ChemAxon pKa Predictor – Published Experimental Benchmarks

Why these benchmarks

A pKa model is only as trustworthy as its behavior across independently assembled datasets. We benchmarked the latest ChemAxon pKa model against three widely cited public collections. The same datasets used in the open source pKa modeling study by Mansouri et al. Together they span pharma in-house curation, expert literature compilation, and large-scale public aggregation.

Data set processing

Environment

- cxcalc CLI
- v25.5.0

① Novartis curated dataset

In-house pharma reference, small but rigorously curated. 280 monoprotic compounds in total.

→ *Considered subset: Basic compounds (168)*

Statistical metrics

n = 168	Pearson	R^2	MAE	RMSE
v25.5.0	0.913	0.761	0.779	1.037

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	39	23%
$0.25 \leq \Delta < 0.5$	28	17%
$0.5 \leq \Delta < 1$	55	33%
$1 \leq \Delta \leq 2$	39	23%
$\Delta > 2$	7	4%

→ *Considered subset: Acidic compounds (112)*

Statistical metrics

n = 112	Pearson	R^2	MAE	RMSE
v25.5.0	0.877	0.757	0.875	1.259

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	35	31%
$0.25 \leq \Delta < 0.5$	17	15%
$0.5 \leq \Delta < 1$	22	20%
$1 \leq \Delta \leq 2$	22	20%
$\Delta > 2$	15	13%

② AvLiLuMoVe dataset from five literature sources

Diverse measurement protocols across five sources: Avdeef (135), Liao (113), Luan (66), Morgenthaler (174), and Vertex (124).

→ *Considered subset: Basic compounds (413)*

Statistical metrics

n = 413	Pearson	R^2	MAE	RMSE
v25.5.0	0.945	0.886	0.446	0.679

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	175	42%
$0.25 \leq \Delta < 0.5$	112	27%
$0.5 \leq \Delta < 1$	88	21%
$1 \leq \Delta \leq 2$	26	6%
$\Delta > 2$	12	3%

→ *Considered subset: Acidic compounds (133)*

Statistical metrics

n = 133	Pearson	R^2	MAE	RMSE
v25.5.0	0.952	0.901	0.564	0.772

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	50	38%
$0.25 \leq \Delta < 0.5$	28	21%
$0.5 \leq \Delta < 1$	32	24%
$1 \leq \Delta \leq 2$	17	13%
$\Delta > 2$	6	5%

③ ChEMBL-DataWarrior combined unique dataset

Large-scale public aggregation combines unique molecules from ChEMBL and DataWarrior.

→ *Considered subset: Basic compounds (3,586)*

Statistical metrics

n = 3,586	Pearson	R^2	MAE	RMSE
v25.5.0	0.891	0.766	0.779	1.086

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	946	26%
$0.25 \leq \Delta < 0.5$	765	21%
$0.5 \leq \Delta < 1$	911	25%
$1 \leq \Delta \leq 2$	664	19%
$\Delta > 2$	300	8%

→ *Considered subset: Acidic compounds (2,405)*

Statistical metrics

n = 2,405	Pearson	R^2	MAE	RMSE
v25.5.0	0.876	0.730	0.868	1.343

Ranged delta distribution

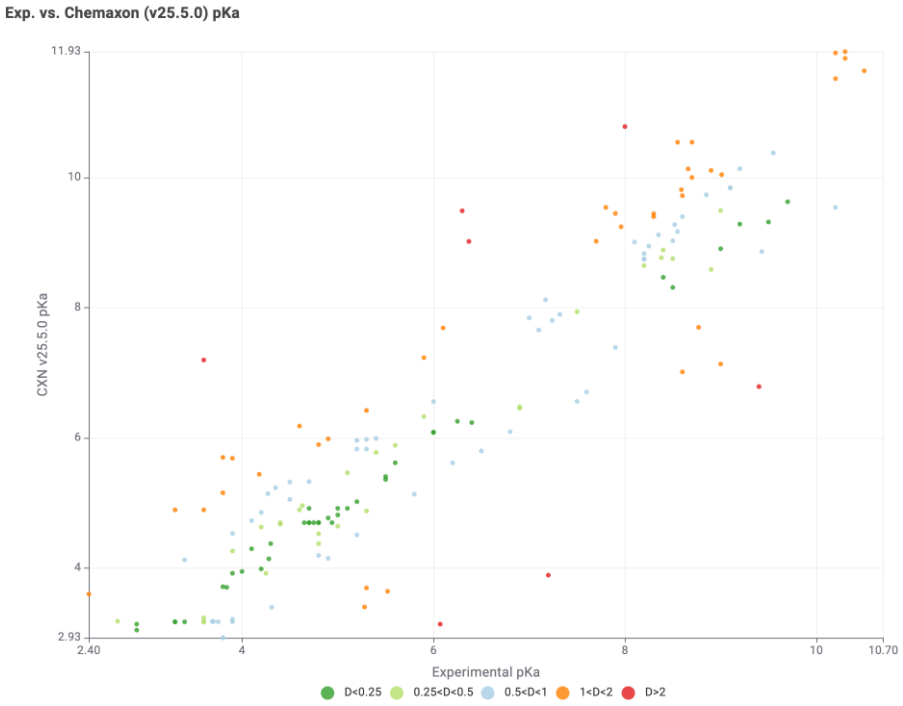
	Occurrence	
$\Delta < 0.25$	672	28%
$0.25 \leq \Delta < 0.5$	483	20%
$0.5 \leq \Delta < 1$	552	23%
$1 \leq \Delta \leq 2$	451	19%
$\Delta > 2$	247	10%

