Molecular Properties of the Solubility Panel

Descriptor table and ESOL solubility brief

Zatoona Labs · October 2026

Summary

The primary result is a row-preserving descriptor table for 80 supplied molecules, alongside standardized structures and flags. Seventy-nine structures were readable and described; 78 received an aqueous-solubility estimate. Delaney's ESOL equation was compared with the supplied measurements without refitting: 74 eligible pairs gave an RMSE of 0.997 log units, MAE of 0.762, and bias of +0.097 (prediction minus measured). In a 15-compound control set scored in the same run, RMSE was 1.545 and bias was -0.703, showing that an individual numerical estimate should not be treated as precise. The table is the lookup result; this brief highlights its applicability and limitations.[3]

1 Findings

Among 79 molecules with calculated descriptors, 73 had no Lipinski rule-of-five violations and six had one violation; none had more than one. Five of 79 exceeded at least one Veber criterion. These are screening rules, not clinical or efficacy judgments.[1, 2]

ESOL estimates span -6.678 to $+1.619$ $\log_{10}$ mol/L among the 78 scored structures. The measured comparison includes 67 in-domain and seven out-of-domain compounds. The in-domain subset had RMSE 0.957 log units, while the all-pair RMSE was 0.997. Figure 1 shows the paired values against identity and a one-log-unit band. The model's in-domain control RMSE of 1.545 log units is larger than the difference between many candidate estimates, so close ranks should not be interpreted as meaningful separations. ESOL is an empirical structure-based estimate, not a measurement.[3]

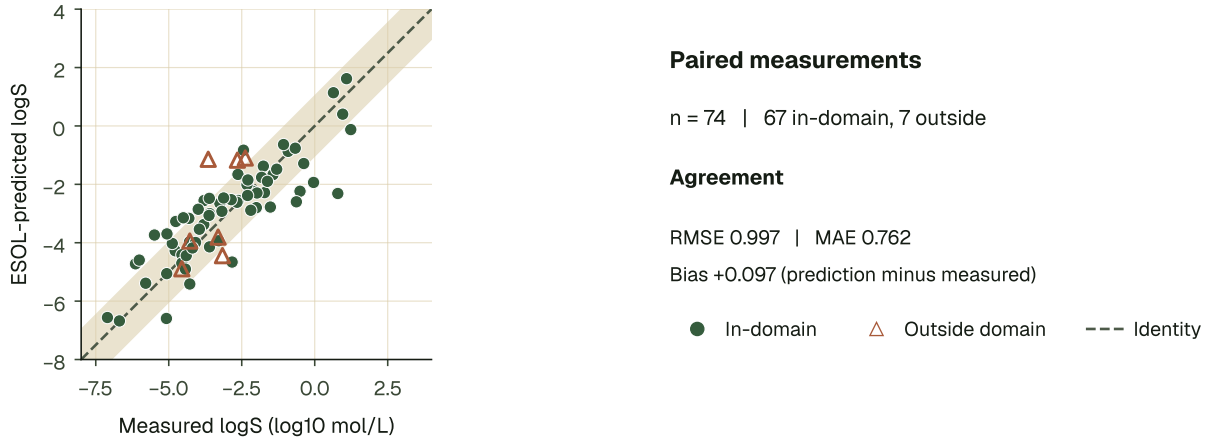


Figure 1. ESOL estimates versus the supplied measured logS values for the 74 eligible pairs. Open triangles mark the seven compounds outside the measured-chemistry domain. The shaded region is within one log unit of the identity line.

2 Flags

Seven of the 78 scored molecules were outside the reference chemistry at the similarity threshold: M006 famotidine, M014 omeprazole, M038 etodolac, M057 riboflavin, M065 sulindac, M078 diltiazem, and M080 allopurinol. Their estimates remain in the table but merit lower confidence; M078 was closest to the cutoff (maximum similarity 0.344 versus 0.353). One entry, M045 acetylsalicylic acid, had invalid SMILES and no calculated properties. M073 (1-dodecylpyridin-1-ium chloride) retained its row but has no ESOL estimate because the standardized parent remains permanently charged.

Standardization neutralized charges or removed counter-ions in five entries (M041, M042, M050, M073, and M079). Properties describe the standardized parent, not necessarily the uploaded salt or charged form. The paired comparison excludes four scorable standardized forms (M041, M042, M050, and M079), whose supplied measurements are not directly comparable to parent structures. M073 has no estimate, and M045 is unreadable. M051 and M079 also share a parent after standardization. These flags prevent interpreting salt measurements as predictions for a different chemical form.

3 Methods

Structures were processed with RDKit 2026.09.1. Molecular weight, Wildman–Crippen cLogP, topological polar surface area, hydrogen-bond donors and acceptors, and rotatable bonds were calculated with RDKit descriptors. Lipinski violations count molecular weight above 500, cLogP above 5, donors above 5, or acceptors above 10; Veber exceedance is polar surface area above 140 $Å^2$ or more than 10 rotatable bonds. Tautomers and stereochemistry were left as supplied. ESOL predicts logS in $\log_{10}$ mol/L from structure using Delaney's published equation, with no fitting to this panel.[1–3]

Applicability was assessed against AqSolDB, a CC0 open measured-solubility reference resource, using ECFP4 Tanimoto similarity to the nearest reference compound; the cutoff was 0.353. This is a chemistry-similarity screen, not a calibrated probability of prediction error.[4, 5]

4 Limitations

The 74-pair comparison is a within-panel benchmark, not independent external validation, and measured logs were not used to fit ESOL. The model control error and mean bias show substantial uncertainty even for scorable structures. Out-of-domain status describes weak similarity to measured reference chemistry, not a claim that a specific estimate is wrong. Standardized salts and charged forms may have solubility behavior different from their neutralized parents. The input provides no permeability or dose, so a Biopharmaceutics Classification System class is not assessable. Results are computational triage only.

References

[1] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46(1-3):3–26. PMID 11259830 10.1016/s0169-409x(00)00129-0

[2] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002;45(12):2615–2623. PMID 12036371 10.1021/jm020017n

[3] Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci. 2004;44(3):1000–1005. PMID 15154768 10.1021/ci034243x

[4] Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Sci Data. 2019;6(1):143. PMID 31395888 10.1038/s41597-019-0151-1

[5] Harvard Dataverse. AqSolDB: A curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. 2019. 10.7910/DVN/OVHAW8