ChemAxon logP Predictor – SangsterLogP Benchmark

Why these benchmarks

The octanol/water partition coefficient (logP) of the neutral species is a foundational descriptor in drug discovery. It influences solubility, permeability, distribution, protein binding, and metabolic stability, and underpins models for logD, PAMPA, and CNS penetration.

The most rigorous test of logP predictor is performance on independent, well-curated experimental references and on brand-new chemistry that cannot be in the model's training distribution.

In May 2026, Cirino *et al.* Published SangsterLogP, the largest publicly available curated logP collection (23,520 unique molecules, logP range –3.8 to 11.7), assembled from James mSangster's lifetime literature compilation of 3,000+ sources and curated by the Tetko group with logD-to-logP conversion and consensus-based outlier removal.

Data set processing

Environment

- cxcalc CLI
- V25.5.0

Reference

- Cirino T., Ermondi G., Charochkina L., Tetko I.V. **2026** *SangsterLogP – the largest publicly available dataset of logP values.* Sci Data (in press).

Tested Subsets

① SangsterLogP Huuskonen

Classical, well-characterized drug-like and small-molecule chemistry. The kind of reference set against which every logP predictor has been compared since the 1990s.

- Classical shake-flask reference compounds
- A high-confidence sanity check of molecule calibration

Statistical metrics

n = 1,849	Pearson	R^2	MAE	RMSE
v25.5.0	0.955	0.911	0.317	0.446

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	971	53%
$0.25 \leq \Delta < 0.5$	540	29%
$0.5 \leq \Delta < 1$	277	15%
$1 \leq \Delta \leq 2$	53	3%
$\Delta > 2$	8	0.4%

② SangsterLogP Prospective

The hardest possible test. Compounds drawn from publications dated 2024 and later, guaranteed not have been in the Chemaxon training data. Spans logP –0.34 to 5.60, mean 2.35, drug-like chemistry.

Statistical metrics

n = 327	Pearson	R^2	MAE	RMSE
v25.5.0	0.910	0.776	0.371	0.477

Ranged delta distribution

	Occurrence	
$\Delta < 0.25$	131	40%
$0.25 \leq \Delta < 0.5$	112	34%
$0.5 \leq \Delta < 1$	76	23%
$1 \leq \Delta \leq 2$	7	2%

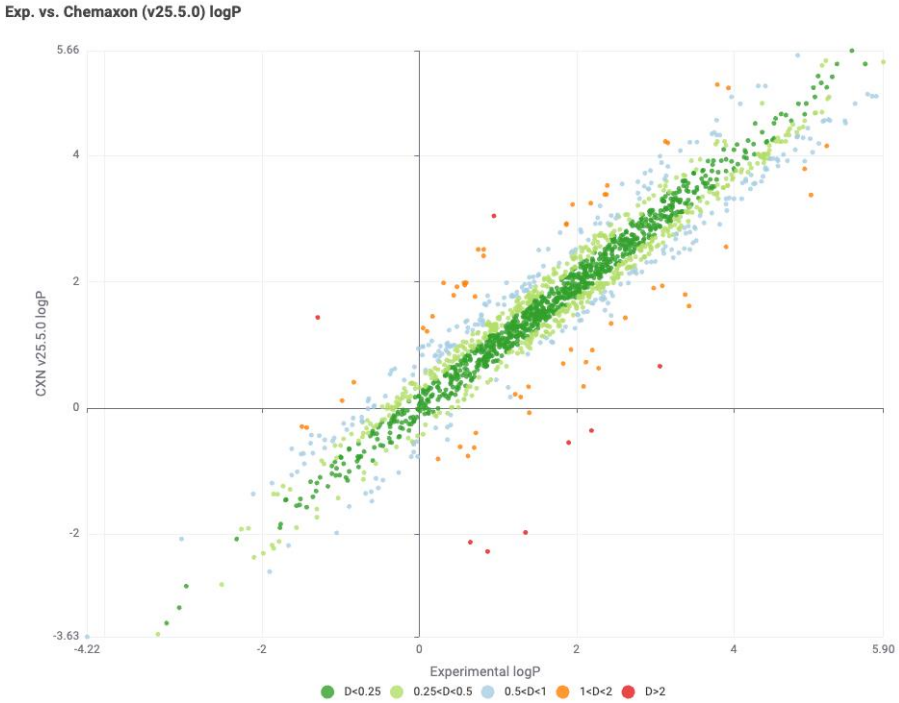
$\Delta > 2$	1	0.3%

Bottom line

- **Essentially perfect calibration on classical chemistry**: on the 1,849 Huuskonen reference compounds the Chemaxon prediction lines up against shake-flask experiment with slope 1.00, intercept 0.00, RMSE 0.45 (at the noise floor of shake-flask logP measurement itself).
- **Robust on brand-new chemistry**: the 327-compound Prospective set, drawn from 2024+ literature and guaranteed out-of-training, gives RMSE 0.48 with 98% of predictions within one logP unit.
- **Fit for purpose** for ranking analogs, lipophilicity flagging in virtual libraries, ADME triage, and as an upstream descriptor for logD7.4 and solubility models, **without post-hoc calibration**.

Scatter Plots

SangsterLogP Huuskonen



SangsterLogP Prospective

