ChemAxon pKa Predictor – AstraZeneca Curated Benchmark

Why this benchmark

pKa determines protonation state at physiological pH and influences solubility, permeability, ADME, and target engagement. To validate the latest Chemaxon pKa model on pharma-relevant chemistry, we tested it against AstraZeneca's curated pKa datasets in ChEMBL. These industrially relevant, peer-visible references cover diverse amines, nitrogen heterocycles, multi-protonation-center scaffolds, and a broad range of acidic functional groups.

Data set processing

Environment

- cxcalc CLI
- v25.5.0

① Most basic pKa (B1) — CHEMBL3301362

Diverse amines, nitrogen heterocycles, scaffolds with multiple competing protonation centers.

Statistical metrics

n = 247	Pearson	R^2	MAE	RMSE
v25.5.0	0.883	0.731	0.721	1.107

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	67	27%
$0.25 \leq \Delta < 0.5$	61	25%
$0.5 \leq \Delta < 1$	64	26%
$1 \leq \Delta \leq 2$	41	17%
$\Delta > 2$	14	6%

② Most acidic pKa (A1) — CHEMBL3301375

Weak-to-strong acids covering carboxylic acids, sulfonamides, phenols, and NH-acidic heterocycles.

Statistical metrics

n = 196	Pearson	R^2	MAE	RMSE
v25.5.0	0.905	0.793	0.761	1.053

Ranged delta distribution

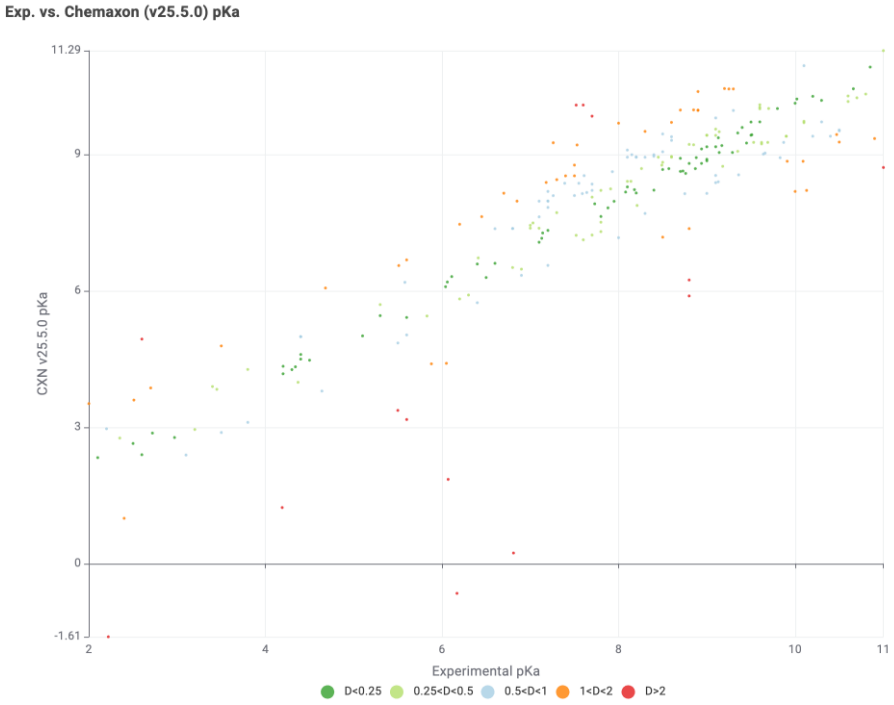
	Occurrence	
$\Delta < 0.25$	47	24%
$0.25 \leq \Delta < 0.5$	45	23%
$0.5 \leq \Delta < 1$	49	25%
$1 \leq \Delta \leq 2$	42	21%
$\Delta > 2$	13	7%

Bottom line

- **Balanced reliability** across acidic and basic centers, same predictor, same workflow.
- **Strong correlation** (Pearson 0.88–0.91) on industrially curated chemistry, not a pre-filtered showcase set.
- **Sub-unit mean-error** in both domains (MAE 0.72–0.76); fit for purpose for ranking analogs, filtering virtual libraries, prioritizing synthesis, and feeding ADME/formulation models.

Scatter Plots

Most basic pKa (B1)



Most Acidic pKa (A1)

