Aqueous Solubility: Key Concepts and Implications for Machine-Learning Models

Aqueous solubility is the maximum amount of a solute that can dissolve in water under defined conditions until a saturated equilibrium between dissolved and undissolved phases is reached. It is a thermodynamic property that depends on temperature, pressure and the chemical nature of both solute and solvent. In pharmaceutical and environmental contexts, solubility is commonly reported as LogS, the base-10 logarithm of molar solubility, which compresses values that may span many orders of magnitude into a compact numeric range for analysis and modelling [1].

At the molecular level, dissolution requires breaking intermolecular interactions in the solid (crystal lattice or amorphous matrix) and forming favourable interactions between solute species and water. Compounds that are strongly crystalline, with tightly packed, low energy lattices, typically show low solubility, whereas amorphous forms and solvates are higher in free energy and can display markedly higher apparent solubility and faster dissolution [5]. Functional groups that can ionise, donate or accept hydrogen bonds, or increase polarity generally enhance interaction with water and thus solubility, while large hydrophobic regions reduce it. For ionisable molecules, pH influences the fraction of ionised versus neutral species. Because ionised forms are usually much more water soluble, the total solubility can change by several orders of magnitude across the physiological pH range, as described quantitatively by the Henderson–Hasselbalch relationship [2].

Measuring equilibrium aqueous solubility is experimentally demanding. The traditional saturation shake-flask method adds excess solid to water or buffer, agitates the suspension at controlled temperature until equilibrium is assumed, then separates undissolved solid and quantifies the dissolved concentration by an analytical method such as HPLC or UV/vis spectroscopy [3].

In early discovery, kinetic solubility assays are widely used for throughput reasons. In such assays, the solute is first dissolved in a co-solvent (often DMSO), then titrated into aqueous medium until turbidity or precipitation is observed. These measurements are useful for flagging compounds likely to precipitate in screening, but they depend strongly on protocol and do not necessarily equal the thermodynamic equilibrium solubility.

Further complexity arises from solid-state effects, pH and temperature. Many organic compounds can exist in multiple polymorphs, solvates, hydrates or amorphous phases, each with different solubility. Solid-form transformations can occur during the experiment, changing the effective solubility with time. The pH-dependence of ionisable compounds means that solubility values cannot be interpreted without the corresponding pH and pKa information [2]. Temperature also matters: most solids become more soluble as temperature increases, so data collected at 25 °C are not directly comparable to data at 37 °C.

Differences in assay type (thermodynamic versus kinetic), pH, temperature, solid form and analytical technique all contribute. This measurement uncertainty imposes a practical upper limit on the accuracy that purely data-driven prediction models can achieve.

These chemical and experimental realities have direct consequences for machine-learning (ML) modelling of solubility. Many ML models treat solubility as a single scalar label depending only on molecular structure, but in practice each label is tied to a specific set of conditions and a particular experimental protocol. If a training set mixes kinetic and thermodynamic solubility data,

or combines measurements obtained at different and often unspecified pH and temperatures, then the model is effectively trained on inconsistent targets and will struggle to generalise reliably [4].

Careful data curation is therefore a prerequisite for robust solubility models. Curated resources such as AqSolDB restrict data to aqueous media, standardise values as LogS, and attempt to remove duplicates and conflicting records [1]. Additional chemistry-aware filters include excluding measurements in mixed or non-aqueous solvents, restricting pH to a defined window, and discarding obviously inconsistent outliers. In parallel, molecular descriptors should reflect the underlying physics of dissolution. Important features include polarity measures (e.g. topological polar surface area), hydrogen-bond donor and acceptor counts, indicators of ionisable functional groups and predicted pKa, as well as lipophilicity parameters such as logP or logD, which tend to correlate inversely with solubility [1,4]. When available, melting point or other solid-state proxies can provide information about lattice strength, which is also linked to solubility [5].

Beyond numeric accuracy, mechanistically informed or hybrid strategies may be advantageous. For example, recent work decomposes the problem into prediction of intermediate physicochemical quantities (such as intrinsic solubility, lipophilicity and melting point) that are then combined in mechanistically motivated frameworks [4]. Such approaches can improve interpretability and define a clearer domain of applicability.

In summary, aqueous solubility is a central but experimentally challenging property that depends on both molecular structure and environmental conditions. The same complexities that make solubility difficult to measure also complicate the construction of predictive ML models. By integrating sound chemical understanding of solid-state behaviour, pH-dependence and experimental uncertainty with systematic data curation and physically meaningful descriptors, it is possible to build more robust and interpretable solubility models. This chemistry-aware approach is essential if solubility predictions are to be trusted and used effectively in drug discovery and related fields.

Future Strategy: From Simple Descriptors to Graph Neural Networks

Looking ahead, the failure of our current models on the "Challenge 2" dataset (which contained highly insoluble "Brick Dust" molecules) highlights a major limitation: standard 2D descriptors cannot see how tightly a molecule packs into a crystal. To fix this, we should move beyond simple fingerprints and adopt **Graph Neural Networks (GNNs)**. A very recent study by Ulrich et al. (2025) showed that GNNs can learn these complex structural patterns directly from the molecular graph, achieving much higher accuracy than traditional methods [6]. Furthermore, we must stop assuming our model can predict everything. We need to implement an **"Applicability Domain"**, essentially a safety zone that flags new compounds that are too chemically different from our training data. As highlighted by Llompart et al., knowing when *not* to predict is just as important as the prediction itself to avoid confident but wrong answers on novel drugs [7]. By combining GNNs with strict safety checks, we can turn our model from a statistical guessing machine into a reliable chemical tool.

References

[1] Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB: a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Scientific Data. 2019; 6:143. DOI: 10.1038/s41597-019-0151-1.

[2] Völgyi, G.; Baka, E.; Box, K. J.; Comer, J. E.; Takács-Novák, K. Study of pH-dependent solubility of organic bases: Revisit of the Henderson–Hasselbalch relationship. Analytica Chimica Acta. 2010; 673(1):40–46. DOI: 10.1016/j.aca.2010.05.022.

[3] Baka, E., Comer, J. E. A., & Takács-Novák, K. (2008). *Study of equilibrium solubility measurement by saturation shake-flask method using hydrochlorothiazide as model compound.* **Journal of Pharmaceutical and Biomedical Analysis, 46**(2), 335–341. https://doi.org/10.1016/j.jpba.2007.10.030

[4] Intrinsic Aqueous Solubility: Mechanistically Transparent Data-Driven Modeling of Drug Substances. Pharmaceutics. 2022; 14(10):2248. DOI: 10.3390/pharmaceutics14102248.

[5] Direct Measurement of Amorphous Solubility. Accessible via PubMed Central at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6750642/.

[6] Ulrich, N., et al. Prediction of the water solubility by a graph convolutional-based neural network on a highly curated dataset. *Journal of Cheminformatics*, 17, 55 (2025).

[7] Llompart, P., et al. Will we ever be able to accurately predict solubility? *Scientific Data*, 11, 303 (2024).