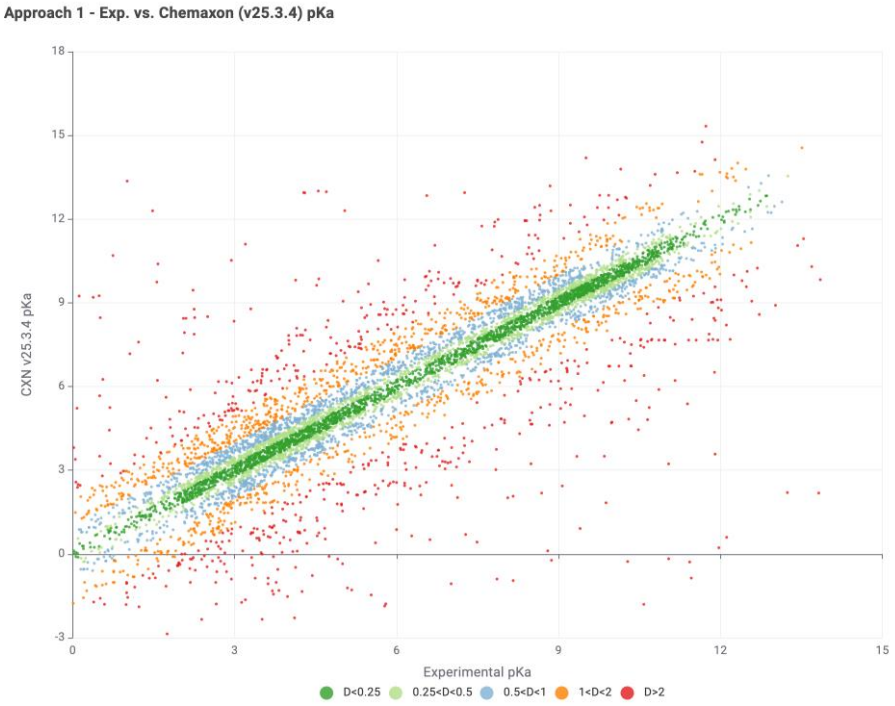
	Occurrence	
$\Delta < 0.25$	2,313	34%
$0.25 \leq \Delta < 0.5$	1,483	22%
$0.5 \leq \Delta < 1$	1,560	23%
$1 \leq \Delta \leq 2$	1,082	16%
$\Delta > 2$	440	6%

Bottom line

- **Approach 1 and Approach 2 yield nearly identical metrics** (R^2 0.79 vs. 0.79, MAE 0.81 vs. 0.83). The model performance is not an artifact of label mapping.
- The **Chemaxon A1/A2/B1/B2 framework is robust against label heterogeneity** in the source. It achieves comparable accuracy whether the IUPAC pKa / pKaH annotations are used explicitly (Approach 1) or ignored (Approach 2). This robustness matters in practice on large public references, where annotation quality varies per entry.

Scatter Plots

Approach 1 (pKa1-2/pKaH1-2, reliable/approximate)



Approach 2 (pKa1-2/pKaH1-2, reliable/approximate)

