An Investigation into Aqueous Solubility

Sadia Chaudhry; Zenab Khan

Introduction: why solubility matters?

Fundamentally, a drug to be efficient must have high bioavailability and for that it needs to dissolve in fluid which is measured in terms of aqueous solubility, which is a key physicochemical property in drug discovery [1]. It affects delivery, absorption, and how a drug interacts and mixes in gastrointestinal fluids.[2] Low aqueous solubility can lead to unexpected drug behavior, and it's estimated that up to 40% of drug candidates have poor solubility, making this one of the most common challenges in modern drug development.[1] Without considering the solubility, handling and administering drugs will not only be inefficient but impractical.[1,2] Since experimental solubility assays are resource intensive and relatively slow, it is expensive to calculate experimental solubility in terms of time and costs.[3] Thus, chemists tend to use in-silico and machine learning approaches to predict solubility of a chemical compound by considering the structure of the compound, where aqueous solubility is expressed as LogS, which is defined as base-10 logarithm of solubility in mol/L under defined aqueous conditions.[3,4]

What exactly is “aqueous solubility” and LogS?

Aqueous solubility can be defined as the equilibrium concentration of a compound which dissolves in water under specific and strict conditions such as temperature, physical state and thermodynamic balance.[5] For drug like molecules this value changes significantly with slight changes in numerous factors such as ionization state, solid form, pH and other experimental factors. In practice, these factors can vary significantly, leading to ambiguity and making it hard to interpret solubility values uniformly, so these solubility values are usually reported as LogS, which compresses the broad spectrum of solubility values into a manageable range, which supports quantitative modelling and analysis. Curated and publicly available datasets such as AqSolDB stores solubility as LogS along with molecular structural information.[3]

How is aqueous solubility measured?

Aqueous solubility is typically determined by bringing an excess of solid compound into contact with an aqueous medium, allowing the system to equilibrate, and then quantifying the dissolved concentration in the clear phase. In experimental settings it is generally measured in two ways, the “shake flask” or equilibrium solubility method, where the chemical compound is incubated in excess water or buffer at specific conditions, especially temperature, prolonged mixing, separation of undissolved material, and analysis of the supernatant by UV spectroscopy, HPLC or LC–MS.[6,7] While, this method lets us directly measure the equilibrium solubility but is time consuming.

Alternatively, kinetic solubility assays are faster, giving high throughput methods usually used in early drug discovery. In these, compounds are usually dissolved in solvents like DMSO, until precipitation forms. Later the undissolved precipitation is calculated using techniques like turbidimetry, nephelometry or other techniques.[6,7] Even though these

techniques are robust and allow faster assessment, there might be different values of solubility between labs, literature sources, and experimental conditions, due to variations in temperature, mixing time, buffer composition, or form of compound.[3]

This experimental variability introduces noise and bias into compiled datasets, making values from different sources not always strictly comparable. Understanding how these factors influence aqueous solubility and methodological inconsistencies is crucial for interpreting and modelling aqueous solubility data accurately. [3,4]

Why do these measures vary so much, and what makes them difficult to deal with?

Public solubility datasets, such as AqSolDB, merge the results from multiple literature sources, most of which uses unique laboratory procedures, experimental conditions, or measurement approaches.[4] While AqSolDB unifies nine publicly available data sources into one curated dataset, namely, 9,982 individual compounds, and normalizes values into a single LogS format, validates molecular identifiers, and provides basic 2D molecular features, it cannot avoid the intrinsic variability associated with the underlying experimental designs.[3]

Measured solubility for the same chemical compound can differ by at least 0.5 LogS units due to the differences in physical state, temperature, equilibrium duration, or assay technique, so chemical compounds are thus categorized into trust classes based on the consistency of their measurements.[7] As a result, even after an intensive review process, this dataset yet contains results of varying certainty, bringing together readings obtained under various, and sometimes unknown conditions, which introduces variability and bias that any cheminformatics or machine learning analysis needs to be aware of.[4]

Implications for Machine Learning pipelines

The discussion above already pointed out that aqueous solubility is not a perfectly defined quantity, but rather it is influenced by factors such as pH, temperature, physical state, and experimental protocols, and that even trusted, curated datasets such as AqSolDB still contain measurements gathered under varying conditions, leading to different levels of reliability.[3,7] As a result, LogS values fed into models include both inherent lab variability and heterogeneity from mixing thermodynamic, apparent and intrinsic solubility data, so assuming every data point is equally precise doesn't reflect reality.[3,4]

When machine learning models are trained on such heterogeneous data without considering reliability groups, experimental conditions or inconsistencies between sources, they may learn patterns that have more to do with measurement noise and dataset specific biases rather than real solubility traits.[3,4] This can lead to models that perform good on test subset of known data but fail to generalize on new unseen solubility measurements obtained under different, strict conditions.[4]

A more chemically informed cheminformatics strategy will focus on higher-reliability subsets of AqSolDB, be explicit about the definition and experimental conditions of the modelled LogS endpoint and evaluate models on independent test sets that resemble wellcontrolled experimental data.[3,4] These steps take into consideration the complex, condition dependent nature of aqueous solubility, boosting reliability, and interpretability of predictions.

Conclusion

In summary, knowing how aqueous solubility is defined, measured and curated reveals that LogS values used for modelling are condition dependent and often contain measurement errors, so they're far from consistent.[3,4] Recognizing these limitations is essential for interpreting model performance and for designing machine learning workflows that rely on reliable, chemically meaningful solubility data.[4] This perspective shapes better solubility predictions and their interpretation in drug development.[3,4]

References

1. Coltescu, A. R., Butnariu, M. & Sarac, I. The importance of solubility for new drug molecules. *Biomedical and Pharmacology Journal* **13**, (2020).
2. Savjani, K., Gajjar, A. & Savjani, J. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm* **2012**, 195727 (2012).
3. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* **6**, (2019).
4. Llompart, P. *et al.* Will we ever be able to accurately predict solubility? *Sci Data* **11**, (2024).
5. Oja, M., Sild, S., Piir, G. & Maran, U. Intrinsic Aqueous Solubility: Mechanistically Transparent Data-Driven Modeling of Drug Substances. *Pharmaceutics* **14**, (2022).
6. Larsson, J. Methods for measurement of solubility and dissolution rate of sparingly soluble drugs. in (2009).
7. Veseli, A., Žakelj, S. & Kristl, A. A review of methods for solubility determination in biopharmaceutical drug characterization. *Drug Dev Ind Pharm* **45**, 1717–1724 (2019).