Atia Tul Wahab (7069720)
Aaryan Jaitly (7068949)

Report

Aqueous solubility refers to "the maximum amount of compound that can dissolve in water under given conditions" **[1]**. This amount is measured when solubility reaction reaches the equilibrium state **[8]**. The most important physicochemical factor that influences the absorption of orally administered drugs is solubility **[8]**. Intrinsic solubility is the solubility of an ionizable compound at equilibrium at a pH where it exists as fully neutral or unionizable form **[8]**. It helps to understand chemical, pharmaceutical and environmental behaviour of a drug **[5]**.

Most drug molecules are ionizable, and their solubility varies with pH. As a result, apparent or total solubility increases when the ionized form is favored, which explains the dependency of solubility on pH values interpreted through the Henderson–Hasselbalch relationship **[8]**. Although the ideal equation can describe simple systems, the solubility also varies depending on the physical state of the compound (polymorphism), which tends to change during experiments. Solubility measurements for amorphous solids are higher as compared to their crystalline forms as they lack long-range molecular order **[2,7]**. Temperature effects kinetics of the system and thus adds another dimension to solubility. For a wide range of pharmaceutical compounds, solubility varies with temperature **[3,4]**.

Accurate solubility measurement requires standardized methodology. The saturation shake-flask method is widely used for determining equilibrium solubility (Völgyi et al., 2010; Janssen et al., 2019). The technique requires careful control of temperature, pH, and sampling because small deviations—especially in filtration or sample handling—can significantly affect the measured concentration **[2,4]**. Despite standardized methods, considerable variability exists in solubility values reported across the literature. Many datasets lack essential experimental details such as temperature, pH, or sample preparation procedures, and repeated measurements of the same compounds can differ substantially **[1]**.

Such variability may arise from crystallinity differences, residual amorphous content, aggregation, self-association, or inconsistencies in analytical detection **[2,7,8]**. These challenges highlight why solubility remains difficult to measure precisely and why curated intrinsic solubility datasets paired with mechanistically transparent modeling approaches are now preferred in prediction and drug development **[3]**.

The **AqSolDb dataset [1]** consists of two primary columns: SMILES and , where represents the logarithm of solubility. Upon generating a histogram or distribution plot (KDE) of the values, the data exhibits significant negative skewness, with a long left tail extending from approximately −6 to −15. The mean and median values are −2.89 and −2.61, respectively. These statistics indicate that the majority of compounds in the dataset possess low water solubility. Most importantly, the 75th percentile remains relatively low at −1.20, suggesting a high bias toward lipid-soluble compounds.

Atia Tul Wahab (7069720)
Aaryan Jaitly (7068949)

To prepare the data for modeling, we employed a **sanitization process** to ensure that each SMILES string follows chemical validity rules and represents a realistic, chemically valid molecule **[9]**. This step is essential because downstream processes, such as descriptor calculation, depend heavily on accurate molecular structures **[10, 11]**. Furthermore, for molecules capable of existing in multiple tautomeric forms, the most stable representation was selected to ensure consistency **[12]**.

Following standardization, descriptor generation was performed. The initial **Mordred** generation gave 1,612 descriptors, which were reduced to 721 after eliminating columns containing null (NA) values. While training on all remaining descriptors resulted in high training accuracy, it likely contributed to subpar performance on the challenge dataset due to overfitting. Additionally, we got **166 MACCS fingerprints**, which are binary vectors representing specific molecular substructures. Unlike continuous descriptors, the MACCS bit vector can only be adjusted by removing specific bits. In this study, the model was trained using the entire bit vector, which may further explain the performance variations across datasets.

Two models were trained using an 80/20 train-test split, a standard approach in machine learning. Initially, The **Support Vector Regression (SVR)** model with a Radial Basis Function (RBF) kernel was used. Fine-tuning was done using leave-one-out cross-validation to optimize the hyperparameters , (epsilon), and (gamma). A grid search identified the optimal parameters. Intuitively, a higher imposes a stronger penalty on errors, while a low implies that only errors very close to the decision boundary are ignored.