Bottom line

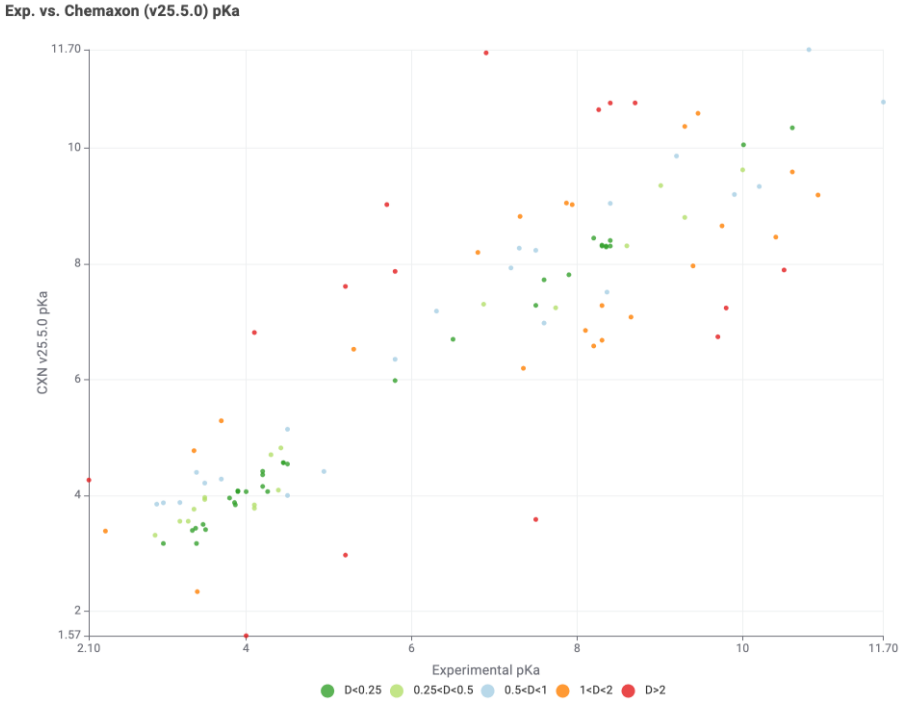
- **Consistent performance** across three datasets with distinct curation philosophies.
- **Best metrics on the cleanest reference data** (AvLiLuMoVe basic: R^2 0.89, MAE 0.44), reflecting a carefully assembled literature compilation.
- **Holds up at scale:** nearly 6,000 ChEMBL-DataWarrior compounds show solid correlation (Pearson ≈ 0.88) and a well-behaved error distribution.
- Same predictor handles acidic and basic centers. No tool switching across the workflow.

Scatter Plots

Novartis basic subset



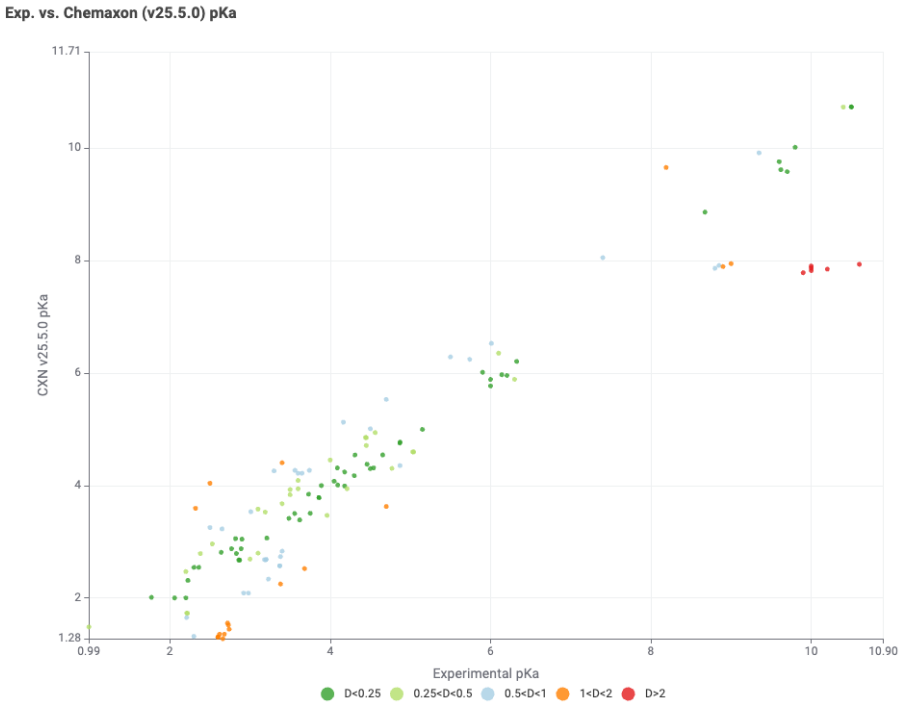
Novartis acidic subset



AvLiLuMoVe basic subset



AvLiLuMoVe acidic subset



ChEMBL-DataWarrior basic subset



ChEMBL-DataWarrior acidic subset