However, SVR struggles with highly skewed and noisy data **[1]**. Consequently, Evaluation of **Random Forest** model was done. As an ensemble of decision trees generated via data resampling, Random Forest is better equipped to handle dataset skewness and noise. It increases model diversity by selecting attributes at random, with the final prediction derived from averaging the outputs of all trees. While Random Forest outperformed SVR on both the training and challenge partitions of the AqSolDB data, both models struggled to accurately represent solubility trends on the external challenge dataset, highlighting the difficulty of generalizing from this specific data distribution. The main challenges for our models are:

1. **Data Quality:** Prioritize dataset homogeneity by filtering based on experimental sources to reduce noise.
2. **Feature Selection:** Retain only descriptors strictly relevant to to avoid performance degradation.
3. **Hyperparameter Tuning:** Tailor tuning strategies to the specific characteristics of the dataset rather than relying on generic cross-validation.

Atia Tul Wahab (7069720)
Aaryan Jaitly (7068949)

References:

1. AqSolDB Consortium. (2019). *AqSolDB: A curated reference set of aqueous solubility data*. *Scientific Data, 6*, 143. https://doi.org/10.1038/s41597-019-0151-1
2. Avdeef, A. (2019). *Solubility of ionizable drugs*. In *Absorption and Drug Development* (pp. 27–68). Wiley. https://doi.org/10.1002/9781118970967.ch3
3. Demirbaş, A., Çelik, A., et al. (2018). *Temperature and solvent effects in the solubility of some pharmaceutical compounds*. *Journal of Molecular Liquids, 272*, 846–855. https://doi.org/10.1016/j.molliq.2018.10.097
4. Janssen, M., Verboom, W., et al. (2019). *Study of equilibrium solubility measurement using the saturation shake-flask method*. *Journal of Pharmacy and Pharmacology, 71*(1), 100–111. https://doi.org/10.1111/jphp.13040
5. Oja, M., Maran, U., & Veski, P. (2022). *Intrinsic aqueous solubility: Mechanistically transparent data-driven modeling*. *Pharmaceutics, 14*(10), 2248. https://doi.org/10.3390/pharmaceutics14102248
6. Sorkun, M. C., Khetan, A., & Er, S. (2019). *AqSolDB: A curated reference set of aqueous solubility data*. *Scientific Data, 6*, 143. https://doi.org/10.1038/s41597-019-0151-1
7. Ueda, K., Takahashi, Y., et al. (2025). *Amorphous solubility advantage: Theoretical and experimental perspectives*. *Journal of Pharmaceutical Sciences*. https://doi.org/10.1016/j.xphs.2024.12.003
8. Völgyi, G., Baka, E., Box, K. J., Comer, J. E. A., & Takács-Novák, K. (2010). *Study of pH-dependent solubility of organic bases: Revisit of the Henderson–Hasselbalch relationship*. *Analytica Chimica Acta, 673*(1), 40–46. https://doi.org/10.1016/j.aca.2010.05.022
9. Chemaxon. (n.d.). SMILES. Chemaxon Documentation. Retrieved November 24, 2025, from https://docs.chemaxon.com/display/docs/formats_smiles.md
10. Datamol. (n.d.). Preprocessing. Datamol Documentation. Retrieved November 24, 2025, from https://docs.datamol.io/stable/tutorials/Preprocessing.html
11. RDKit. (n.d.). Molecular Sanitization. In The RDKit Book. Retrieved November 24, 2025, from https://www.rdkit.org/docs/RDKit_Book.html#molecular-sanitization
12. Oxford Protein Informatics Group (OPIG). (2022, May). Molecular Standardization. BLOPIG. Retrieved November 24, 2025, from https://www.blopig.com/blog/2022/05/molecular-standardization/